MSc Artificial Intelligence
Master Thesis

Improving Completeness in Deep Research Agents through Targeted Enrichment

by
Jesse Wonnink
13213334

February 15, 2026

University Supervisor:
Dr Paul Groth

Examiner:
Dr Paul Groth

External Supervisor:
Jakub Zavrel

Second reader:
Dr Mohammad Alian Nejadi

Universiteit van Amsterdam

Contents

Abstract

Deep research agents represent a significant advance in AI-assisted information synthesis, capable of conducting comprehensive investigations that traditionally required substantial human effort. However, ensuring completeness in automatically generated research reports remains challenging: existing systems rely on ad-hoc query decomposition through prompt engineering, providing no formal guarantees about coverage or diversity, and evaluation frameworks often assess single dimensions rather than holistic report quality.

This thesis addresses these limitations through three primary contributions. First, we propose a multi-dimensional framework that operationalizes completeness as three interdependent aspects: coverage (breadth and relevance of information), grounding (citation accuracy and factual consistency), and presentation quality (clarity and structural coherence). Second, we present HERO (High Enrichment Retrieval Orchestrator), a hierarchical deep research architecture that combines submodular optimization for query diversification with a novel two-stage enrichment mechanism that identifies and addresses information gaps through targeted follow-up investigation. Third, we conduct comprehensive evaluation across both academic (ScholarQABench) and general knowledge (DeepResearchGym) domains, enabling holistic evaluation of Deep Research agents.

HERO achieves state-of-the-art performance on both benchmarks, with the highest coverage metrics (Key Point Recall of 67.63 on DeepResearchGym), strongest grounding (Citation F1: 91.57), and superior presentation quality scores. Ablation studies reveal that submodular optimization and hierarchical enrichment each contribute distinct improvements, with synergistic effects when combined. However, important limitations remain: our analysis reveals systematic sycophantic bias where the system adapts argumentative positions to match query framing, demonstrating that architectural improvements alone cannot overcome inherent behavioral patterns in foundation models.

This work contributes both a concrete system demonstrating measurable improvements in report completeness and a framework for multi-dimensional evaluation of deep research agents. As these systems evolve toward deployment in high-stakes domains, the architectural principles and evaluation methodology established here provide a foundation for building reliable, comprehensive AI research assistants.

Chapter 1

Introduction

Deep Research (DR) agents represent Artificial Intelligence's (AI) evolution from question-answering to autonomous investigation[26]. DR agents are AI systems powered by Large Language Models (LLMs) that integrate dynamic reasoning, adaptive planning, and iterative tool use to acquire, aggregate, and analyze external information [26]. By combining retrieval, reasoning, and iterative planning, these systems can conduct the kind of comprehensive research that typically required days or weeks of dedicated human effort. However, since these reports can now be mass produced, it would be beneficial to have a concrete measure of their completeness. Unlike task-specific Retrieval Augmented Generation (RAG) [60] applications, DR agents tackle open-ended research tasks where the scope and necessary depth are not predefined: they formulate search strategies, pursue follow-up questions based on initial findings, and synthesize information across multiple sources into structured reports [60]. Current approaches to ensuring completeness typically generate multiple queries through techniques like query decomposition, relying on prompt engineering to encourage diversity and relevance. However, this ad-hoc approach provides no guarantees about query diversity or coverage optimization—LLMs often generate redundant queries that reflect common perspectives while missing less obvious, yet relevant angles. Moreover, as these agents become more capable and widely deployed, the lack of standardized evaluation frameworks for measuring completeness becomes increasingly problematic [26].

To address the limitations of ad-hoc query generation, this thesis introduces **HERO**: (**H**igh **E**nrichment **R**etrieval **O**rchestrator), a DR agent designed to systematically improve research completeness. HERO moves beyond simple query decomposition by implementing two core architectural principles. First, it employs a hierarchical enrichment strategy that breaks down complex topics into focused sub-investigations, allowing for adaptive and deep exploration of the information space. Second, it leverages an optimization method to generate query sets with provable diversity and coverage guarantees, ensuring that the research process explores a wide range of distinctly relevant perspectives rather than redundant ones.

1.1 Problem Statement

1.1.1 Completeness

Completeness as a construct is inherently multifaceted and difficult to measure: in complex research tasks, there is no definitive ground truth for what constitutes a 'complete'

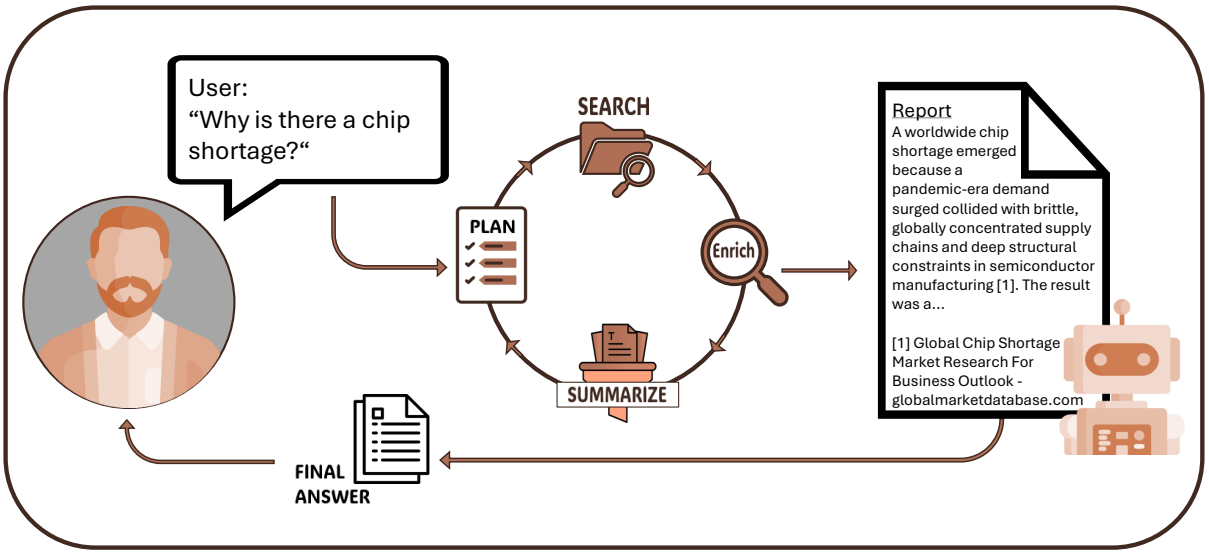

Figure 1.1: Deep research agent workflow: agents iteratively search, plan, synthesize, and enrich information before producing comprehensive reports. This thesis addresses how to systematically improve the completeness of such outputs.

answer, and any benchmark will necessarily impose arbitrary boundaries on scope and depth. Unlike traditional QA evaluation, which assesses factual correctness of short answers, Deep Research (DR) agents must generate comprehensive, structured reports that synthesize information across multiple sources.

Existing RAG evaluation frameworks often assess multi-hop reasoning [18]. Multi-hop are question that cannot be answered in 'one step'. E.g. 'what is the name of the father of the president of Argentina in 1980.' One would need at least two steps to find this answer efficiently. their benchmarks target question-answering tasks rather than long-form report generation. Similarly, current DR agent benchmarks focus primarily on task-specific performance metrics (e.g., QA accuracy) rather than holistic assessment of research reports as integrated artifacts [26]. Expert-level benchmarks such as GAIA [43], ARC-AGI [9], and Humanity's Last Exam [51] have been proposed for evaluating DR agents; however, these datasets measure only answer accuracy, making them ill-suited for assessing the quality of comprehensive research reports. A clear framework for evaluating completeness in DR reports is therefore necessary.

We propose that completeness in DR agent outputs consists of three interdependent dimensions:

1. **Coverage**: Ensuring all relevant aspects of the research question are addressed with detail commensurate to query expectation

2. **Grounding**: Factual accuracy supported by proper citations and verifiable source attribution

3. **Presentation Quality**: Narrative coherence, readability, and structural organization

These dimensions are necessary and sufficient: high coverage without grounding produces unverifiable content; accurate grounding without adequate coverage leaves critical questions unanswered; poor presentation quality renders even complete and well-grounded

information unusable in practice. Any additional considerations for completeness, such as multi-modal integration (e.g., charts, diagrams) and source criticism, fit under the umbrella of Presentation quality and grounding respectively. A complete DR agent report should strive to jointly optimize all three dimensions.

1.2 Research Questions

This thesis is guided by a primary research question, which is broken down into two specific subquestions:

RQ: ***How can we improve completeness in deep research agents?***

RQ.1: **System Effectiveness:** Does the proposed HERO architecture improve completeness in generated research reports compared to existing baseline deep research agents?

RQ.2: **Algorithmic Contributions:** What are the individual and combined contributions of submodular query optimization and hierarchical enrichment to the observed improvements in coverage?

1.3 Thesis Contributions

To answer the research questions, this thesis makes the following contributions:

- **Failure mode analysis:** A qualitative investigation and categorization of common failure modes in baseline deep research agents, informing the design of our proposed system (Appendix A).
- **Completeness framework:** An operationalization of completeness as three interdependent dimensions—coverage, grounding, and presentation quality—providing a structured approach to evaluating deep research systems (Section 1.1).
- **HERO system:** A modular, multi-agent architecture that explicitly separates query generation, information extraction, synthesis, and enrichment into specialized components with hierarchical information isolation (Chapter 4).
- **Submodular query optimization:** Application of facility location objectives to query decomposition, providing formal guarantees for balancing relevance and diversity in information retrieval (Section 2.5.4).
- **Hierarchical enrichment mechanism:** A novel two-stage enrichment approach that identifies and fills informational gaps in generated reports through targeted follow-up queries, improving coverage depth (Section 4.2.3).
- **Comprehensive benchmark evaluation:** A multi-dimensional assessment across DeepResearchGym and ScholarQABench, combining two distinct corpora and eight query domains to enable cross-domain validation and cross-benchmark metric analysis (Section 2.3.

Chapter 2

Background and Related Work

2.1 Historical Context

Deep Research (DR) agents represent the convergence of several research traditions whose developments have enabled the generation of comprehensive, well-grounded research reports. The pursuit of completeness in information systems emerged from High Recall Information Retrieval (HRIR), which traditionally originated in legal e-discovery contexts where achieving near-complete recall (95–100%) is crucial for constructing an accurate case [63]. The field matured through evaluation frameworks established by the TREC Legal Track (2006–2011) and later the TREC Total Recall Track (2015–2016) [54], in which machine learning-based recall outperformed manual human search. Two notable advances were Technology-Assisted Review (TAR) [20] and its successor, Continuous Active Learning (CAL) [12], which iteratively update models after each human relevance judgment, achieving higher recall with 50–80% less review effort than traditional or manual search approaches.

Beyond legal applications, HRIR has expanded to medical systematic reviews—where exhaustive literature searches across multiple databases are required to avoid publication bias and ensure valid evidence synthesis [42]. However, HRIR systems operate primarily at the document level, identifying which documents are relevant rather than extracting and synthesizing specific content. This document-level focus represents a fundamental difference from the content-level completeness required by modern DR agents, who must not only find relevant sources but also address all aspects of a research question with appropriate depth and detail.

Meanwhile, the field of HRIR remained relatively disconnected from developments in general information retrieval, where sparse retrieval methods were soon outpaced by the introduction of Dense Passage Retrieval (DPR) [34]. This shift proved essential for future search systems built on LLMs generating informed queries, as it enabled better semantic matching [35] and more sophisticated query optimization methods.

With the advent of large language model systems [8, 48], the field of IR was transformed. A significant advance came with RAG systems [38]. RAG combines a retrieval component with a generative language model: first, relevant text passages are retrieved from a knowledge base, then these passages are injected into the LLM's context to ground its generated response [18]. With this development, the output scope moved from the document level to content-level relevance; rather than returning entire documents for users to read, RAG systems extract relevant content chunks and synthesize them into coherent, contextually appropriate responses. The retrieved passages serve as evidence

and source material that the LLM uses to generate accurate, grounded answers while reducing hallucination [59].

However, early RAG systems faced several limitations, including robustness issues where incorrect information could severely influence results and the recognition that a singular query is often insufficient for questions requiring multiple reasoning steps (multi-hop questions). The need for intermediate reasoning and reflection—allowing systems to iteratively refine their understanding—led to the development of agentic AI systems [70, 60], which provided the foundation for modern Deep Research agents.

2.2 Deep Research Agents

Deep Research agents extend agentic RAG frameworks by focusing specifically on open-ended research tasks that require comprehensive information gathering, synthesis, and report generation [27]. Unlike task-specific RAG applications that answer discrete questions, DR agents must formulate research strategies, often pursue multiple lines of inquiry, and produce structured, comprehensive outputs that synthesize findings from diverse sources.

2.2.1 Taxonomy of Existing Systems

Deep Research agents can be analyzed along several principal dimensions. This taxonomy is derived from the framework proposed by [26] (see also [27] for an alternative taxonomy).

Workflow Architecture: Static versus Dynamic Systems

A **static system** follows a structured pipeline with a predetermined sequence of operations, possibly with iterative components, but with a clearly defined beginning and end. Examples include Agent Laboratory [57], which uses a fixed sequence where agents with different roles apply predetermined tools until reaching a target paper count, and LitLLM [1], which follows a rigid query→retrieve→rank→synthesize pipeline with pre-structured output templates. OpenScholar [5] also exemplifies this through its fixed segment→encode→retrieve→rerank→generate sequence, though it includes iterative refinement within this structure.

In contrast, a **dynamic system** does not have a predefined endpoint but iterates, strategizes, and reconsiders on the fly based on intermediate findings. OpenAI's Deep Research [45] and Gemini DeepResearch [19] exemplify this approach by adaptively deciding when to stop or continue exploration in specific areas. PaperQA2 [61] demonstrates dynamic behavior through its autonomous agent that decides when to search more, rewrite keywords, or traverse citations based on intermediate findings. Open-source implementations like gpt-researcher [17] use parallel subquery decomposition with dynamic aggregation, while node-DeepResearch [3] implements an iterative query→search→read→reason loop that continues until a cost budget is exhausted. deep-research [74] introduces user-defined depth and breadth parameters to control exploration scope, fully dynamic systems like open-deep-research [56] operate on time budgets rather than fixed parameters.

For dynamic systems, budget constraints are essential to prevent unbounded exploration. These limiters can be expressed in various forms: maximum research rounds, parallel query limits, maximum thinking time, or token budgets.

Agent Composition: Single-Agent versus Multi-Agent Systems

Single-agent systems integrate planning, tool invocation, and execution within a unified LLM that makes all decisions. OpenScholar [5] demonstrates this through a unified model handling retrieval, re-ranking, and generation with built-in reference checking, while PaperQA2 [61] uses a single agent to autonomously manage keyword rewriting and citation traversal decisions. This streamlined approach enables clearer analysis and explanation but places significant demands on the LLM's computational capabilities, as managing too many tools can overwhelm a single agent [53].

Conversely, **multi-agent systems** leverage multiple specialized agents to collaboratively execute subtasks. PaSa [22] exemplifies an early dual-agent design: a Crawler Agent handles query generation and citation traversal while a Selector Agent filters for relevance, allowing one agent to correct the other's errors. ReSP [31] employs separate 'reasoner' and 'retriever' agents supported by dual memory structures (global evidence and local memory). `GPT-Researcher` [17] assigns parallel agents to research different subqueries simultaneously before aggregating findings.

These multi-agent configurations effectively handle complex, parallelizable research tasks, thereby enhancing flexibility and scalability in open-ended research scenarios. However, they introduce increased coordination complexity, additional communication costs, and multiple levels of memory and context management.

Tool Integration and Functional Capabilities

Beyond their core research objectives, many agent systems incorporate specialized tools to extend their capabilities. Core capabilities include 1) *Code interpreters* that enable agents to execute scripts during inference for data processing, algorithm verification, and model simulation—OpenAI Deep Research [45] uses internal code execution for data analysis and generating visualizations; (ii) *Data analytics modules* that transform raw retrievals into structured insights through statistical computation, interactive visualizations, and quantitative model evaluations; and 2) *Multimodal processing and generation tools* that enable agents to integrate, analyze, and generate heterogeneous data including text, images, audio, and video within a unified reasoning pipeline—`gpt-researcher` [17] supports incorporating diverse sources including images into its reports. [68] generates complete graphs by preselecting an appropriate data structure and accompanying plot. PaperQA2 [61] implement specific academic tools for exploring references based on citation frequency. Interfaces such as DeerFlow [14] use Model Context Protocol [23] to facilitate modular tool integration, enabling dynamic expansion of agent capabilities while maintaining the same architecture.

2.3 Evaluation Approaches for Deep Research Systems

Evaluating the performance of systems designed for long-form report generation, such as those provided by Deep Research agents, necessitates specialized benchmarks that build on, but also extend, traditional IR metrics focused solely on document relevance and accuracy[26]. Traditional multi-hop benchmarks [60] serve as useful proxies for the reasoning and decomposition steps required in DR tasks. However, we argue these datasets are insufficient for evaluating comprehensive DR reports. While multi-hop benchmarks assess whether systems can synthesize knowledge across sources, complete DR reports

must simultaneously achieve factual accuracy, comprehensive coverage, and narrative coherence.

Several benchmarks have been proposed specifically for literature review generation or DR agent evaluation:

PaperQA2 Benchmark [61]: Contributes a benchmark dataset designed to evaluate tasks pertinent to its framework, such as answering questions over scientific corpora and generating synthesis reports. This dataset is useful for assessing systems operating in similar scientific QA and summarization contexts, but is domain specific and also not generalizable

KIWI Benchmark [39]: Focuses on evaluating longer-form generated answers and is distinguished by its development process, which involved iterative refinement based on feedback from professional domain experts. This emphasis aims to align the benchmark's evaluation criteria more closely with real-world quality expectations. The questions are still Mostly QA-style however.

The development of robust and comprehensive benchmarks remains a critical challenge and an active area of research.

2.4 Query Generation and Diversification

An important component of modern RAG architectures is query generation and decomposition [60]. In the pursuit of completeness, a single search query is often insufficient; multiple queries exploring different angles or reasoning steps are required to obtain a comprehensive understanding. Early advances were made by [50], which decomposed multi-hop questions [69] into a composition of subquestions that could be answered independently. A diverse array of decomposition methods subsequently emerged [18, 25, 15, 75, 71, 16].

However, these methods primarily represent engineering solutions without principled frameworks for optimization. There is no informed guidance on what constitutes optimal query diversification, beyond "more different angles is likely better." This lack of formalization is particularly problematic for achieving completeness: queries must be sufficiently diverse to avoid excessive overlap and redundant retrieval, yet sufficiently relevant to the original question to stay on the right level of granularity.

2.4.1 Submodular Optimization in Information Retrieval

Submodular optimization provides a principled framework for set selection under diversity constraints. The foundational work by [44] established that greedy algorithms achieve a $(1 - 1/e)$-approximation. In IR, this can be framed as an relevance vs. diversity control. Early applications include [2], which employed submodular objectives to diversify query results. Similarly it has been applied to recommender systems [6], where submodular optimization balances recommendation accuracy with diversity across item categories. These applications demonstrate that submodular function could be a useful tool in the selection of queries.

Despite these successes in result diversification, submodular optimization has not previously been applied to query decomposition in deep research agents.

2.5 Positioning of HERO

Having established the landscape of deep research agents and their evaluation challenges, we now position HERO within this taxonomy and describe how its design addresses key limitations in existing approaches. HERO builds directly on the evaluation frameworks of OpenScholar [5] and DeepResearchGym [11], the submodular optimization theory from [**?**], and incorporates principles from recent work on submodular query generation [67].

2.5.1 Workflow Architecture: Controlled Iteration with Adaptive Depth

Within the taxonomy established in Section 2.2.1, HERO occupies a hybrid position. The system employs a structured multi-turn pipeline with predetermined maximum iterations ($t_{\max}$), providing the cost predictability and clear termination conditions of static systems. However, within this bounded structure, the enrichment mechanism enables adaptive depth—pursuing targeted follow-up investigation when initial summaries reveal information gaps. This design provides deterministic cost bounds while maintaining the adaptive exploration capabilities that characterize dynamic systems.

The bounded iteration approach addresses a common limitation in fully dynamic systems: without clear stopping criteria, research processes can become computationally prohibitive or, under an agent's purview, stop too early. Conversely, purely static pipelines are not adaptive enough to different levels of complexity. HERO's hybrid approach balances these concerns through controlled iteration with adaptive refinement.

2.5.2 Agent Organization: Specialized Modules with Hierarchical Isolation

HERO implements a multi-agent architecture (Section 2.2.1) where specialized components handle distinct research phases: query generation, information extraction, synthesis, and enrichment. This modularity enables parallel processing of multiple subqueries while maintaining focused expertise within each agent.

A key architectural principle is hierarchical information isolation—each subquery pipeline operates independently, accessing only its own retrieved documents and intermediate summaries. This isolation serves two purposes: (1) it prevents context window overflow that can occur when agents process accumulated information from all parallel pipelines, and (2) it maintains query diversity by preventing pipelines from converging on similar angles based on observing each other's findings. The centralized OrchestrationState coordinates these parallel operations while maintaining global citation tracking and result aggregation. The complete architecture is detailed in Chapter 4.

2.5.3 Query Decomposition: Submodular Optimization

HERO addresses the query selection through submodular optimization, specifically employing a facility location objective that provides explicit control over the diversity-relevance trade-off. This approach offers several advantages: (1) provable approximation guarantees through greedy selection, (2) mathematical control over query similarity through the α parameter, and (3) applicability to both main query decomposition and enrichment query selection. To our knowledge, HERO represents the first application of

submodular optimization to query decomposition in deep research agents, though related work by [67] has explored similar principles in query generation contexts. The formal problem formulation is presented in Chapter 3.

2.5.4 Evaluation Strategy: Multi-Dimensional Completeness Assessment

The evaluation framework employed in this thesis addresses limitations identified in Section 2.3 through three design choices. First, we operationalize completeness as three interdependent dimensions—coverage, grounding, and presentation quality—rather than relying on single metrics that may miss important quality aspects. Second, we evaluate across both academic (ScholarQABench [5]) and general knowledge (DeepResearch-Gym [11]) domains to assess generalization beyond specialized contexts. Third, we conduct systematic cross-benchmark analysis to understand how metric operationalization affects comparative assessment (Appendix A.4).

This multi-dimensional, cross-domain approach enables comprehensive evaluation of where and how HERO's architectural choices impact system performance. The experimental design and metrics are detailed in Chapter 4.3.1, with results presented in Chapter 5.1.5.

Chapter 3

Problem Formulation

To operationalize the pursuit of completeness, we formalize the task of a DR agent as an optimization problem over an iterative research process, subject to practical budget constraints.

3.1 High-Level Objective

Given a user query q_{user}, a document corpus $\mathcal{D}$ (which may comprise web documents, academic papers), and a computational budget B, the objective is to generate a research report R and citation set $\mathcal{C} = \{(c_i, d_i)\}_{i=1}^{m}$, where each citation pairs a claim c_i from R with its source document $d_i \in \mathcal{D}$, that maximizes a completeness function Φ:

$$(R^*, \mathcal{C}^*) = \arg\max_{(R,\mathcal{C})} \Phi(R, \mathcal{C} \mid q_{\text{user}}) \quad \text{subject to} \quad \text{Cost}(R, \mathcal{C}) \leq B \tag{3.1}$$

The completeness function Φ quantifies the latent construct of completeness based on three dimensions: **Coverage**, whether all relevant aspects expected in q_{user} are addressed; **Grounding** is the factual accuracy and support of $\mathcal{C}$; and **Presentation Quality** entails presentation coherence and readability. These dimensions will be quantified through concrete metrics (detailed in Section 4.3.1).

3.2 HERO's Setup

HERO approximates a solution to the optimization problem defined in Equation 3.1 through a structured, multi-turn research process, with a budget constraint.

3.2.1 Budget Constraints

The budget constraint B is operationalized through a dual limit on the research process:

$$T \leq T_{\max} \tag{3.2}$$

$$|\mathcal{Q}_t| \leq k_t \quad \forall t \in \{1, \ldots, T\} \tag{3.3}$$

$$|\mathcal{Q}'_i| \leq k'_i \quad \forall q_i \in \mathcal{Q}_t, \forall t \tag{3.4}$$

where $T \leq T_{\max}$ is the number of research turns executed, k_t bounds the number of queries selected in turn t, and k'_i restricts the enrichment queries generated for each subquery.

3.2.2 Submodular Optimization

For each turn $t \in \{1, \ldots, T\}$, HERO generates a candidate pool V_t of size $n = m \cdot k_t$ queries via prompt engineering, where $m \geq 1$ is a candidate multiplier. An optimal subset $\mathcal{Q}_t \subseteq V_t$ is selected via a facility location objective that maximizes coverage [13]:

$$\mathcal{Q}_t^* = \arg\max_{\mathcal{Q}_t \subseteq V_t} f_\alpha(\mathcal{Q}_t) \quad \text{subject to} \quad |\mathcal{Q}_t| = k_t \tag{3.5}$$

where the submodular objective function is defined as:

$$f_\alpha(\mathcal{Q}_t) = \sum_{j \in V_t} \max\left(\alpha \cdot \text{sim}(\mathbf{e}_0, \mathbf{e}_j), \max_{q_i \in \mathcal{Q}_t} \text{sim}(\mathbf{e}_j, \mathbf{e}_i)\right) \tag{3.6}$$

Here, $\mathbf{e}_j = \phi(q_j)$ denotes the embedding of query q_j, $\mathbf{e}_0 = \phi(q_{\text{user}})$ is the original query embedding, sim is cosine similarity, and $\alpha \in [0, 1]$ is a relevance weight hyperparameter. The function f_α is monotone submodular, admitting a greedy algorithm [44]. The same optimization is applied to select enrichment queries $\mathcal{Q}_i'$ subject to $|\mathcal{Q}_i'| \leq k_i'$.

3.2.3 Research Pipeline

For each selected subquery $q_i \in \mathcal{Q}_t$, HERO executes a research pipeline function Π:

$$\Pi(q_i, q_{\text{user}}, \mathcal{D}) \to (S_i', \mathcal{C}_i') \tag{3.7}$$

This pipeline Π operates in two stages:

1. **Initial Synthesis:** Retrieve relevant documents from $\mathcal{D}$ and synthesize an initial summary with citations:

$$(S_i, \mathcal{C}_i) = \text{Synthesize}(q_i, \mathcal{D}) \tag{3.8}$$

2. **Enrichment:** Analyze S_i for information gaps and, if necessary, generate follow-up queries $\mathcal{Q}_i'$ (subject to $|\mathcal{Q}_i'| \leq k_i'$) to produce an enriched version:

$$(S_i', \mathcal{C}_i') = \text{Enrich}(S_i, \mathcal{C}_i, q_i, q_{\text{user}}, \mathcal{D}) \tag{3.9}$$

where $S_i' \supseteq S_i$ with additional information gathered through follow-up queries, and $\mathcal{C}_i' \supseteq \mathcal{C}_i$ includes the potentially expanded citation set.

The complete pipeline architecture is detailed in Section 4.

3.2.4 Final Report Generation

After completing all research turns, HERO synthesizes the final report R and consolidated citation set $\mathcal{C}$ from all enriched summaries:

$$(R, \mathcal{C}) = \text{Write}(\{S_i' : q_i \in \mathcal{Q}_t, t \in \{1, \ldots, T\}\}) \tag{3.10}$$

where Write represents the final aggregation and report generation by HERO's answer writer module (Section 4.3.1).

Chapter 4

Architecture

4.1 HERO: High Enrichment Retrieval Orchestrator

HERO implements a hierarchical, multi-turn research architecture where independent subquery pipelines explore different aspects of the user question in parallel. Each pipeline produces enriched research summaries that are synthesized into a comprehensive final report. The system maintains global state through a centralized OrchestrationState that coordinates parallel execution.

Figure 4.1 illustrates the complete system architecture. We describe each component in detail below.

All prompts used for agent modules are provided in Appendix F.

4.1.1 System Overview

The system operates in iterative research rounds (up to $t_{\max}$). Each round begins with the Query Generator (step 3) decomposing the user question into diverse subqueries using submodular optimization. These subqueries execute in parallel through independent pipelines, each producing an enriched summary. The Answer Writer then synthesizes all summaries into the final report.

4.1.2 Hierarchical Information Isolation

A key architectural principle of HERO is that each subquery pipeline operates as an independent research branch with isolated context. All agents are designed to maintain focus on their designated research area through mechanisms such as targeted information extraction instructions and submodular optimization for query diversity. This architectural choice keeps redundancy in search—the most expensive operation as shown in Table 5.2—in check.

For instance, when researching environmental impacts of vertical farming, one pipeline investigating "water efficiency" only accesses documents and summaries from its own retrieval—it never sees information gathered by the parallel pipeline researching "energy consumption." This isolation prevents pipelines from converging on similar angles and ensures each maintains its designated focus.

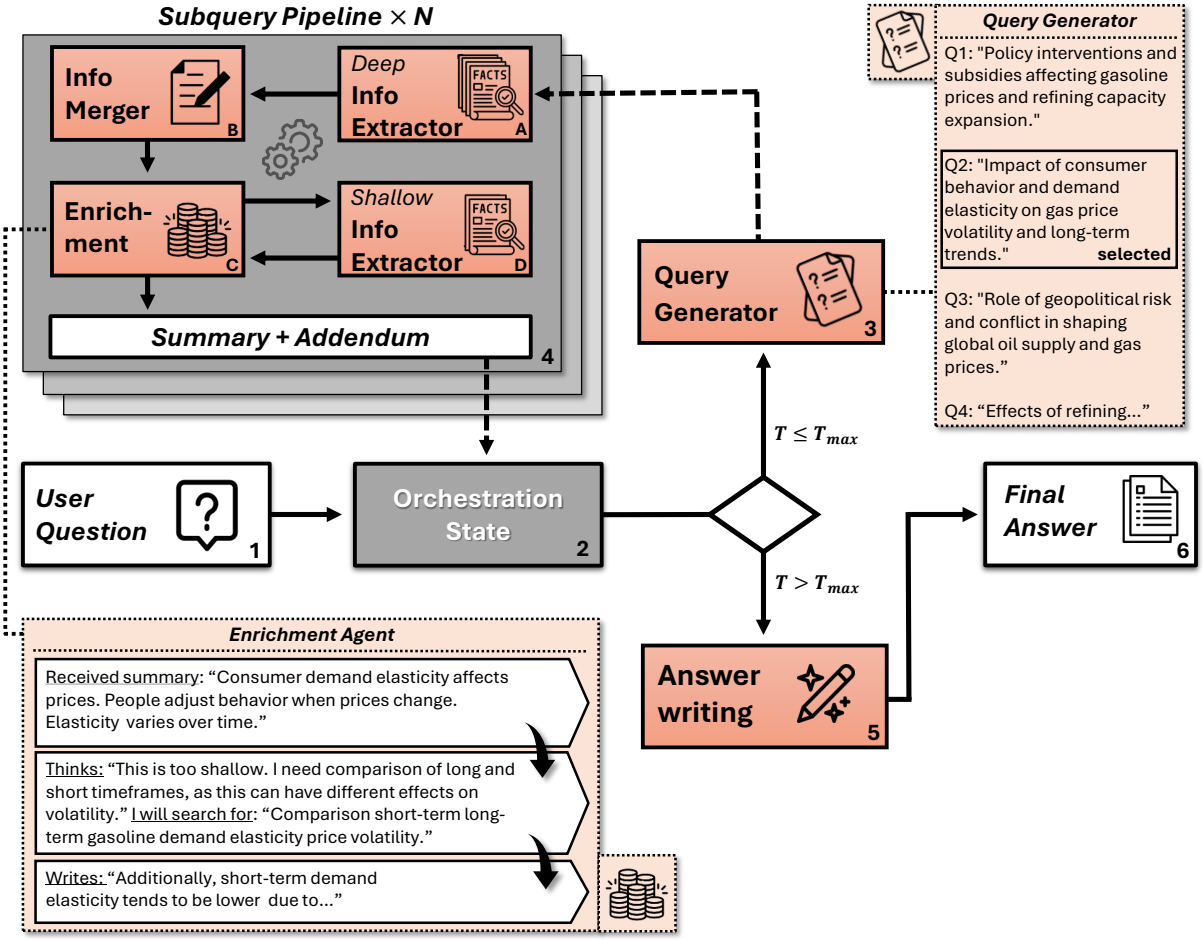

Figure 4.1: HERO system architecture showing multi-turn research flow with parallel subquery pipelines. Numbered steps (1-6) indicate execution sequence; letters (A-D) denote pipeline stages.

4.1.3 Parallel Execution

HERO executes multiple subquery pipelines simultaneously for each research turn, with the number determined by submodular query selection. The centralized OrchestrationState coordinates these parallel operations while maintaining consistency through atomic state updates.

4.1.4 Query Generator

The Query Generator (step 3 in Figure 4.1) initiates each research turn by decomposing the user question into diverse, targeted research queries. Its inputs vary by research round: in the first round the query generator only receives the user question. On subsequent rounds it will also receive previous subqueries and (enriched) summaries from previous turns.

The agent generates a candidate pool of potential queries. This over-generation provides the submodular optimizer with sufficient diversity to select from. Each candidate includes both the query text and search instructions that guide the Information Extractor

Through submodular optimization (Section 3.5) a set of subqueries is selected. Importantly, the generated subqueries decompose the user question rather than attempting to answer it directly. The user question is addressed only through synthesis of subquery results.

4.2 Subquery Pipeline

4.2.1 Information Extraction

Each subquery executes through a four-stage pipeline (steps A→D in Figure 4.1). The Information Extractor performs both deep (step A) and shallow (step D) retrieval.

- **Deep Extraction**: Deep search for the subquery, retrieving broadly relevant documents
- **Shallow Extraction**: Targeted retrieval for enrichment queries (Section 4.2.3), focusing on specific gaps identified during the enrichment phase

Both strategies use the same extraction agent but differ in search scope and instruction specificity. The distinction enables efficient resource allocation—deep searches explore broadly while shallow searches fill targeted gaps with more precise search terms and instructions. The agent can be considered a reranker. It selects only relevant excerpts, claims, or factoids with accompanying source, that are directly relevant to its subquery and instructions. Finally, there is early stopping built in, where the agent stops searching if fewer than 10% of documents in a retrieval batch yield new citations compared to previously found total sources. We allow one additional low-yield batch before terminating.

4.2.2 Information Merger

The Information Merger (step B) synthesizes all extracted information into an intermediate research summary for the subquery.

The agent receives the subquery and the list of extracted information items from the Information Extractor. It then connects the loose facts and excerpts into an information-dense, coherent paragraph that addresses the subquery. Citations are deduplicated and the agent has the purview to remove ones it deems redundant.

This intermediate summary serves two purposes: (1) it provides structured knowledge for the Enrichment Agent to analyze, and (2) it will be used by the answer writer.

4.2.3 Enrichment

The Enrichment Agent (step C) analyzes merged summaries to identify unexplored aspects, contradictions, or areas with insufficient evidence.

The agent receives: 1) The original user question, 2) the specific subquery focus, 3) the initial research summary from the Information Merger and, 4) the set of other subqueries which have been investigated

This input structure enables the agent to understand the current research context and explored directions without overloading its context window with detailed information from other pipelines. Importantly, the agent does *not* see summaries from other subqueries—it only knows their topics, preserving hierarchical information isolation.

Based on its analysis, the agent may generate enrichment queries to address identified gaps. If multiple enrichment queries are proposed, submodular optimization selects a diverse subset to prevent redundant searches. These enrichment queries are delegated to the Information Extractor for shallow, targeted retrieval.

The agent then synthesizes the newly retrieved information into an enrichment paragraph and appends it to the original summary, which we will call the enriched summary.

4.3 Answer Synthesis

4.3.1 Answer Writer Module

The Answer Writer (step 5) implements final report generation. This agent synthesizes accumulated research into comprehensive, well-structured answers that directly address the user question.

Unlike the isolated subquery pipelines, the Answer Writer is the only component that sees all enriched summaries. The agent performs synthesis, identifying connections, resolving contradictions, and organizing information into a coherent narrative structure, which answers the user question firstly, and then paints the complete picture.

Chapter 5

Experimental Setup

5.1 Datasets and Metrics

Since our model is designed to be a general application DR-agent, the benchmark suite takes two angles. We analyze its performance both in a general setting, by looking at typical internet questions and using a controlled snapshot-of-the-internet environment [11]. Simultaneously, for academic research purposes, we use a large academic corpus [62] and an accompanying suite of queries for different domains.

Table 5.1 presents our benchmark composition, comprising 458 queries across eight distinct evaluation datasets. To ensure fair comparison with published baselines, we maintain the original evaluation protocol of each respective paper with the note that the similarly named metrics are not comparable between benchmarks(see Appendix A.4.

5.1.1 DeepResearchGym

Our system is evaluated on DeepResearchGym [11] using the first 100 queries from the Researchy Questions dataset [55]. These queries, extracted from commercial search logs, represent complex, non-factoid information needs requiring multi-document synthesis, with users averaging 15.85 clicks across 6.31 unique documents per query.

Corpus

The ClueWeb22-B corpus [49], comprising 87 million English web documents is used. It was indexed via MiniCPM-Embedding-Light dense retrieval [24] with DiskANN [30] approximate nearest neighbor search. We utilize the API key generously provided by the [11] authors, who host the aforementioned setup.

Evaluation Metrics

We adopt the complete DeepResearchGym evaluation protocol [11], which employs `GPT-4.1-mini` [47] as judge (validated with $\kappa = 0.87$ inter-annotator agreement). We organize our presentation of these metrics according to the three dimensions of our completeness framework.

Coverage. *Key Point Recall (KPR)* measures the proportion of ground-truth key points, extracted from pages clicked by users in real search sessions—that are substantiated by

the report: KPR $= \frac{1}{M}\sum_{j=1}^{M} c_j$, where $c_j = 1$ if key point j is supported. Three qualitative metrics assessed on 0–10 scales complement this: *Depth* (comprehensiveness of analysis), *Breadth* (range of subtopics covered), and *Balance* (consideration of multiple perspectives).

Grounding. *Key Point Contradiction (KPC)* quantifies misinformation as the proportion of ground-truth key points the report contradicts: KPC $= \frac{1}{M}\sum_{j=1}^{M} d_j$, where $d_j = 1$ if the report contradicts key point j. Attribution quality is evaluated through *Citation Recall* (proportion of factual claims with at least one citation, $\frac{N_{\text{cited}}}{N_{\text{total}}}$) and *Citation Precision*[1] (average support quality across cited claims, $\frac{1}{N_{\text{cited}}}\sum_{i=1}^{N_{\text{cited}}} s_i$, where $s_i \in \{0, 0.5, 1\}$ reflects no, partial, or full support). The *Support* metric (0–10 scale) assesses overall evidence quality and source credibility.[2]

Presentation Quality. Two 0–10 scale metrics evaluate report quality: *Clarity* (structural organization, logical flow, and linguistic fluency) and *Insightfulness* (synthesis quality and analytical nuance).

Dataset	Task Format	Discipline	N	Metrics
peS2o Corpus				
SciFact	Binary	Biomed	50	Corr, Cite
PubMedQA	Binary	Biomed	50	Corr, Cite
QASA	Short	CS	50	Corr, Cite
ScholarQA-CS	Long	CS	50	Corr, Cite
ScholarQA-BIO	Long	Biomed	50	Cite
ScholarQA-NEURO	Long	Neuro	50	Cite
ScholarQA-MULTI	Long	Mixed	50	Cite, LLM-5
ClueWeb22-B Corpus				
Researchy Questions	Long	General	100	KPR, Cite, LLM-10
Total			**508**	

Table 5.1: Full Benchmark Suite composition. Metrics: Corr (Correctness) measured via accuracy for binary tasks or ROUGE-L/rubric scores for long-form tasks; Cite (Citation quality) measured via precision, recall, and F1; KPR (Key Point Recall); LLM-5 evaluates relevance, coverage, and organization on a 5-point-Likert scale; LLM-10 evaluates six quality dimensions (clarity, depth, balance, breadth, support, insightfulness) on a 10 point scale.

5.1.2 ScholarQABench

We evaluate our system on ScholarQABench, the comprehensive benchmark suite established in OpenScholar [5], which encompasses both single-paper and multi-paper synthesis tasks across multiple scientific disciplines. We use every dataset to get the full coverage as is proposed in OpenScholar. Due to computational cost and initial saturation, we used the first 50 queries of each dataset.

[1]The static corpus lacks URL-based retrieval, defaulting to live web crawling, which resulted in 40% miss rate. Manual corpus search recovered 99.3% of cited documents in their snapshot state.

[2]We adjusted the claim-extraction prompt to account for our agent's citation format, which places citations at sentence-end rather than inline with specific claims.

Datasets

The suite comprises seven datasets with varying complexity levels. SciFact [65] contains biomedical claims requiring True/False verification. PubMedQA [33] provides yes/no questions about clinical research findings. QASA [37] focuses on detailed single-paper understanding in AI or Machine Learning, and expects ∼50 word answers.

The multi-paper ScholarQA tasks demand synthesis across multiple sources. ScholarQA-CS features computer science related queries with general and expert-annotated rubrics. The answers expected have a relative length penalty losing 5% on answers longer than 600 words. ScholarQA-BIO and ScholarQA-NEURO comprise a large set of expert-generated literature review questions in biomedicine and neuroscience. ScholarQA-MULTI represents the most comprehensive task, with queries and expert-written answers across computer science, physics, and biomedicine, each answer requiring approximately 56 minutes of expert annotation time and explicit prohibition of LLM assistance during creation.

Corpus and Retrieval

We utilize the peS2o v3 dataset [62], comprising 45 million open-access scientific papers derived from the Semantic Scholar Open Research Corpus [41]. This corpus spans publications through October 2024 across all major scientific disciplines. We optimize OpenScholar's approach slightly, filtering short shards and using a slightly improved chunking distribution algorithm, which results in approximately 253 million chunks with a mean length of 218.15 and a standard deviation of 65.10.

For retrieval, we reuse OpenScholar's approach of encoding these passages using a Contriever bi-encoder that was continually pre-trained on the peS2o corpus [29, 62]. The dataset was hosted on the Snellius HPC cluster [64] and was queried with a setup adapted from the OpenScholar retrieval configuration [5]. We did not include the partial web search that OpenScholar used, to keep the data environment more controlled.

Evaluation Metrics

We follow the complete ScholarQABench evaluation protocol [5], but organize metrics by completeness.

Coverage. *Correctness* varies by task type: accuracy for binary classification (SciFact, PubMedQA), and semantic similarity via ROUGE-L and BERTscore [7] for short-form generation (QASA). For ScholarQA-CS, the expert-annotated *Correctness (Rubric Score)* evaluates the presence of key ingredients (60% weight) alongside fixed criteria (40% weight). For ScholarQA-MULTI, the `Prometheus V2` LLM-as-judge [36] evaluates *Content Quality* on a five-point Likert scale, with relevance and coverage criteria directly measuring this dimension.

Grounding. We compute *Citation Recall* ($R_c = \frac{N_{\text{supp}}}{N_{\text{claims}}}$) as the proportion of sentences that are verifiably supported by their citations, determined by a `Flan-T5-XL` NLI model [10, 72]. [3] *Citation Precision* ($P_c = \frac{N_{\text{nec}}}{N_{\text{cites}}}$) measures the proportion of necessary citations via leave-one-out testing; a citation is deemed necessary if its removal causes the claim to become unsupported (see Chapter E for explanation)

[3] Following [5], only sentences exceeding 50 characters are evaluated, and we truncate to a maximum of three citations per sentence.

Presentation Quality. For ScholarQA-MULTI, the `Prometheus V2` model evaluates report *organization* on a five-point Likert scale as part of its Content Quality assessment.[4]

5.1.3 Baseline Models

For both benchmarks, we reuse the values reported in DeepResearchGym [74] and OpenScholar [5], with the caveat that our sample size is smaller for all datasets.

5.1.4 HERO Configuration

Configuration parameters were adjusted based on task complexity and expected answer length (Table 5.3). For long-form research questions (DeepResearchGym and ScholarQA-MULTI), we used the maximum parameter values that the Answer Writer can handle while maintaining quality and citation accuracy. For shorter, more focused question types, we reduced parameters to prevent unnecessary computational cost. ScholarQA-CS, which expects intermediate-length answers, used settings between these extremes.

Dataset characteristics influenced search depth configuration. PES2O's smaller, fragmented chunks (academic paper sections) required deeper retrieval compared to FineWeb's more holistic web pages. This resulted in higher token usage for PES2O runs, as shown in Table 5.2.

Table 5.2: Average token usage per agent by retriever type (thousands of tokens).

Agent	Model	FineWeb	PES2O
Query Generator	gpt-4.1-mini	1.6k	3.2k
Info Extractor	gpt-4o-mini	2.2k	7.3k
Info Merger	gpt-4.1-mini	2.1k	3.1k
Enrichment	gpt-4.1-mini	1.3k	1.8k
Answer Writer	gpt-5	9.4k	9.5k
Total per subquery		**17**	**44**

For submodular optimization, we set $\alpha = 0.6$ for main query generation to balance relevance and diversity. For enrichment queries, we increased α to 0.65 to prioritize staying close to the subquery focus and prevent topic drift. Both agents used a $k = 3$ candidate multiplier. These values were determined through preliminary experiments on the DeepResearchGym benchmark.

Table 5.3: Agent configuration across datasets.

Configuration	DeepResearchGym	ScholarQA-MULTI	ScholarQA-CS	Other Datasets[†]
Max Turns	3	3	2	2
Queries per Turn	3	3	3	2
Enrichments per Query	2	2	1	1
Deep Search Depth	24	80	80	80
Shallow Search Depth	8	40	40	40

[†] SciFact, PubMedQA, QASA, ScholarQA-BIO, ScholarQA-NEURO.

5.1.5 Ablation Study Design

To address RQ.2 regarding the individual contributions of submodular optimization and hierarchical enrichment, we conducted ablation experiments with three system configurations on a random sample of DeepResearchGym queries.

[4] Since `GPT-4-Turbo` is deprecated, we use `gpt-4.1-mini` [47] for ScholarQA-CS rubric evaluation.

- **HERO-Full**: Complete system with submodular query optimization ($\alpha = 0.6$, $\alpha' = 0.65$, $m = 4$ candidate multiplier) and hierarchical enrichment ($k' = 2$ enrichment queries per subquery)
- **HERO-NoEnrichment**: Submodular optimization retained, enrichment stage disabled ($k' = 0$)
- **HERO-NoSubmodular**: Hierarchical enrichment retained ($k' = 2$), queries selected via uniform random sampling from candidate pool instead of submodular optimization

All configurations used a single research turn ($T = 1$) with $k = 4$ subqueries selected per turn ($m = 4$ multiplier). This protocol captures the direct contribution of each component before the logarithmic improvement from multiple iterations would reduce measurable differences. We evaluate on 20 queries per configuration and report KPR as primary metric.

Chapter 6

Results

6.1 DeepResearchGym Results

HERO achieves the strongest performance on DeepResearchGym across coverage and presentation quality dimensions (Table 6.1). It shows strong grounding performance too, with the highest F1 score.

6.1.1 Coverage.

HERO achieves the highest KPR at 67.63, indicating it successfully identifies and incorporates the most relevant information from search results. Figure 6.1 shows HERO scores highly on three coverage-related quality dimensions: Breadth (9.80), Depth (9.41), and Balance (8.88), averaging 9.36 across these metrics. Its answers demonstrate appropriate complexity for user requests while covering most relevant points.

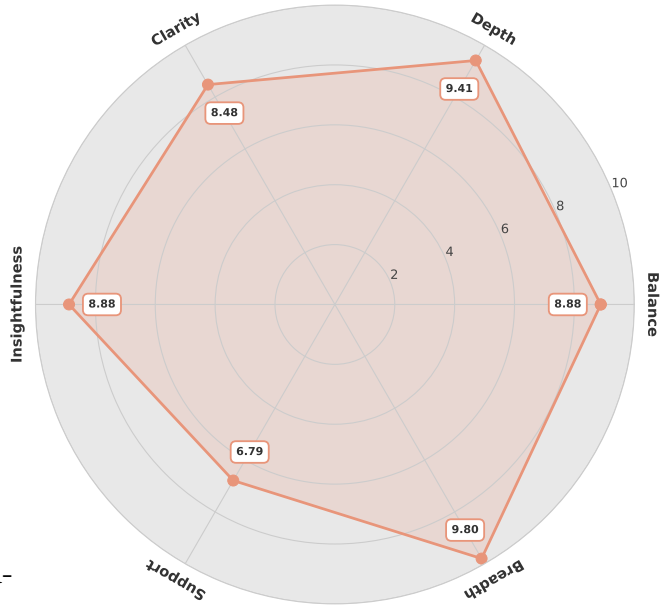

Figure 6.1: Report Quality performance across six dimensions (0–10 scale) on DeepResearchGym. Evaluated using GPT-4.1-mini as judge.

6.1.2 Grounding.

HERO demonstrates exceptional grounding with 99.35% citation recall, substantiating nearly all factual claims with proper attribution. This exceeds the already >90% rates of OpenDeepSearch (94.82%) and GPT-Researcher (90.82%). Recall paired with a 84.92 on precision, which is slightly below GPT-Researcher (85.36), and taken together, HERO has the highest F1 score at 91.57. Key Point Contradiction (KPC) shows a different signal. HERO's score of 2.34 is the worst among all baselines (lower is better), consistent with a trend for all models ($r = 0.72$, $n = 6$) where broader coverage is associated with a higher contradic-

Table 6.1: Performance comparison on DeepResearchGym [11] (first 100 queries). Key Point Recall (KPR) and Key Point Contradiction (KPC) measure report coverage and factual consistency with ground-truth key points. Precision, Recall, and F1 measure citation faithfulness. Clarity and Insightfulness assess report quality on a 10-point scale. All metrics evaluated using GPT-4.1-mini as judge. Abbreviations: Cov. = Coverage, Prec. = Precision, Rec. = Recall, Qual. = Presentation Quality, Clar. = Clarity, Ins. = Insightfulness. *Systems not tailored for long-report generation.

	Cov.		**Grounding**			**Qual.**	
System	KPR	KPC	Rec.	Prec.	F1	Clar.	Ins.
HERO	**67.63**	2.34	**99.35**	84.92	**91.57**	**84.8**	**88.8**
GPT-Researcher [17]	64.67	1.42	90.82	**85.36**	87.96	83.70	78.01
OpenDeepSearch [4]	42.81	0.84	94.82	81.32	87.64	61.48	49.51
HuggingFace-DeepSearch [28]	35.22	1.35	0.10	0.10	0.20	58.34	52.36
Search-o1* [40]	29.93	**0.38**	–	–	–	30.31	37.87
Search-R1* [32]	4.95	0.80	–	–	–	9.07	11.18

tion rate. On the Radar chart 6.1 support score is 6.79, although there is no comparison, since we don't have the other baseline model scores (unreported in DeepResearchGym). Manual inspection confirms the low Support stems from HERO citing sources the judge deemed insufficiently authoritative, an expected limitation given HERO does not filter for source quality during retrieval.

6.1.3 Presentation Quality.

HERO achieves the highest scores on rubric-based presentation quality metrics (Table 6.1). Figure 6.1 shows HERO leads in both Clarity (8.48) and Insightfulness (8.88), with a strong margin (+10.79 percentage points) on Insightfulness compared to the next-best baseline, indicating well-structured and analytically rich reports.

6.2 ScholarQABench Results

HERO achieves exceptional performance on scholarQABench, on coverage and presentation quality, while it has mixed results on grounding.

6.2.1 Coverage.

HERO achieves perfect accuracy (100%) on both PubMedQA and SciFact single-paper classification tasks. QASA shows lower ROUGE-L scores (14.2). It is related to HERO's choice of different terminology, as the complementary BERTScore (86.4) indicates high similarity to the expected answer. On ScholarQA-CS, HERO achieves a rubric score of 68.9, outperforming the strongest baseline (OS-GPT4o at 57.7) by 11.2 percentage points. The LLM-5 score exceeds the average of all baselines by 30.92 percentage points. HERO scores near the metric ceiling (Table: 6.2, with HERO consistently scoring near-maximum on organization and relevance dimensions. Occasional lower coverage scores occur when

HERO provides different conceptual depth or examples than expected by ground truth, comparable to omitted keypoints in DeepResearchGym.

6.2.2 Grounding.

Citation F1 scores are comparatively middle of the pack across most single-paper datasets while high across most multi-paper datasets (Appendix Table C.1 shows the recall and precision independently).

These citation patterns differ from DeepResearchGym due to fundamental differences in how the benchmarks operationalize citation metrics (see Appendix A.4).

Table 6.2: LLM-5 metric breakdown for HERO on ScholarQABench multi-paper datasets.

Dimension	Score
Coverage	4.88
Organization	5.0
Relevance	4.96
Average (LLM-5)	**4.95**

We verified potential data contamination after observing high single-paper correctness performance. While Openscholar [5] reportedly filtered contaminated contexts, we identified 22 leaked chunks in SciFact, though only 5 appeared in outputs, excluding them reduced F1 with 0.54 percentage points)

6.2.3 Presentation Quality.

The organization dimension directly assesses Presentation Quality and HERO achieves a perfect score (5.0) 6.2.

6.3 Ablation Study Results

To isolate the individual contributions of submodular optimization and hierarchical enrichment to coverage performance, we evaluated three system configurations on DeepResearchGym using a single research turn ($T = 1$, $k = 4$ subqueries). This simplified protocol captures direct component effects before logarithmic improvements from multiple iterations would attenuate the signal.

Table 6.4 reveals that both components contribute substantially and approximately equally to coverage performance. Removing enrichment decreases KPR by 3.6 percentage points, while removing submodular optimization decreases KPR by 5.3 percentage points.

	Single-paper Datasets						Multi-paper Datasets					
	Pub		Sci		QASA		CS		Multi		Bio	Neu
Model	Corr	Cite	Corr	Cite	Corr	Cite	Corr	Cite	LLM-5	Cite	Cite	Cite
HERO	**100.0**	71.9	**100.0**	55.2	14.2 (86.4)	55.4	**68.9**	47.3	**4.95**	**60.5**	**66.9**	**72.6**
Llama3-8B	61.5	0.0	66.8	0.0	14.3	0.0	41.9	0.0	3.79	0.0	0.0	0.0
+OSDS	75.2	63.9	75.5	36.2	18.6	47.2	46.7	26.1	4.22	25.3	38.0	36.8
OS-8B	76.4	68.9	76.0	43.6	23.0	56.3	51.1	47.9	4.12	42.8	50.8	56.8
Llama3-70B	69.5	0.0	76.9	0.0	13.7	0.0	44.9	0.0	3.82	0.0	0.0	0.0
+OSDS	77.4	71.1	78.2	42.5	22.7	63.6	48.5	24.5	4.24	41.4	53.8	58.1
OS-70B	79.6	74.0	82.1	47.5	**23.4**	**64.2**	52.5	45.9	4.03	54.7	55.9	63.1
GPT4o	65.8	0.0	77.8	0.0	21.2	0.0	45.0	0.1	4.01	0.7	0.2	0.1
+OSDS	75.1	73.7	79.3	47.9	18.3	53.6	52.4	31.1	4.03	31.5	36.3	21.9
OS-GPT4o	74.8	**77.1**	81.3	**56.5**	18.7	60.4	57.7	39.5	4.51	37.5	51.5	43.5
PaperQA2	–	–	–	–	–	–	45.6	**48.0**	3.82	47.2	56.7	56.0
Perplexity	–	–	–	–	–	–	40.0	–	4.15	–	–	–

Note: Corr measures Coverage for CS (presence of key ingredients) and Grounding for Pub/Sci/QASA (factual accuracy). Cite measures Grounding (Citation F1). LLM-5 combines Coverage and Presentation Quality dimensions.

Table 6.3: Performance comparison on ScholarQABench [5] with 50 samples per dataset. Pub and Sci stand for PubMedQA and Scifact, while CS, Multi, Bio, and Neu indicate ScholarQA-CS, ScholarQA-MULTI, ScholarQA-BIO, and ScholarQA-NEURO, respectively. Corr indicates correctness metrics (accuracy for PubMedQA and SciFact, ROUGE-L for QASA with BERTScore using SciBERT shown in brackets, overall rubric score for ScholarQA-CS). Cite indicates citation F1. LLM-5 indicates the average score of Cov (coverage), Org (organization), Rel (relevance) as evaluated by Prometheus V2.

Table 6.4: Ablation study results on DeepResearchGym (N=20 queries, mean ± SD). Key Point Recall (KPR) measures the proportion of ground-truth key points substantiated by the report. All metrics evaluated using GPT-4.1-mini as judge. Configurations test individual component contributions: HERO-Full includes both submodular optimization and hierarchical enrichment; NoEnrichment disables enrichment ($k' = 0$); NoSubmodular uses random query selection instead of submodular optimization.

Configuration	**KPR**
HERO-Full	**66.0 ± 16.7**
NoEnrichment	62.4 ± 21.0
NoSubmodular	60.7 ± 21.1

Chapter 7

Discussion

7.1 Overview

This thesis investigated completeness in DR agents and attempted to improve them through HERO, a model optimized with submodular optimization 2.4.1 and an Enrichment stage 4.2.3, in order to achieve the best completeness results. Here, completeness consists of coverage, grounding, and presentation quality. We set out to find answers to RQ How to improve completeness in DR agents, with sub-research questions: firstly RQ.1, which asks whether the proposed HERO architecture improve completeness in generated research reports compared to existing baseline DR agents? Secondly RQ.2, which questions what the individual contributions of submodular query optimization and enrichment are, to the observed improvements in coverage?

7.2 System Effectiveness

To answer RQ.1, we evaluate across all three dimensions to address the completeness evaluation of HERO.

7.2.1 Coverage Performance

HERO achieves the highest coverage performance across both benchmarks (Table 6.1, Table 6.3). It achieves the highest KPR, and complementarily, rates 9+ on breadth, depth, and balance dimensions (Figure 6.1). On ScholarQABench, HERO substantially outperforms all baselines on the rubric score, with near-ceiling on coverage and relevance (Table 6.2).

This demonstrates that HERO's architecture effectively retrieves and integrates diverse information streams to produce comprehensive research reports, outperforming all baselines as reported in [5, 11].

Limitations and Future Work

Underlying LLM We utilize a mixture of current OpenAI models [47], which likely improved over baselines in data processing and synthesizing capabilities. A full analysis would require an updated run of the compared models with their current versions. On the other hand, the architecture of the current model could have been set up comparably, with weaker LLMs and more hierarchical layers, or more surgical, parallelized

tasks. It would even be interesting and a valid approach to fully leverage the control, submodular optimization offers, and use the current setup with many more lightweight query-pipelines. In a way, spawning an army of agentic workflows that all research a diverse separate direction.

7.2.2 Grounding Performance

HERO leads on DeepResearchGym's recall and second on precision, while scoring middle of the road on the ScholarQABench academic corpus citation metrics. HERO has exceptionally high recall on both datasets, whereas the high score on DeepResearchGym is likely inflated, since recall doesn't verify the veracity of the grounding, solely the presence. The ScholarQABench is likely a better proxy. On these the average F1 citations performance is in sharp contrast to the near perfect accuracy on the single-paper datasets. What seems to happen is that HERO processes vast amounts of knowledge, so it learns the answer in its intermediate summaries, but it chooses to include different chunks as citations in the final answer, opting for more general descriptive chunks, then specific ones that answer the query ultimately. A task mismatch, as the multi-paper datasets, which are the type of output HERO is build for, show better performing results.

Limitations and Future Work

Citation Verification Module It could be argued that HERO could be improved by doing a citation-verification step. OpenScholar [5] has a simple implementation, where when a claim is unsubstantiated they do a quick search for verification, but a more current one is proposed in e.g. [52]. For our hierarchical setup, it would be an easy addition and likely best to append as a post-processing step, as the several stages of information merging could lead to claims becoming malformed and losing their veracity after all. It would only work on 'simple' citations, as synthesized points would require another deeper research.

Key Point Contradiction scores Regrettably, HERO's KPC is worst of all baselines. However, we would like to caveat it mostly happened in philosophically or sociologically themed questions, which can be contradictory in nature to begin with. It seems to be an artifact of HERO's prompting, which is declarative, combined with the peculiarities of this benchmark.

To investigate this pattern, we categorized all keypoints in relation to their queries as being *pro* (supporting the angle of the query), *neutral* (not leaning either way), or *anti* (opposing the framing of the question). Figure 7.1 shows the contradiction rate for these categories. When HERO encounters keypoints that oppose or criticize a topic, it is more than twice as likely to contradict them compared to supportive or factual arguments (3.42 vs. 1.53 percentage points).

For example, on death penalty queries, this manifests as stance-flipping. When asked the neutral question 'should the death penalty be legal? (See Appendix **??** for all full reports mentioned)', HERO's answer aligned with arguments against the death penalty (supporting all 7 anti-keypoints while contradicting 4 of 8 pro-keypoints). However, when asked the PRO-framed question 'why should the death penalty be allowed?', HERO reversed its position, now aligning with pro-death-penalty arguments (supporting 10 of 12 pro-keypoints while contradicting 2 of 4 anti-keypoints). This demonstrates that HERO

adapts its argumentative position to match query framing [46, 66], a sycophantic bias common in LLMs [58] that persists despite the enrichment agent's explicit instructions to find contradictory evidence.

Other instances where our model fails are: 'is the criminal justice system fair?' (KPR = 31%, KPC = 69%). HERO focuses on global, instead of the expected US-centric answers, or 'what role did Indians play in the wars for empire?' (KPR = 20%, KPC = 20%), where our model completely focuses on Asian Indian people's influence in the British Empire, which is a misinterpretation of the intended question on Native American participation in the US Army.

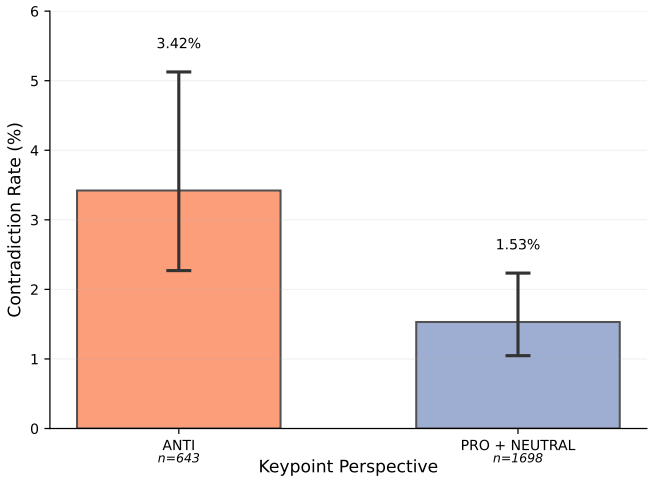

Figure 7.1: Contradiction rates by keypoint perspective. Keypoints opposing the query framing (ANTI) are contradicted at 3.42% compared to 1.53% for supportive and neutral keypoints combined (PRO + NEUTRAL).

7.2.3 Presentation Quality Performance

HERO is the best model for Presentation quality in academic and general setting. It gives the most complete, balanced and insightful answers in both benchmarks on all relevant metrics.

Limitations and Future Work

Answer Length Arguably, these metrics are not painting a full picture. Both clarity and insight, of DeepResearchGym, and organization and coverage of ScholarQABench are rubric rated items. The latter also have a high quality ground truth example, which is posited to be make the LLM response more reliable [5]. While there is ample evidence that supports the qualitative validity [21], there is similar evidence for a length bias: 'longer answers are rated higher [76].' We are not in posession of the generated reports of the baseline models, but we could venture the guess that our model, which has a focus on being more complete, has longer outputs. It's not 1-to-1 certain whether this is actually necessarily what is valued by users. One could imagine the addition of explanatory graphs or tables, would be a great addition to make it more appealing for users. A great case study is [68], where they initially plan the type of plot that is sensible for the research and let the agent fill it in. In general it seems DR agents will steer more towards multi-modality as a core feature [26].

7.3 Algorithmic Contributions

RQ.2 asked what submodular query optimization and hierarchical enrichment each contribute to coverage improvements.

The ablation study (Section 6.3) shows both components improve coverage. However, two caveats apply. First, the NoEnrichment configuration does less computa-

tional work by design, so this measures effectiveness rather than efficiency. Second, the single-turn protocol may favor breadth (submodular optimization) over depth (enrichment)—multiple research rounds might show different patterns. For example in Appendix C.2, the effect of the enrichment phase is shown on 9-subquery runs.

Despite these limitations, the results clearly demonstrate that both components contribute to HERO's coverage performance. Diverse query selection establishes broad topical coverage, while enrichment adds targeted depth by identifying and addressing specific gaps. Neither component alone achieves the performance of the full system.

Limitations and Future Work

The current study did not investigate the full range of possibilities. Hyperparameter tuning, or using different searches with different α's would be interesting to see the actual measurable effect in a noisy field as IR.

7.4 Conclusion

This thesis asked how to improve completeness in deep research agents. We demonstrated that hierarchical architectures combining submodular query optimization with multi-stage enrichment systematically improve all three dimensions of completeness—coverage, grounding, and presentation quality. HERO achieves state-of-the-art performance on both benchmarks, with the highest coverage metrics and superior presentation quality, providing empirical evidence that principled query diversification paired with targeted enrichment successfully addresses the completeness challenge. Simultaneously, we show that each module is essential. Taken together, we can safely assert we have answered the main RQ. HERO's architecture and design choices lead to a complete answer.

This work makes three primary contributions. First, HERO as a concrete system demonstrating that formal optimization methods can be effectively integrated into agentic research workflows. Second, a multi-dimensional evaluation framework that assesses completeness through coverage, grounding, and presentation quality—a more comprehensive approach than single-metric evaluation. Third, a benchmark suite combining DeepResearchGym and ScholarQABench that enables cross-domain validation across both general and academic research contexts, addressing the field's need for more rigorous evaluation standards for deep research agents.

As deep research agents grow into more multi-modal heavy use agents, the challenges identified will be at the forefront of development. HERO offers a promising addition for building reliable, comprehensive research assistants that balance breadth with accuracy.

Appendix A

Exploratory Empirical Problem Analysis

A.1 Introduction

Initially, to pinpoint clearly where the baseline model was failing, we performed a qualitative customer study. Zeta-Alpha had requested and received a set of evaluations from one of their clients in chemistry research. They deployed our baseline model agent, connected to their internal database. We analyzed these, categorized the issues, and used this as a foundation for our understanding of how to improve the agent system.

A.2 Methodology

Twenty evaluators were asked to give a set of 'typical' questions they would use a deep research agent for in their internal database. We then generated the reports for these questions and asked the clients to freely evaluate them. Four of the clients did this, and we used their evaluations to find out where the reported issues came from. Since we only had the final reports of the customers and not the trace the agent used to get to its final system, we had to rerun the reports again manually, up to a maximum of three times, in order to recreate the issue as the user saw it in their report. The repetition was sometimes necessary due to natural stochasticity in outcomes of Large Language Models (LLMs), especially multi-agent systems. Certain issues therefore fall into several categories if it was hard to establish the root cause of the evaluated issue (e.g., a false claim in the report can arise in several of the steps the agent makes in the agent system).

After internal deliberation, we decided to categorize the issues into four categories, in line with [73]:

1. Relevant information not retrieved
2. Relevant information retrieved but ignored by LLM
3. Irrelevant information retrieved and included
4. Report structure issues

Unfortunately, due to client privacy, the actual qualitative analysis cannot be shared. For inquiries about the underlying analysis, please contact Zeta-Alpha.

A.3 Results and Discussion

Category 3, *irrelevant information retrieved and included*, is by far the most occurring category and is associated with precision in the traditional information retrieval (IR) sense [24]. The complicated nature of chemistry reports seems to lead to the agent getting confused at several potential locations in the pipeline. Often when this occurred, it was also due to data integrity issues. Multiple chunks get retrieved and analyzed in batches, which led to the agent sometimes recognizing chunks as being related, which is what led to issues. For example, a relevant 'introduction' chunk and a wrongly retrieved irrelevant chunk would be retrieved and concatenated, and it would throw off the information extractor to see the irrelevant chunk as relevant.

There are also a couple of cases of category 4, which often resulted from the final answer agent writing a badly structured report on the source material, or cases where users asked a very specific question that the agent wasn't necessarily prompted well for, e.g., a client asked for a list of contact info, and the agent wrote a full report on those people.

Finally, the relevant info not retrieved or the relevant info retrieved but lost in the pipeline was not too prevalent, with only nine cases between them. Trace analysis showed that this was often due to the agent not asking the right subquery, which led it to fixate on those results, a common issue in RAG systems [18]. An example is where the agent didn't really pay heed to the user's word choice of 'quantify,' such that the agent just gave a qualitative answer instead of a quantitative answer. Finally, often the agents just did not propagate the relevant information upstream because they deemed it irrelevant. However, no clear pattern emerged as to why this happened.

Table A.1: Total Frequency of Error Categories with Ambiguity Distribution

Error Category	**Total Frequency**[a]
1: Relevant info not retrieved	3
2: Relevant info retrieved but ignored	2
3: Irrelevant info retrieved and included	23
4: Report structure issues	8
1 or 2: Ambiguous between retrieval/ignorance[b]	4

[a] Frequencies are tallied from 35 distinct issues. Issues with an ambiguous root cause are counted in each potential category they could belong to. Therefore, the sum of frequencies (40) is greater than the total number of unique issues.

[b] This category represents cases where, due to the inability to reconstruct the original agent trace, it was impossible to determine if relevant information was never retrieved (Category 1) or retrieved but subsequently ignored (Category 2).

[c] Note removed as it was referenced but not defined in the original.

A.4 Implications for System Design

Due to the omnipresence of category 3, for our final design, we performed an explorative search depth analysis, where we looked at the effectiveness of searching depth versus actual useful retrieved information. Although HERO already has early stopping built in if the next batch of chunks it visits don't actually contain relevant information, this system is too simple, as even though the Information Extractor can deem a chunk relevant, that chunk can still disappear higher up the chain due to it being less relevant or duplicated with other information. Besides, we elected to reduce the number of chunks visited in a batch, as the amount was probably too heavy for the underlying `gpt-4o-mini` agent to properly assess.

A substantial limitation from this qualitative study is the completeness aspect; only two users had the domain knowledge and gave clear indication the agent system omitted information that should have been in the report.

Appendix B

Cross-Benchmark Citation Metric Comparison

Metric	DeepResearchGym	ScholarQABench
Citation Recall		
Definition	Proportion of factual claims with ≥1 citation	Proportion of sentences with NLI-verified support
Scope	LLM-extracted factual claims only	All sentences exceeding 50 characters
Validation	Presence check (binary)	NLI model verification (entailment)
Citation Precision		
Definition	Average support quality across all citations	Proportion of citations that are necessary
Scoring	Graduated (0, 0.5, 1) based on LLM judgment	Binary (necessary vs. redundant)
Method	Direct assessment of support quality	Leave-one-out ablation testing
Philosophy	Encourages comprehensive attribution	Encourages parsimonious attribution

Table B.1: Operationalization of citation metrics across benchmarks. These differences reflect distinct evaluation philosophies rather than inconsistencies, and make direct cross-benchmark comparison invalid.

While both DeepResearchGym and ScholarQABench report citation recall and precision, these terms operationalize fundamentally different constructs. Table B.2 summarizes the key methodological differences.

DeepResearchGym's citation recall measures the *presence* of citations for extracted factual claims, while precision evaluates support quality with graduated scores (0, 0.5, 1) averaged across all claim-citation pairs. In contrast, ScholarQABench's recall requires NLI-verified support for all sentences exceeding 50 characters, while precision identifies necessary citations through leave-one-out ablation testing.

These methodological differences preclude direct numerical comparison across benchmarks. However, they provide complementary perspectives on citation quality: DeepResearchGym incentivizes comprehensive attribution (rewarding thoroughness), while ScholarQABench rewards parsimonious citation (penalizing redundancy). Additionally, the

scope differs—DeepResearchGym evaluates only LLM-identified factual claims, whereas ScholarQABench evaluates all sentences above a character threshold, which may systematically lower recall scores in the latter. In comparison, Recall of ScholarQABench is more similar to precision in DeepResearchGym, as both metrics match the claim to the citation.

Table B.2: Operationalization of citation metrics across benchmarks. These differences reflect distinct evaluation philosophies rather than inconsistencies, and make direct cross-benchmark comparison invalid.

Metric	**DeepResearchGym**	**ScholarQABench**
Citation Recall		
Definition	Proportion of factual claims with $\geq$1 citation	Proportion of sentences with NLI-verified support
Scope	LLM-extracted factual claims only	All sentences exceeding 50 characters
Validation	Presence check (binary)	NLI model verification (entailment)
Citation Precision		
Definition	Average support quality across all citations	Proportion of citations that are necessary
Scoring	Graduated (0, 0.5, 1) based on LLM judgment	Binary (necessary vs. redundant)
Method	Direct assessment of support quality	Leave-one-out ablation testing
Philosophy	Encourages comprehensive attribution	Encourages parsimonious attribution

Appendix C

Additional Results

C.1 Citation Recall and Precision ScholarQABench

Table C.1: Citation recall and precision scores per ScholarQABench dataset.

Dataset	Recall	Precision
Single-paper datasets		
PubMedQA	74.00	70.00
SciFact	56.00	54.33
QASA	55.17	55.67
Single-paper average	*61.72*	*60.00*
Multi-paper datasets		
Neuro	74.92	70.47
Bio	68.91	64.99
Multi	63.25	57.89
CS	46.72	47.86
Multi-paper average	*63.45*	*60.30*
Overall Average	**62.59**	**60.15**

C.2 Enrichment Mechanism

Beyond quantitative ablation, examining full system runs (N=22, T=3 rounds, k=3 sub-queries per round, k'=2 enrichment queries) reveals *how* enrichment contributes to coverage. Figure C.1 shows the distribution of citations by pipeline stage across all queries.

This pattern confirms the enrichment mechanism's intended function: after initial broad retrieval establishes baseline coverage, the Enrichment Agent analyzes intermediate summaries to identify specific gaps or areas requiring deeper investigation, then generates targeted follow-up queries that surface additional relevant information. The substantial proportion of unique enrichment-sourced citations indicates this adaptive depth strategy successfully complements the breadth-first approach of diverse initial queries.

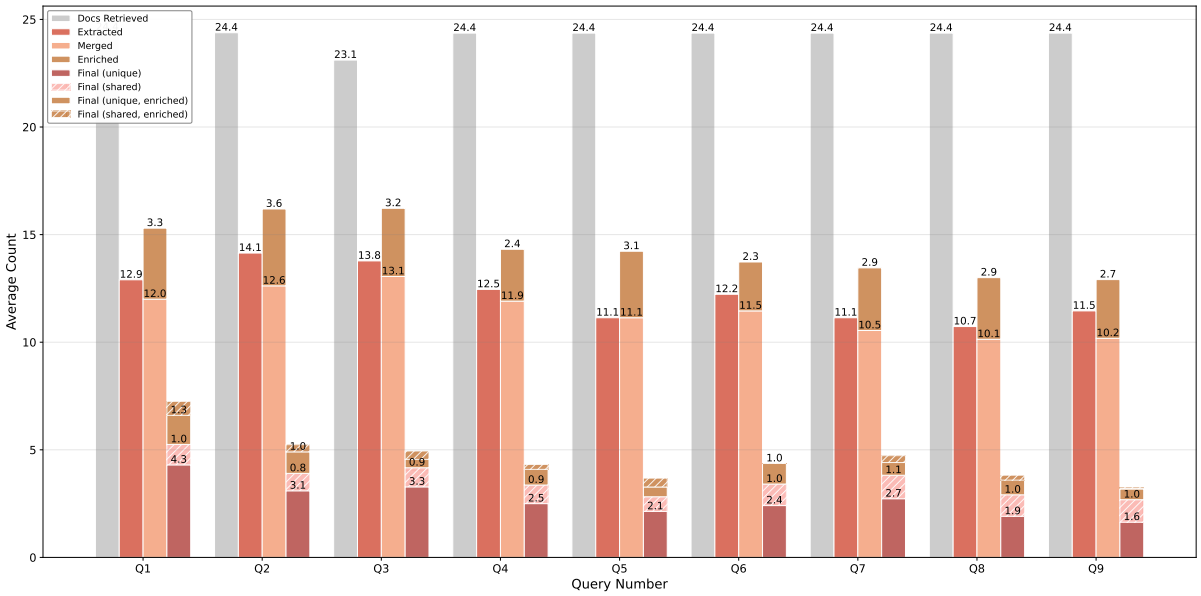

Figure C.1: Distribution of citation sources across pipeline stages in full HERO runs. The figure shows that enrichment queries identify information not captured by initial subquery searches, contributing unique citations that are important enough for the final report.

Appendix D

Full Reports

:

Listing D.1: Raw Appendix Content

]

```
Should the death penalty be legal?

No. The best available evidence shows the death penalty does not
    provide a unique public-safety benefit over long prison terms,
    while it carries irreversible risks of wrongful execution,
    imposes far higher financial and social costs than alternatives
    , and often harms---rather than helps---victims' families, all
    amid viable substitute punishments and global legal trends away
     from capital punishment [1]. Countries and U.S. states that
    abolished the death penalty have not experienced the predicted
    rise in murder, undermining its supposed deterrent value [2].
    The irreversibility of execution in a system that demonstrably
    convicts the innocent makes legalization ethically indefensible
     in practice [3].


Public safety and deterrence: no unique benefit from execution

Decades of research find the death penalty does not deter homicide
     better than life imprisonment, and overall crime levels do not
     significantly differ between places with capital punishment
    and those without it [1]. Comparative analyses show no
    meaningful crime reduction advantages in death-penalty
    jurisdictions, weakening the core deterrence claim [4]. Cross-
    state comparisons reinforce this: differences in homicide
    trends do not consistently track the presence or absence of
    capital punishment, challenging the premise that executions
    make communities safer [5].

Real-world outcomes after abolition show the sky does not fall:
    states that have ended capital punishment have not seen the
    forecasted murder spikes, and some evidence suggests police and
```

the public are safer in non-death states over long periods and many executions [2]. A study of more than three decades of FBI homicide data concluded that after 1,600 executions, safety outcomes were better in states without the death penalty, indicating no public-safety gain from maintaining executions [6]. Isolated historical counterexamples exist---such as Kansas's murder rate rising after abolition in the late 1960s and 70s and Alabama's rate remaining largely unchanged after reinstatement---but these outliers underscore that broader social and policy factors, not execution policy, drive violent crime trends [7].

Where abolition coincides with reforms in policing, sentencing, and rehabilitation, responses to violent crime tend to improve, suggesting the real levers of safety lie in system design rather than executions [8]. Because over half of state prison populations are incarcerated for violent offenses, right-sizing punishments and reallocating resources to focus on serious violence is central to reducing incarceration and improving safety after abolition [9]. Prioritizing serious crime through targeted policing and corrections can help mediate violent crime trends in the absence of capital punishment [10].

Justice, morality, and victims' needs

Proponents argue capital punishment deters heinous crime and demands "a life for a life," claiming it is a just response for the worst offenses and more effective than life imprisonment [11]. This retributive view holds that those who commit the most serious murders forfeit their right to life, and some advocates maintain execution deters would-be offenders more than incarceration does [12]. Surveys also show a residual public desire in many places to retain the death penalty for victims' families, though support has declined from 68% in 2001 to 55% in 2022 in one poll [13].

Opponents counter that no credible statistical evidence confirms a deterrent effect and that abolitionist countries often have lower homicide rates than retentionist ones, making execution neither necessary nor uniquely effective for safety [14]. They argue the death penalty violates the right to life and constitutes cruel, inhumane punishment, promoting a cycle of vengeance rather than justice [15]. Concerns about arbitrariness and discrimination---including racial bias and inconsistent application---further undermine its claim to proportional, evenhanded justice [12]. For families of murder victims, the reality is mixed: lengthy capital trials and appeals often force relatives to relive trauma for years, and many do not find the promised closure after an execution, complicating grief rather than resolving it [16]. Empirical

work documents that, contrary to expectation, some victims' families report no peace or closure after execution, indicating the death penalty may not reliably serve their healing needs [17].

Error and irreversibility: wrongful convictions are systemic, not rare

Error rates in capital cases are too high to justify an irreversible punishment. A PNAS study estimates about 4.1%---roughly 1 in 25---of death-sentenced defendants were falsely convicted and would likely be exonerated if they remained on death row indefinitely [3]. The Death Penalty Information Center reports that one in every 8.3 people sentenced to death since the 1970s has been wrongfully convicted and later exonerated, underscoring systemic fallibility [18]. DPIC has documented 185 death-row exonerations across 29 states and 118 counties since 1973, demonstrating that these errors are widespread, not confined to a few jurisdictions [18].

Misconduct and structural flaws drive many errors: nearly 70% of wrongful conviction cases involve police, prosecutorial, or official misconduct, including use of knowingly false testimony and racial bias, suggesting routine rather than exceptional breakdowns in due process [19]. Over 550 capital cases have been reversed for prosecutorial misconduct or led to misconduct-related exonerations, representing more than 5.6% of all death sentences imposed since 1972 [20]. Individual cases humanize these failures: Clemente Aguirre-Jarquin alleged a negligent, bias-tainted investigation after his exoneration, and Christopher Tapp was coerced into a false confession under threats of the death penalty, exemplifying how capital leverage can distort truth-finding [21]. The case of Griffin---executed in 1985 while maintaining innocence---is cited as a likely wrongful execution, illustrating the irrevocable harm when the state errs [22].

These realities have catalyzed reforms and abolition. Illinois abolished its death penalty in 2012, citing wrongful convictions and public discomfort with the risk of executing the innocent, reflecting a direct link between documented errors and legislative change [23]. China now requires all death sentences from local courts to be reviewed and ratified by the Supreme Court to reduce wrongful executions, indicating global recognition that heightened safeguards are necessary but still imperfect [24].

Fiscal costs and opportunity costs

Capital cases cost dramatically more than non-capital prosecutions, diverting scarce public safety resources from prevention, policing, and victim services. Per-case costs for death penalty prosecutions range from $1 million to $7 million, compared with about $500,000 for cases resulting in life imprisonment including incarceration costs [1]. Jurisdictions routinely find death cases up to 10 times more expensive than comparable non-death cases, compounding fiscal strain without proven public safety payoff [15].

Alternatives that emphasize supervision and treatment are not only cheaper but can deliver equal or better safety outcomes for appropriate populations, as community-based approaches cost vastly less than incarceration while maintaining or improving outcomes [25]. When budgets are finite, the death penalty's premium price tag crowds out investments in homicide clearance rates, trauma-informed victim services, and evidence-based violence reduction strategies that drive safety more reliably than executions do [1].

Alternatives that work: life imprisonment and restorative justice, tailored to offense severity

Life imprisonment with no parole, or with a substantial social-safety period, is widely implemented and considered the most rational substitute for capital punishment in modern systems [26]. Legal and policy analyses endorse life imprisonment as a more egalitarian and consistent sanction for the worst crimes, addressing incapacitation and proportionality without irreversibility [27]. Many industrialized countries have replaced the death penalty with long prison terms, demonstrating a workable model for severe punishment that preserves the possibility of correcting wrongful convictions [28].

For nonviolent offenses---and for rehabilitatable populations---restorative and non-custodial sentences reduce reoffending and save substantial public funds, complementing life terms reserved for the gravest crimes [29]. Meta-analyses and program evaluations show restorative justice can cut reoffending by up to 20% relative to non-participants, indicating meaningful public-safety gains when appropriately applied [30]. Economic studies estimate that providing restorative justice in 70,000 adult cases could save PS185 million over two years through reduced reoffending, with returns of PS9 to the justice system and PS14 in broader social value per PS1 invested [31]. Pre-court restorative conferencing for youth is projected to save nearly PS275 million over a lifetime, with costs recouped in the first year, reinforcing the fiscal and criminological case for diversion in suitable cases [31].

These approaches are not panaceas and show mixed effectiveness for violent crime, which is why a tailored sentencing framework is essential after abolition [29]. System actor reluctance to extend alternatives to people charged with violent offenses remains a barrier, requiring sustained policy attention and public education to expand evidence-based options where appropriate [32]. The upshot is a dual-track model: life without parole or very long terms for the most serious homicides, paired with restorative and rehabilitative sanctions for nonviolent and lower-risk cases, which jointly outperform execution on safety, cost, and error-correction grounds [27].

Public opinion and legal trends

Public support for capital punishment is declining in the United States, and preferences shift further away from death when alternatives are explicit. Gallup finds support down from 80% in 1994 to 64% recently, with opposition at a near 40-year high, indicating eroding consensus for executions [33]. When offered life imprisonment as an alternative, only 36% of Americans prefer the death penalty, showing that many favor non-capital punishments when given a concrete option that still ensures incapacitation [34]. A national poll found 61% preferred punishments other than death for murder, with 39% specifically endorsing life without parole, further underscoring the shift [35].

Internationally, attitudes are heterogeneous but trends favor abolition, and global law is moving steadily in that direction. As of December 31, 2023, 112 countries had abolished the death penalty for all crimes, with 9 more abolishing for ordinary crimes, reflecting a broad normative movement away from executions [36]. Within Africa, over 80% of countries have abolished the death penalty in law or practice, and as of October 2024, 26 African Union states had formally abolished it while many others maintained moratoria, marking continental momentum despite national variation [37]. Some countries retain high support---for example Singapore, where 77.4% support executions for serious crimes and 87.9% believe it deters drug trafficking---but these views coexist with rising executions in parts of Asia and do not settle the empirical deterrence question [38]. In Japan, polling shows strong support, yet research suggests many views rest on limited or inaccurate information, implying that public education can shift attitudes over time [39].

Legal reforms mirror public unease with error and excess. California imposed a moratorium in 2019 and ordered the dismantling of death row, signaling institutional retreat from

executions in the nation's largest death-sentencing state [40]. Illinois's abolition in 2012 and China's mandatory Supreme Court review of death sentences illustrate different systems' responses to wrongful-execution risks, pointing toward heightened safeguards or outright repeal as the prudent path [23].

The hidden social costs: harms to families and exonerees

Death penalty processes can prolong and complicate grief for victims' families through years of appeals, with many reporting that executions did not bring the expected closure, countering a common rationale for retention [16]. Families of executed or condemned individuals also suffer stigmatized, disenfranchised grief and mental health burdens, with children experiencing educational and emotional harms that ripple across generations [41].

When wrongful convictions are uncovered, the damage is profound and enduring: between 1989 and 2019, innocent people served over 20,800 years in prison, and exonerees face severe psychological and social challenges with limited support, compounding the tragedy beyond the walls of the prison [42]. Exonerees commonly confront long-term unemployment, housing insecurity, and inadequate compensation, revealing systemic gaps in remedy even after innocence is proven [43]. These social costs highlight a moral hazard unique to capital punishment because execution forecloses all possibility of correction, apology, or restitution [3].

Conclusion: why legality fails the test

The death penalty fails on the metrics that matter: it does not uniquely deter murder relative to life imprisonment, it magnifies the risk of irreversible injustice in a system with documented and systemic error, it drains resources better spent on proven safety strategies and victim support, and it often does not deliver the closure its proponents promise [1]. Abolition does not produce the feared crime surge and can open the door to smarter reforms that reduce violence and incarceration together, aligning public safety with human rights and fiscal responsibility [2].

Given the evidence and viable alternatives---life without parole for the worst crimes and restorative, rehabilitative sanctions where appropriate---the death penalty should not be legal. Modern justice can protect the public, honor victims, and correct its own mistakes without resorting to an irreversible punishment that the evidence does not justify [27].

References

[1] 11 - thewomancondemned.com - https://www.thewomancondemned.com/2011/11/why-death-penalty-is-stupid.html

[2] Experience Shows No Parade Of Horribles Following Abolition Of The Death Penalty - deathpenaltyinfo.org - https://deathpenaltyinfo.org/experience-shows-no-parade-of-horribles-following-abolition-of-the-death-penalty

[3] 4 Percent Of Those Sentenced To Die Are Innocent - northdallasgazette.com - https://northdallasgazette.com/2015/03/10/4-percent-of-those-sentenced-to-die-are-innocent/

[4] The End Of The Rope A Critical Analysis Of Capital Punishment With Regards To Deterrence Theory - ijlmh.com - https://ijlmh.com/paper/the-end-of-the-rope-a-critical-analysis-of-capital-punishment-with-regards-to-deterrence-theory/

[5] Death Penalty Essay Argumentative Essay Sample - essayshark.com - https://essayshark.com/examples/death-penalty-essay-argumentative-essay-sample/

[6] Dp3 Study After 1600 Executions The - dppolicy.substack.com - https://dppolicy.substack.com/p/dp3-study-after-1600-executions-the

[7] Do States With Capital Punishment Have Lower Crime Rates - civil-war.net - https://www.civil-war.net/do-states-with-capital-punishment-have-lower-crime-rates/

[8] The Next Step In Criminal Justice Reform Ending Excessive Punishment For Violent Crimes - davisvanguard.org - https://davisvanguard.org/2019/04/the-next-step-in-criminal-justice-reform-ending-excessive-punishment-for-violent-crimes/

[9] Violent Crime Shouldnt Be Left Out Of Prison Reform Debate 2 - justicepolicy.org - https://justicepolicy.org/press/violent-crime-shouldnt-be-left-out-of-prison-reform-debate-2/

[10] No Turning Back - prisonfellowship.org - https://www.prisonfellowship.org/2016/05/no-turning-back/

[11] Death Penalty - emilyandblair.com - https://emilyandblair.com/death-penalty/

[12] Arguments For And Against The Death Penalty - deathpenaltyinfo.org - https://deathpenaltyinfo.org/curriculum/high-school/about-the-death-penalty/arguments-for-and-against-the-death-penalty

[13] The Two Sides Of Death Penalty - empowermag.net - https://www.empowermag.net/post/the-two-sides-of-death-penalty

[14] Essay - auafs.com - https://auafs.com/careers/essay/pros-and-cons-of-the-death-penalty-an-argumentative-essay.html

[15] Death Penalty Usa - lifespark.org - https://lifespark.org/death-penalty-usa/

[16] The Closure Myth - tennesseedeathpenalty.org - https://tennesseedeathpenalty.org/the-facts/the-closure-myth/

[17] Index?Id=11 - theadvocatesforhumanrights.org - https://www.theadvocatesforhumanrights.org/News/A/Index?id=11

[18] Death Penalty Abolition Group Charges Wrongful Conviction Rates Means U S Should Abolish The Death Penalty -

davisvanguard.org - https://davisvanguard.org/2022/01/death-penalty-abolition-group-charges-wrongful-conviction-rates-means-u-s-should-abolish-the-death-penalty/
[19] New Dpic Innocence Report Feb 2021 - americanbar.org - https://www.americanbar.org/groups/committees/death_penalty_representation/publications/project_blog/new-dpic-innocence-report-feb-2021/
[20] Documenting Prosecutorial Misconduct Reversals And Exonerations In Capital Cases - deathpenaltyinfo.org - https://deathpenaltyinfo.org/stories/documenting-prosecutorial-misconduct-reversals-and-exonerations-in-capital-cases
[21] Exonerees In Florida Idaho Murder Cases Initiate Lawsuits For Wrongful Prosecution - deathpenaltyinfo.org - https://deathpenaltyinfo.org/exonerees-in-florida-idaho-murder-cases-initiate-lawsuits-for-wrongful-prosecution
[22] Why Are So Many Innocents Put On Death Row - tucmag.net - https://www.tucmag.net/everyday-people/why-are-so-many-innocents-put-on-death-row/
[23] Documento.Php?Id=16300509 - handsoffcain.info - https://www.handsoffcain.info/documento.php?id=16300509
[24] 10 - deathpenaltythailand.blogspot.com - https://deathpenaltythailand.blogspot.com/2006/10/china-changes-law-on-death-penalty.html
[25] Why Is A Tough On Crime Agenda So Appealing By Caleb Ratzlaff An Intern At Ccjc - ccjc.ca - https://ccjc.ca/why-is-a-tough-on-crime-agenda-so-appealing-by-caleb-ratzlaff-an-intern-at-ccjc/
[26] 7777 - tojqi.net - https://tojqi.net/index.php/journal/article/view/7777
[27] An Excerpt From Are Prisons Obsolete 24084 - bookforum.com - https://www.bookforum.com/politics/an-excerpt-from-are-prisons-obsolete-24084
[28] Latest Project Topics Materials And Research Ideas - eduprojects.ng - https://eduprojects.ng/law/justification-for-and-the-abolition-of-capital-punishment-under-human-rights-law/latest-project-topics-materials-and-research-ideas
[29] Alternatives To Custody - blogs.lse.ac.uk - https://blogs.lse.ac.uk/politicsandpolicy/alternatives-to-custody/
[30] Restorative Justice Will Cover The Country - restorativejustice.org - https://restorativejustice.org/rj-archive/restorative-justice-will-cover-the-country/
[31] The Economic Case For Restorative Justice - covrj.uk - https://covrj.uk/the-economic-case-for-restorative-justice/
[32] Alternative Prosecutorial Responses Violent Crime - innovatingjustice.org - https://www.innovatingjustice.org/publications/alternative-prosecutorial-responses-violent-crime
[33] Poll Shows Decreasing Support Death Penalty - eji.org - https://eji.org/news/poll-shows-decreasing-support-death-penalty/
[34] More Indicators Of The Falling Support For The Death Penalty - worldcoalition.org - https://worldcoalition.org/document/more-indicators-of-the-falling-support-for-the-death-penalty/

[35] Poll Finds Growing Opposition To Death Penalty - new.finalcall.com - https://new.finalcall.com/2010/11/30/poll-finds-growing-opposition-to-death-penalty/
[36] The Countries Ditching The Death Penalty Baltimores Violent Crime Drops 36 Denmarks Culture Vitamins - squirrel-news.net - https://squirrel-news.net/news/the-countries-ditching-the-death-penalty-baltimores-violent-crime-drops-36-denmarks-culture-vitamins/
[37] Triggers For The Abolition Of The Death Penalty In Africa A Southern African Perspective - worldcoalition.org - https://worldcoalition.org/document/triggers-for-the-abolition-of-the-death-penalty-in-africa-a-southern-african-perspective/
[38] Untitled 2 - wethecitizens.net - https://www.wethecitizens.net/untitled-2/
[39] The Public Opinion Myth Why Japan Retains The Death Penalty - worldcoalition.org - https://worldcoalition.org/document/the-public-opinion-myth-why-japan-retains-the-death-penalty/
[40] Death Penalty Justified Abolished - theperspective.com - https://www.theperspective.com/debates/politics/death-penalty-justified-abolished
[41] Grief Loss And Treatment For Death Row Families - worldcoalition.org - https://worldcoalition.org/document/grief-loss-and-treatment-for-death-row-families/
[42] How To Overcome The Trauma Of A Wrongful Conviction - highriselegalfunding.com - https://www.highriselegalfunding.com/faqs/how-to-overcome-the-trauma-of-a-wrongful-conviction/
[43] Tragic Justice Wrongfully Convicted Prisoners Die Shortly After Exoneration - prisonlegalnews.org - https://www.prisonlegalnews.org/news/2017/mar/9/tragic-justice-wrongfully-convicted-prisoners-die-shortly-after-exoneration/

D.1 why should the death penalty be allowed

Why the death penalty should be allowed---if narrowly and rigorously constrained to the worst crimes

A narrowly tailored death penalty can be justified on three core grounds---retribution proportional to the most heinous offenses, incapacitation that permanently prevents the most dangerous offenders from killing again, and a plausible (though empirically debated) deterrent effect---provided it is administered with rigorous safeguards that minimize error and bias and is reserved for the gravest crimes such as willful first-degree murder and terrorism [1][2]. This position reflects enduring public and political judgments about justice and safety even as overall national support has softened, and it remains consistent with the continued retention of capital

punishment by many countries that see it as compatible with their legal and cultural frameworks [3][4].

Retributive justice and the moral logic for capital punishment

At the heart of pro-death-penalty reasoning is retributive justice: some crimes are so egregious that only the ultimate punishment matches their moral gravity, delivering deserved condemnation and a sense of justice for victims' families and communities [1][2]. Philosophers such as Louis P. Pojman argue that capital punishment is morally justified because it proportionally punishes the worst offenses and helps maintain moral order by expressing society's collective judgment against the gravest wrongs [5]. Supporters also maintain that executions can provide a sense of closure and justice to victims' families, reinforcing social norms against extreme violence and validating their suffering in the eyes of the law [6].

Incapacitation and public safety: preventing the worst harms

Capital punishment guarantees that convicted murderers will not kill again in free society or within prison, a risk that persists under life imprisonment because some lifers do commit violence behind bars and, in rare cases, may be released under future policy changes [7]. Proponents highlight that executing convicted killers prevents further murders by offenders who would otherwise remain capable of harming others, including prison staff and inmates, thereby offering a distinct public-safety benefit beyond what life sentences can universally assure [8]. The incapacitation rationale gains urgency when considering evidence that people with serious violent priors have sharply elevated risks of violent reoffending, such as the finding that offenders with prior homicide convictions were 1,467 percent more likely to commit homicide again in a youth cohort study, even as individual risk predictions can be fallible and must be treated with caution [9][10].

Deterrence: theory, mixed evidence, and a prudential case for retention

Supporters contend that the fear of execution can deter would-be offenders, especially for premeditated crimes where offenders weigh consequences more than crimes of passion [11]. One prominent study reported that each additional execution might reduce homicides by about five, while each commutation could correspond with an increase of roughly five, suggesting a potential---but complex---relationship between execution

activity and murder rates [12]. Political advocates also claim that capital punishment can deter violent misconduct beyond murder, including assaults in prison, by signaling that the most severe consequences remain on the table for egregious violence [13].

At the same time, leading reviews caution that the total deterrence literature is mixed and not dispositive: the National Research Council concluded that existing research does not establish whether capital punishment decreases, increases, or has no effect on homicide, and many criminologists remain skeptical that increasing executions or shortening death-row stays would materially affect crime [14][15]. Given this uncertainty, a defensible policy rationale focuses on the combined value of incapacitation, retributive justice, and a plausible deterrent effect, rather than deterrence alone, when justifying a narrow death penalty for the most heinous offenses [16].

Democratic legitimacy: public and political judgments about the worst crimes

In democracies, punishment policy reflects moral judgments about justice and safety, and a substantial share of the public still supports capital punishment, even as overall support has declined from historical peaks [3]. For example, one survey found that 55 percent supported the death penalty as the only way to combat crimes, down from 68 percent in 2001, while other polling shows notable support for life without parole, underscoring a durable yet contested mandate for retaining capital punishment alongside severe non-capital alternatives [3][17]. Politically, candidates and officeholders often emphasize tough stances on capital punishment for first-degree murder to align with voters' moral sentiments, reflecting the salience of retribution and protection in public debate [1]. Support varies across states and is associated with broader ideological and racial-attitude patterns, explaining why some jurisdictions strongly retain capital punishment while others abolish it [18].

International practice and human rights debates

Globally, retentionist countries justify capital punishment primarily on deterrence, retribution, and public safety, often applying it to crimes such as murder, terrorism, treason, and certain severe drug offenses [19]. The international landscape remains divided: roughly 90 nations have abolished the death penalty while a near-equal number retain it, and some states have recently expanded or reintroduced it to confront terrorism

and drug trafficking, demonstrating that many legal systems still view it as compatible with their institutional and cultural norms [4][20]. Proponents argue that, when bounded by due process, capital punishment can coexist with commitments to law and order, though this claim remains contested within the broader human rights community and has prompted ongoing interpretive debates under regional and international instruments [13][21].

Fairness, error reduction, and safeguards: how to make "allowed" ethically defensible

If the death penalty is to be allowed, it must be conditioned on robust safeguards that minimize error and bias, an expectation made more plausible by improvements in forensic science and prosecutorial practices over recent decades [1]. Advances such as modern DNA testing, high-resolution imaging, and digital forensics have strengthened post-conviction review and exoneration efforts, as illustrated by Gary Dotson's landmark DNA exoneration and the William Richards case, where new forensic testimony undermined earlier circumstantial proof [22][23][24]. Beyond forensics, procedural innovations like race-blind charging systems---piloted in San Francisco using machine learning and offered free to other prosecutors---aim to limit racial bias at charging, improving the fairness of decisions that can lead to capital exposure [25][26]. While wrongful convictions persist and their true number is uncertain, these technological and procedural safeguards materially reduce risk and are conditions many supporters cite in morally defending a retained, tightly constrained death penalty [27][26]. Some proponents further argue that, with layered appeals and reviews in place, the residual risk of error is a grave but necessary moral burden to uphold justice for the most heinous crimes [28].

Costs, opportunity costs, and a narrow-use framework

It is important to acknowledge that capital punishment is consistently more expensive than life without parole because of longer, more complex trials, constitutionally mandated appeals, and higher incarceration costs on death row [29]. Empirical cost studies find, for example, that Kansas death penalty cases cost 70 percent more than comparable non-death cases, Maryland capital cases average about $3 million each, and California's system costs roughly $137 million annually versus $11.5 million without capital punishment, with public defenders in Nevada logging an average of 1,211 additional hours per capital case [30][31]. Law enders in Nevada logging an average of 1,211 additional hours per capital case [30][31]. Law enforcem in

Nevada logging an average of 1,211 additional hours per capital case [30][31]. Law enforcemnt officials have also criticized capital punishment for diverting scarce investigative resources from solving other crimes, reinforcing the argument that, if retained, the death penalty should be used sparingly and strategically to avoid undermining broader public-safety goals [32].

A prudential path forward is to allow the death penalty only for the rarest and most aggravated murders where retributive proportionality and incapacitation are most compelling, while investing the bulk of public-safety resources in prevention, trauma-informed services, and effective non-capital enforcement, which many survivors and voters prefer [33]. This balanced approach preserves the moral and incapacitative rationale for capital punishment in exceptional cases without crowding out evidence-based strategies that reduce violence upstream [34].

Bottom line

The case for allowing the death penalty rests on its retributive proportionality for the most heinous crimes, its unique incapacitation value, and a plausible---though empirically contested---deterrent effect, all within a strict framework of due process, forensic rigor, and bias-reducing safeguards [1][11]. Democratic support remains significant even as opinions diversify, and many nations continue to retain capital punishment on grounds of justice and public safety, reinforcing the view that a tightly bounded death penalty can be morally and politically defensible in exceptional cases [3][4]. The ethical path is not maximal use, but narrow allowance: reserve it for the worst crimes, insist on the strongest procedural and scientific protections, and pair it with robust investments in prevention and victim-centered services so society exacts ultimate punishment only where justice and safety truly demand it [25][34].

References

[1] The Death Penalty Debate Pkqqvuyvj - bartleby.com - https://www.bartleby.com/essay/The-Death-Penalty-Debate-PKQQVUYVJ

[2] Capital Punishment Is Not So Easy Way 633052 - essaygpt.hix.ai - https://essaygpt.hix.ai/essay/capital-punishment-is-not-so-easy-way-633052

[3] The Two Sides Of Death Penalty - empowermag.net - https://www.empowermag.net/post/the-two-sides-of-death-penalty

[4] Latest Project Topics Materials And Research Ideas - eduprojects.ng - https://eduprojects.ng/law/justification-for-and-the-abolition-of-capital-punishment-under-human-rights-law/latest-project-topics-materials-and-research-ideas

[5] Wecabrio.Com - https://wecabrio.com/louis-p-pojman.html

[6] Essay - auafs.com - https://auafs.com/careers/essay/pros-and-cons-of-the-death-penalty-an-argumentative-essay.html
[7] Capital Punishment - hubvela.com - https://hubvela.com/hub/pros-cons/capital-punishment/
[8] Death Penalty Justified Abolished - theperspective.com - https://www.theperspective.com/debates/politics/death-penalty-justified-abolished
[9] Crimehistory - news.iastate.edu - https://www.news.iastate.edu/news/2018/12/18/crimehistory
[10] 93755 - daklakonline.com - https://daklakonline.com/world/texas-execution-centers-on-a-jurys-assessment-of-future-dangerousness/93755/
[11] Agreement On Death Penalty - mesotheliomalungcancernet.com - http://mesotheliomalungcancernet.com/agreement-on-death-penalty
[12] Does Death Penalty Reduce Crime Rates - consensus.app - https://consensus.app/questions/does-death-penalty-reduce-crime-rates/
[13] Death Penalty - emilyandblair.com - https://emilyandblair.com/death-penalty/
[14] 201204.Php?Iddocumento=16304879&Mover=0 - english.nessunotocchicaino.it - https://english.nessunotocchicaino.it/archivio_news/201204.php?iddocumento=16304879&mover=0
[15] Capital Punishment In The Usa 65379 - paperdue.com - https://www.paperdue.com/essay/capital-punishment-in-the-usa-65379
[16] 0 9 Support - find-your-support.com - https://find-your-support.com/0-9-support/10-reasons-to-support-capital-punishment.html
[17] Justice - samples.essaysprofessors.com - https://samples.essaysprofessors.com/justice/death-penalty.html
[18] In Many States Support For The Death Penalty Is Driven By Racist Attitudes - blogs.lse.ac.uk - https://blogs.lse.ac.uk/usappblog/2023/10/04/in-many-states-support-for-the-death-penalty-is-driven-by-racist-attitudes/
[19] Arguments For And Against The Death Penalty - essaywritingshub.com - https://essaywritingshub.com/questions/arguments-for-and-against-the-death-penalty/
[20] Resurgence - bruxelles2019.ecpm.org - http://bruxelles2019.ecpm.org/resurgence/
[21] The Death Penalty And Human Rights - pro-papers.com - https://pro-papers.com/samples/law/death-penalty/the-death-penalty-and-human-rights
[22] What Role Does Forensic Technology Play In Wrongful Death Cases - conmylaw.com - https://www.conmylaw.com/blog/2024/11/what-role-does-forensic-technology-play-in-wrongful-death-cases/
[23] First Look Innocence After Guilt The Role Of Post Conviction Dna Testing In The Us - ishinews.com - https://www.ishinews.com/first-look-innocence-after-guilt-the-role-of-post-conviction-dna-testing-in-the-us/
[24] Forensic Advances Raise New Questions About Old Convictions - kanw.com - https://www.kanw.com/2013-03-20/forensic-advances-

raise-new-questions-about-old-convictions
[25] Is Race Baked Into The Criminal Justice System - hadaraviram.com - https://hadaraviram.com/2019/06/13/is-race-baked-into-the-criminal-justice-system/
[26] Artificial Intelligence Will Soon Be Responsible For Reducing Implicit Bias In The San Francisco Das Office - witnessla.com - https://witnessla.com/artificial-intelligence-will-soon-be-responsible-for-reducing-implicit-bias-in-the-san-francisco-das-office/
[27] How Hard Is It To Right A Wrongful Conviction In Canada - thebigstorypodcast.ca - https://thebigstorypodcast.ca/2024/01/24/how-hard-is-it-to-right-a-wrongful-conviction-in-canada/
[28] The Debate Over The Death Penalty And Its Ethical Implications - localnewsherald.com - https://www.localnewsherald.com/the-debate-over-the-death-penalty-and-its-ethical-implications/
[29] The Financial Implications Of The Death Penalty - topstocksinsider.com - https://topstocksinsider.com/2023/07/20/the-financial-implications-of-the-death-penalty/
[30] Death Penalty Cost - amnestyusa.org - https://www.amnestyusa.org/issues/death-penalty/death-penalty-facts/death-penalty-cost/
[31] James Holmes Death Penalty Costs 3X More Than A Life Sentence - mic.com - https://www.mic.com/articles/32255/james-holmes-death-penalty-costs-3x-more-than-a-life-sentence
[32] Public Safety - indianaabolition.org - http://indianaabolition.org/resources/public-safety/
[33] A Game Plan For Communities Ready To Tackle Gun Violence - governing.com - https://www.governing.com/now/a-game-plan-for-communities-ready-to-tackle-gun-violence
[34] Youth Development - equityiskey.org - https://equityiskey.org/youth-development/

D.2 Query: is the criminal justice system fair

```
Is the criminal justice system fair?
```

```
The criminal justice system aspires to "Equal Justice Under Law,"
   but the best available evidence shows it is not fair in
   practice, particularly across race and socioeconomic status,
   with bias now also surfacing in newer algorithmic tools that
   aim to guide decisions [1]. Public perceptions mirror the data:
    fewer than half of U.S. adults believe the system is fair to
   most Americans, and confidence is even lower regarding fairness
    to poor, Black, and Hispanic Americans [2]. There has been
   some measurable progress in reducing certain disparities since
```

the early 2000s, yet entrenched policies and practices continue to produce large, durable inequities from arrest through sentencing and beyond [3][4].

Racial disparities across the pipeline: from police contact to prison

Across stages of the process---stops, arrests, charging, sentencing, and incarceration---racial disparities are large and persistent, indicating systemic rather than isolated problems. Black people are imprisoned at five times the rate of white people, and in several states the Black-to-white disparity exceeds 10 to 1, underscoring that geography can amplify inequity [5][6]. The burden is concentrated among young Black men: one in every ten Black men in his 30s is incarcerated on any given day, and Black men ages 18-19 face imprisonment at 11.8 times the rate of white peers [1][7]. Although Black and Hispanic people constitute roughly a quarter of the U.S. population, they make up 58% of the approximately 2.3 million people imprisoned, reflecting disproportionate impact at scale [8].

Disparities are evident from first contact with law enforcement through misdemeanor arrests and jail admissions, where implicit and explicit bias influence decisions and intensify early-stage inequities [9]. In California, about one-third of traffic stops involve Black drivers---roughly twice the share of stops involving white drivers---illustrating how over-policing feeds cumulative disadvantage downstream [10]. Over-policing and disparate treatment perpetuate negative stereotypes and erode trust between communities and police, which undermines cooperation and public safety over time [11][12].

The disparities begin early in life and extend into youth justice. Black youth are five times more likely than white youth to be incarcerated, and American Indian youth are three times more likely, indicating that unequal treatment is embedded at juvenile stages as well [13]. Arrest data reinforce the pattern: African Americans accounted for 27% of arrests in 2016---double their share of the population---and over a quarter of drug arrests in 2015 despite evidence that drug use rates do not fully explain these gaps [14].

Socioeconomic status and the bail-to-conviction pipeline

Socioeconomic status shapes outcomes at virtually every juncture, but pretrial policies---especially cash bail---create some of the starkest wealth-based inequities. More than 60% of defendants are detained pretrial because they cannot afford bail, an arrangement that penalizes poverty and contradicts the presumption of innocence [15]. Pretrial detention increases the probability of conviction by 13 to 14 percentage points, largely by pressuring guilty pleas among people seeking release from jail, which magnifies the impact of poverty on case outcomes regardless of culpability [16][17]. The economic shock of pretrial detention, including job loss, eviction, and homelessness, ripples through families and communities, further entrenching disadvantage that began as an inability to pay [18][19].

These patterns contribute to widespread poverty traps in low-income communities, where fines and fees for low-level offenses create cycles of debt that often exceed any public safety benefits they were meant to provide [20]. The broader macroeconomic toll is staggering: imprisonment imposes about $1 trillion in annual costs on the U.S. economy, much of it borne by already disadvantaged communities [21]. Family-level harms are pervasive; many families cannot meet basic needs when a member is incarcerated, and communities of color---already overrepresented behind bars---bear a disproportionate share of these burdens [22].

Judicial discretion, implicit bias, and unequal sentencing

Even when formal legal standards are the same, judicial discretion can widen racial gaps in sentencing. After the Federal Sentencing Guidelines were rendered advisory in United States v. Booker (2005), Black defendants received sentences nearly two months longer on average than comparable white defendants, a roughly 4% increase tied to greater judicial discretion [23]. Disparities were larger among judges appointed after Booker, suggesting that shifts in discretion changed how similarly situated defendants were treated by race [23]. Experimental and observational work shows that implicit racial bias can influence conscious behavior in courtrooms, even though legal doctrine focuses on explicit discrimination, which helps explain why disparities persist without overtly racist intent [24]. Independent analyses have identified individual federal judges who impose harsher sentences on Black and Hispanic defendants, underscoring the role of decision-maker bias in

aggregate disparities [25].

At the same time, legal standards make it hard to remedy biases that are subtle or systemic: the Supreme Court's McCleskey v. Kemp decision requires proof of conscious racial intent, a threshold ill-suited to address implicit bias that pervades complex decision processes [26]. Some studies find limited evidence linking racial bias to sentencing disparities in particular district courts, indicating that the size and direction of racial gaps can vary by case type, setting, and defendant characteristics, rather than being uniform everywhere [27]. The coexistence of evidence for bias with heterogeneous findings across contexts points to a system where discretion, structure, and local practice interact to produce uneven but enduring inequities [28].

Prosecutorial power: charging, plea bargaining, and mandatory minimums

Prosecutors exercise pivotal discretion that can either correct or compound earlier disparities, and the evidence shows both dynamics. In North Carolina, prosecutors' charging decisions shifted over time; they initially compounded stop-based disparities in the late 1990s and early 2000s but later attenuated them from 2015 to 2019, demonstrating the potential for policy and practice to move toward equity within the same jurisdiction [29]. Yet at the federal level, Black defendants are more likely to be charged with offenses carrying mandatory minimums, and Hispanic males are also more likely to receive mandatory minimum sentences, amplifying racial and ethnic disparities through charging choices that constrain judicial discretion downstream [30][31]. In New York County, racial and ethnic disparities appeared at multiple decision points, with the clearest gaps in initial bail requests, showing how prosecutorial recommendations can shape early case trajectories with lasting effects [32].

Researchers disagree about the magnitude of prosecutorial bias in driving overall prison disparities, with some studies suggesting race and class do not systematically change charging decisions in experimental settings, even as real-world administrative data show persistent racial differences in outcomes across places and stages [33][34]. Prosecutors also report viewing racial disparities as systemic and largely beyond their control, which can limit internal accountability for closing gaps during plea bargaining and sentencing recommendations [35]. The American Bar Association stresses

that systemic racism pre-dates first contact and continues through charging and plea bargaining, placing a responsibility on prosecutors to proactively identify and correct inequities within their own practices [36].

Algorithmic "risk" tools: new technology, old inequities

The turn to algorithmic risk assessments has not solved fairness problems and, in some cases, has deepened them. The widely used COMPAS tool labeled Black defendants "high risk" at roughly twice the rate of white defendants, producing racially disparate errors that can influence bail and sentencing decisions [37]. ProPublica's 2016 investigation similarly found that non-reoffending Black defendants were nearly twice as likely as white defendants to be misclassified as high risk, illustrating how biased errors can accumulate within a punitive system [38]. The core technical challenge is a fairness paradox: in the presence of group disparities, it is mathematically impossible for a single algorithm to satisfy all reasonable fairness criteria at once, so reducing one disparity often increases another [38].

Longitudinal evaluations show judges do respond to algorithmic scores---imposing longer sentences on "high-risk" defendants and shorter ones on "low-risk" defendants---yet these tools have not produced net reductions in incarceration or clear public safety gains, raising doubts about promised benefits [39]. Racial disparities persist in both the scores and their application, with judges more likely to follow leniency recommendations for white defendants, which undermines equal treatment even when using the same tool [39]. Many systems are trained on historical crime and enforcement data, embedding past over-policing of poor neighborhoods and communities of color directly into model inputs and outputs, which biases the tools before a judge ever sees a score [40][41].

Policy responses are emerging but remain incomplete. A presidential executive order directed federal agencies to prevent algorithmic discrimination, and the FTC has urged firms to test models for discriminatory outcomes, yet the U.S. lacks comprehensive AI auditing laws, leaving critical gaps in oversight [42][43]. At least 20 states have incorporated risk tools into sentencing, even as many researchers urge excluding inputs correlated with race, redesigning algorithms to equalize key outcomes, or rejecting their use in criminal decisions altogether to avoid amplifying bias [44][45].

Legitimacy and trust: what the public and victims see

Fairness is not only about outcomes; it is also about whether people experience the process as just. Only 47% of Americans believe the justice system is fair to most people, and perceived fairness is even lower regarding poor, Black, and Hispanic Americans, indicating deep legitimacy challenges for courts and law enforcement [2]. Racial gaps in confidence are stark: just 27% of Black adults report high confidence in police, compared to 56% of white adults, and trust eroded further following high-profile police shootings in minority communities [46][47]. In the U.K., a majority of British-born Black, Asian, and Minority Ethnic people believe the justice system discriminates against them, illustrating that legitimacy concerns transcend national context [48].

Victims' perceptions highlight the importance of voice and participation: burglary victims in Minnesota reported that opportunities to participate in proceedings strongly shaped their sense of fairness, with different preferences for rehabilitation, compensation, or punishment influencing what "justice" meant to them [49]. Cuts to legal aid further undermine perceived fairness in court, particularly for those least able to navigate complex procedures, linking resource access to legitimacy as well as outcomes [50]. Procedural justice---fair, respectful treatment by authorities---predicts trust and cooperation, which are essential for reporting crimes and serving as witnesses or jurors [51][52].

Indigenous and international evidence: overrepresentation and mistrust

Outside the U.S., similar patterns of overrepresentation and mistrust highlight systemic problems in how justice systems treat marginalized communities. In Canada, Indigenous people accounted for 25% of accused in criminal courts in 2016 despite being only 5% of the population, and Indigenous accused were 55% less likely than white accused to see their charges withdrawn, dismissed, or discharged, indicating differential treatment in case resolution [53]. Indigenous people are also about twice as likely to mistrust their local police service as non-Indigenous people, with particularly high levels of mistrust in Saskatchewan and the territories, reflecting the

cumulative effects of historical injustices and contemporary practices on legitimacy [54]. These data points echo global concerns that justice systems can reproduce social hierarchies unless guarded against bias at each decision point [55].

The economic fallout: lost earnings, family disruption, and intergenerational harm

Fairness in criminal justice cannot be separated from its economic consequences for individuals, families, and communities. People with convictions earn at least 16% less than peers, those who have been to prison lose roughly half of their earning potential, and aggregate lost wages for justice-involved people exceed $372 billion annually, creating a feedback loop between punishment and poverty [56]. Families struggle to meet basic needs when a member is incarcerated, with two-thirds reporting difficulty affording essentials and heavy reliance on low-wage work or public assistance [57]. Children bear long-term costs: in 2015, about 3.5% of U.S. children had a parent behind bars, and Black children were more than five times as likely as white children to experience parental incarceration, which is linked to educational disruptions and mental health challenges [58][59].

These harms concentrate in communities already facing disadvantage, where nearly 70% of Black men without a high school diploma will experience incarceration by midlife, entrenching neighborhood instability and limiting intergenerational mobility [60]. Reentry compounds the cycle: roughly one-third of the more than 620,000 people released annually return to prison at some point, and 65% remain unemployed four years after release, showing how post-sentence exclusion perpetuates inequality beyond formal punishment [61][62].

Progress and promising reforms---alongside barriers and pushback

There is real, measurable progress in some areas, but it has not yet closed large gaps. From 2000 to 2016, the Black-white state imprisonment disparity fell from 8.3-to-1 to 5.1-to-1, and by 2020, Black adults were imprisoned at 4.9 times the rate of white adults, down from 8.2 times around 2000, reflecting changes in enforcement and sentencing, especially for drug crimes [63][3]. The disparity for drug-related imprisonment

decreased by about two-thirds between 2000 and 2019, and shifts in drug enforcement helped reduce the number of African American men in state prisons even as white male incarceration rose, suggesting policy leverage points exist [4][64]. At the same time, researchers trace many enduring disparities to punitive policies of the war on drugs and racialized stereotypes in decision-making, and there is ongoing policy resistance that threatens to stall or reverse gains, indicating the need for sustained, structural reforms rather than episodic adjustments [65][66].

Bail reform has shown potential to reduce wealth-based detention. Illinois became the first state to eliminate wealth-based pretrial detention, and international experience in countries like Germany, Ireland, and Canada shows non-monetary alternatives can secure appearances and safety without cash, supporting adoption of personal recognizance and supervised release models in U.S. jurisdictions [67][68]. Non-cash options such as personal recognizance bonds directly target the core inequity of the cash bail system by allowing release without payment, a critical shift given that roughly three-quarters of jail detainees are held pretrial in the U.S. [69][70].

Community trust-building efforts underscore the importance of procedural justice and reconciliation. The National Initiative for Building Community Trust and Justice found that police leadership is critical for effective training in procedural justice and implicit bias, though such trainings require substantial resources, and its reconciliation pillar applied lessons from transitional justice to acknowledge harms and repair relationships [71][72]. The Safety and Justice Challenge supports changing not just decisions but the values guiding them, and tools like the Measuring Community Trust Tool help agencies assess and improve legitimacy, especially when incorporating the voices of those most affected by the system [73][74]. Including directly impacted communities in prosecutorial and policy reform agendas has been identified as essential to advancing racial equity and rebuilding safety [75].

Behavioral health diversion is another fairness-enhancing pathway. With mental illness and substance use disorders highly prevalent among people in jail and prison, mental health and treatment courts can reduce recidivism by addressing root causes rather than relying solely on incarceration, and increased access to drug treatment correlates with lower crime rates [76][77]. Still, access to quality, integrated treatment remains uneven by race and ethnicity, which risks reproducing disparities unless collaboration across justice, mental health, and substance use systems is deliberately designed for equity [78][79].

Bottom line

Taken together, the evidence shows that the criminal justice system is not fair in practice because race and wealth substantially shape who is stopped, detained, charged, convicted, and imprisoned, and because newer tools like risk algorithms often reflect and reinforce those inequities rather than curing them [14][37]. While certain gaps have narrowed since the early 2000s and promising reforms---such as bail changes, procedural justice initiatives, and treatment-diversion programs---demonstrate viable paths forward, the persistence of large disparities, the documented role of discretion and implicit bias, and the heavy socioeconomic fallout mean the system still falls short of equal justice under law [3][23]. A fairer system will require aligning policy, practice, and technology around equity at every decision point, sustained accountability for discretionary actors, and meaningful investment in community trust, behavioral health, and economic mobility for people and places most affected by mass punishment [36][73].

References

[1] Sentencing Bias - sentencing.net - https://sentencing.net/sentencing/sentencing-bias

[2] 47 See U S Justice System As Fair To Most Americans 34 Don T - rasmussenreports.com - https://www.rasmussenreports.com/public_content/politics/general_politics/june_2011/47_see_u_s_justice_system_as_fair_to_most_americans_34_don_t

[3] Focusing On Intervention Instead Of Incarceration Could Change The Biggest Racial Gap In Prisons - theirl.xyz - http://theirl.xyz/13527/focusing-on-intervention-instead-of-incarceration-could-change-the-biggest-racial-gap-in-prisons/

[4] Multi State Report - counciloncj.foleon.com - https://counciloncj.foleon.com/reports/racial-disparities/multi-state-report

[5] Racism In The Criminal Justice System 6Ae359 - essaygpt.hix.ai - https://essaygpt.hix.ai/essay/racism-in-the-criminal-justice-system-6ae359

[6] Sentencing Project Report Racial Disparities In Incarceration - eji.org - https://eji.org/news/sentencing-project-report-racial-disparities-in-incarceration/

[7] Is The American Criminal Justice System Flawed - mailletcriminallaw.com - https://www.mailletcriminallaw.com/blog/is-the-american-criminal-justice-system-flawed/

[8] Artificially Intelligent Criminal Justice Reform - vinayiyengar.com - https://www.vinayiyengar.com/2017/06/23/

artificially-intelligent-criminal-justice-reform/
[9] Fairness Of The Justice System - nyc.equityindicators.org - https://nyc.equityindicators.org/themes/justice/fairness-of-the-justice-system/
[10] 5 Ways The Criminal Justice System Has Changed Over The Years 2263.Shtml - attorneyhelp.org - https://www.attorneyhelp.org/guide/5-ways-the-criminal-justice-system-has-changed-over-the-years-2263.shtml
[11] 6353 Racial Discrimination In The Criminal Justice System 13 - carreraremote.com - https://carreraremote.com/6353-racial-discrimination-in-the-criminal-justice-system-13/
[12] African Americans And The Criminal Justice System 57993B - essaygpt.hix.ai - https://essaygpt.hix.ai/essay/african-americans-and-the-criminal-justice-system-57993b
[13] New Report Shows Persistent Racial And Ethnic Disparities In Youth Incarcerations - nationofchange.org - https://www.nationofchange.org/2021/02/04/new-report-shows-persistent-racial-and-ethnic-disparities-in-youth-incarcerations/
[14] A Report To The Un Reveals Deep Racial Disparities In American Criminal Justice System - psmag.com - https://psmag.com/social-justice/a-report-to-the-un-reveals-deep-racial-disparities-in-american-criminal-justice-system/
[15] Us Commission Civil Rights Releases Report Civil Rights Implications Cash Bail - usccr.gov - https://www.usccr.gov/news/2022/us-commission-civil-rights-releases-report-civil-rights-implications-cash-bail
[16] 6 - chicagounbound.uchicago.edu - https://chicagounbound.uchicago.edu/jle/vol60/iss3/6/
[17] Cash Bail Studies Reveal Racism Classism In The System - sentencing.net - https://sentencing.net/legislation/cash-bail-studies-reveal-racism-classism-in-the-system
[18] A Simple Plan For Investigating - taxweb.info - https://taxweb.info/a-simple-plan-for-investigating/
[19] If You Think You Understand Then Read This - legalwebs.info - https://legalwebs.info/if-you-think-you-understand-then-read-this/
[20] Why Inequitable And Burdensome Court Issued Fines And Fees Are A Health Issue And What Health And Policy Leaders Can Do About It - camdenhealth.org - https://camdenhealth.org/resources/why-inequitable-and-burdensome-court-issued-fines-and-fees-are-a-health-issue-and-what-health-and-policy-leaders-can-do-about-it/
[21] The Many Costs Of Incarcerating Low Level Criminal Offenders - coreycohen.com - https://www.coreycohen.com/blog/2020/11/the-many-costs-of-incarcerating-low-level-criminal-offenders/
[22] Opportunities For Finhealth Innovation And Impact In Criminal Justice - finlab.finhealthnetwork.org - https://finlab.finhealthnetwork.org/opportunities-for-finhealth-innovation-and-impact-in-criminal-justice/
[23] 664 - chicagounbound.uchicago.edu - https://chicagounbound.uchicago.edu/law_and_economics/664/

[24] 689 - scholarship.law.stjohns.edu - https://scholarship.law.stjohns.edu/faculty_publications/689/
[25] Recent Study On Federal Sentences Identifies Most Discriminatory Federal Judges - thefederaldocket.com - https://thefederaldocket.com/recent-study-on-federal-sentences-identifies-most-discriminatory-federal-judges/
[26] Racial Sentencing Disparities - shortform.com - https://www.shortform.com/blog/racial-sentencing-disparities/
[27] Does Racial Bias Explain The Black White Sentencing Gap Across U S Courts By Michael T Light And Karl Vachuska - sociology.wisc.edu - https://sociology.wisc.edu/2024/04/03/does-racial-bias-explain-the-black-white-sentencing-gap-across-u-s-courts-by-michael-t-light-and-karl-vachuska/
[28] Subtle And Systematic Discrimination In Sentencing Practices - coggle.it - https://coggle.it/diagram/ZuYgYXdV1iFRTgfN/t/subtle-and-systematic-discrimination-in-sentencing-practices
[29] Brokers Of Bias In The Criminal System Do Prosecutors Compound Or Attenuate Racial Disparities In Policing - hls.harvard.edu - https://hls.harvard.edu/bibliography/brokers-of-bias-in-the-criminal-system-do-prosecutors-compound-or-attenuate-racial-disparities-in-policing/
[30] Index.Cfm?Fuseaction=Ask This.View&Askthisid=550 - niemanwatchdog.org - http://www.niemanwatchdog.org/index.cfm?fuseaction=Ask_this.view&askthisid=550
[31] Sage Journal Articles - edge.sagepub.com - https://edge.sagepub.com/stohrcorrections/student-resources/chapter-5/sage-journal-articles
[32] 4021 - academicworks.cuny.edu - https://academicworks.cuny.edu/gc_etds/4021/
[33] Baughman Co Authors Study On Prosecutors Race And Class Biases - resgestae.utah.edu - https://resgestae.utah.edu/issues/summer-2020/baughman-co-authors-study-on-prosecutors-race-and-class-biases/
[34] Papers.Cfm?Abstract Id=4728007 - papers.ssrn.com - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4728007
[35] Equity Blogpost - rstreet.org - https://www.rstreet.org/commentary/equity-blogpost/
[36] Prosecutors Confront Bias - americanbar.org - https://www.americanbar.org/news/abanews/aba-news-archives/2023/02/prosecutors-confront-bias/
[37] Machine Learning Compas Racism Policing Fairness - dev.massivesci.com - https://dev.massivesci.com/articles/machine-learning-compas-racism-policing-fairness/
[38] Executive Summary Of Algorithmic Injustice - tna-dev.tbfdev.com - http://tna-dev.tbfdev.com/publications/executive-summary-of-algorithmic-injustice
[39] Hdsi 5Th Anniversary Keynote Megan Stevenson 11 3 - hsph.harvard.edu - https://www.hsph.harvard.edu/biostatistics/2021/11/hdsi-5th-anniversary-keynote-megan-stevenson-11-3/
[40] The Danger - pretrialrisk.com - https://pretrialrisk.com/the-danger/

[41] Racial Bias In Criminal Risk Scores Is Mathematically Inevitable - psmag.com - https://psmag.com/news/racial-bias-in-criminal-risk-scores-is-mathematically-inevitable/
[42] Guiding Principles Address Bias Healthcare Algorithms - biologicalsciences.uchicago.edu - https://biologicalsciences.uchicago.edu/news/guiding-principles-address-bias-healthcare-algorithms
[43] ?Oref=Wt Next Story - washingtontechnology.com - https://www.washingtontechnology.com/opinion/2023/06/confronting-biases-embedded-ai-and-mitigating-risks/387468/?oref=wt-next-story
[44] 08 - mattmangino.com - http://www.mattmangino.com/2019/08/gatehouse-risk-assessment-under.html
[45] Bias In Bias Out?Ref=Publicsquare.Uk - yalelawjournal.org - https://www.yalelawjournal.org/article/bias-in-bias-out?ref=publicsquare.uk
[46] How Can Law Enforcement Build Trust With Black Americans - okjusticereform.org - https://www.okjusticereform.org/blog/how-can-law-enforcement-build-trust-with-black-americans
[47] Justice Collaboratory Holds Inaugural Conference - law.yale.edu - https://law.yale.edu/yls-today/news/justice-collaboratory-holds-inaugural-conference
[48] Building Trust How Our Courts Can Improve Criminal Court Experience Black Asian And - justiceinnovation.org - https://www.justiceinnovation.org/publications/building-trust-how-our-courts-can-improve-criminal-court-experience-black-asian-and
[49] Victim Understanding Of Fairness Burglary Victims In Victim Offender Mediation - restorativejustice.org - https://restorativejustice.org/rj-archive/victim-understanding-of-fairness-burglary-victims-in-victim-offender-mediation/
[50] Search?Page=47 - justiceinnovation.org - https://justiceinnovation.org/search?page=47
[51] Making The System Work - sharedsafety.us - https://sharedsafety.us/making-the-system-work/
[52] 5B876A101De9Cd7Eca257B16000E0F16?Opendocument - abs.gov.au - https://www.abs.gov.au/ausstats/abs@.nsf/Previousproducts/5B876A101DE9CD7ECA257B16000E0F16?opendocument
[53] Justice Canada Study Finds Courts Stacked Against Indigenous Accused - aptnnews.ca - https://www.aptnnews.ca/national-news/justice-canada-study-finds-courts-stacked-against-indigenous-accused/
[54] Justice Policing - afn.ca - https://afn.ca/rights-justice/justice-policing/
[55] The Balance Of Criminal Law And Indigenous Peoples Rights - hawparvilla.my.id - https://hawparvilla.my.id/the-balance-of-criminal-law-and-indigenous-peoples-rights/
[56] Criminal History Continues Punishment Far Beyond Prison Term - legalexaminer.com - https://www.legalexaminer.com/politics/criminal-history-continues-punishment-far-beyond-prison-term/
[57] Prison Costs Families - motherjones.com - https://www.motherjones.com/criminal-justice/2015/09/prison-costs-families/
[58] When A Parent Is Incarcerated Partners And Children Also Pay

A Price - prb.org - https://www.prb.org/resources/when-a-parent-is-incarcerated-partners-and-children-also-pay-a-price/
[59] Affordable Bail Bonds In Miami County - delaughterbailbonds.com - https://delaughterbailbonds.com/blog/tag/affordable-bail-bonds-in-miami-county/
[60] For 50 Years Mass Incarceration Has Hurt American Families Heres How To Change It - source.washu.edu - https://source.washu.edu/2021/10/for-50-years-mass-incarceration-has-hurt-american-families-heres-how-to-change-it/
[61] Work And Opportunity Before And After Incarceration - brookings.edu - https://www.brookings.edu/articles/work-and-opportunity-before-and-after-incarceration/
[62] Dei For Formerly Incarcerated People - resources.workable.com - https://resources.workable.com/stories-and-insights/dei-for-formerly-incarcerated-people
[63] Racial Disparities In Us Prisons Decline Study - alaturkanews.com - https://www.alaturkanews.com/2019/12/03/racial-disparities-in-us-prisons-decline-study/
[64] Study Racial Disparities Shrink In State Prison Populations - scrippsnews.com - https://www.scrippsnews.com/us-news/crime/study-racial-disparities-shrink-in-state-prison-populations
[65] Race And Sentencing Disparity - academyforjustice.asu.edu - https://academyforjustice.asu.edu/resource/race-and-sentencing-disparity/
[66] Report Reveals That Racial Disparities In Incarceration Persist Despite Progress - darealprisonart.news - https://darealprisonart.news/report-reveals-that-racial-disparities-in-incarceration-persist-despite-progress/
[67] Ending Money Bond - ransom-lawfirm.com - https://ransom-lawfirm.com/ending-money-bond/
[68] Only Two Countries Have For Profit Bail Systems - bailproject.org - https://bailproject.org/learn/only-two-countries-have-for-profit-bail-systems/
[69] Breaking The Chains Bail Bond Alternatives - bestbailbondscopperascove.com - https://bestbailbondscopperascove.com/breaking-the-chains-bail-bond-alternatives/
[70] Pretrial Transformation And Abolition - journals.law.harvard.edu - https://journals.law.harvard.edu/crcl/pretrial-transformation-and-abolition/
[71] National Initiative Building Community Trust And Justice - ojp.gov - https://www.ojp.gov/library/publications/national-initiative-building-community-trust-and-justice
[72] Building Community Trust And Justice - nacole.org - https://www.nacole.org/building_community_trust_and_justice
[73] Engaging Community Members In Criminal Justice Reform - macfound.org - https://www.macfound.org/press/perspectives/engaging-community-members-in-criminal-justice-reform
[74] Transforming Police Community Relations Using The Innovative Trust Measurement Tool - operationbluebeat.com - https://operationbluebeat.com/news/transforming-police-community-

relations-using-the-innovative-trust-measurement-tool
[75] Justleadershipusa And Association Of Prosecuting Attorneys Collaborate To Issue Historic Report On Advancing Race Equity In The Criminal Legal System - jlusa.org - https://jlusa.org/media-release/justleadershipusa-and-association-of-prosecuting-attorneys-collaborate-to-issue-historic-report-on-advancing-race-equity-in-the-criminal-legal-system/
[76] About - samhsa.gov - https://www.samhsa.gov/criminal-juvenile-justice/about
[77] Latest In A Series Of Policy Briefs On Public Safety Data Shows Substance Abuse Treatment Reduces Crime - justicepolicy.org - https://justicepolicy.org/press/latest-in-a-series-of-policy-briefs-on-public-safety-data-shows-substance-abuse-treatment-reduces-crime/
[78] Understanding And Reducing Inequities In Substance Use Treatment - c-dias.org - https://www.c-dias.org/presentations/understanding-and-reducing-inequities-in-substance-use-treatment/
[79] Expansion Of An Integrated Treatment Program For Multi Diagnosis And High Risk Populations - dualdiagnosis.org - https://dualdiagnosis.org/research/expansion-of-an-integrated-treatment-program-for-multi-diagnosis-and-high-risk-populations/

D.3 Query: what role did indians play in the wars for empire?

Indians in the wars for empire: indispensable soldiers, suppliers, financiers---and ultimately resisters and nation-builders

Indians were indispensable to the British Empire's wars, providing vast manpower, materiel, and money while bearing disproportionate social and economic costs that later fueled resistance and decolonization [1].

Across the two world wars, about 3.8 million soldiers from the Indian subcontinent served under the British flag, a scale of participation that made India central to imperial war-making [2].

Beyond manpower, India shipped 170,000 animals and 3.7 million tonnes of supplies in World War I, underscoring its role as the empire's logistical spine [3].

Financially, Indian soldiers and the Indian state delivered an extraordinary subsidy---roughly PS8 billion in today's value as a one-time war contribution and PS2.4 billion annually thereafter during imperial wars---cementing India's place as the empire's fiscal shock absorber [4].

These contributions came amid acute suffering, epitomized by the Bengal Famine of 1943, in which an estimated three million

people died as imperial wartime priorities eclipsed local food security [5].
- At the same time, exposure to global battlefields and racial hierarchies catalyzed anti-colonial mobilization, including the formation of the Indian National Army to fight the British during World War II [6].

Fighting the empire's wars: global combat roles and costly sacrifices

- Indian forces fought on nearly every front the empire deemed strategic, from Flanders and Gallipoli to East Africa, Palestine, and Mesopotamia in World War I, and from North Africa and Italy to Burma in World War II [7].
- Over one million Indian soldiers were deployed in World War I alone, often rushed in as the British Expeditionary Force's only available professional reinforcement in 1914-15 [8].
- In total, about 1.5 million Indian volunteers or active soldiers participated in World War I, with roughly 50,000-62,000 deaths and 65,000-67,000 serious injuries, reflecting the human toll of "imperial fire brigade" service across distant theatres [9].
- The Indian Army acted as an imperial fire brigade for the British, shifting rapidly from Belgium and France to Egypt, East Africa, Gallipoli, Palestine, Mesopotamia, and even Singapore to suppress revolt and hold imperial lines [10].
- At the outset, it was one of the only trained professional forces able to reinforce the British Expeditionary Force, making its early-war role decisive for imperial survival [11].
- Indian troops were frequently assigned the most dangerous jobs---digging trenches, riding dispatch, and manning guns---often for meager pay of Rs 11 per month, yet still winning early Victoria Crosses for gallantry [12].
- More than 2.5 million Indians served in World War II, forming one of history's largest volunteer armies, a force that absorbed intense combat across Asia, Africa, and Europe [13].
- Casualties in World War II exceeded 67,000 killed, more than 34,000 wounded, and over 67,000 captured, a reminder that the scale of sacrifice rose with the scale of the war [14].
- Even as Indian public opinion was often apathetic or hostile to imperial wars, British commanders remarked on the Indian Army's loyalty, a paradox that defined the politics of service under empire [15].

Across the services: army, air, and navy under imperial command

- Indian formations reinforced the Western Front and campaigns across the Middle East and Africa, with around 800,000 Indian soldiers engaged in active fighting during World War I [16].
- They deployed as tailored expeditionary forces, such as the Indian

Expeditionary Force B from Bombay to East Africa in late 1914, reflecting how India underwrote imperial flexibility across oceans [17].

Indian participation also extended to the air: four Indian volunteers flew with the Royal Flying Corps in World War I, with Indra Lal Roy posthumously awarded the Distinguished Flying Cross after destroying ten enemy aircraft [18].

Institutionally, the Royal Indian Air Force was established on 8 October 1932 and attached to the RAF in World War II, earning the "Royal" prefix in 1945 for distinguished service [19].

At sea, Indian personnel crewed escorts and minesweepers in the Royal Indian Navy during World War II, providing critical protection in convoy and littoral operations [20].

Near the war's end, the RIN mutiny of February 1946---encompassing roughly 20,000 sailors, 78 of 88 ships, and 20 shore establishments---signaled a crisis of legitimacy for British rule [21].

This naval revolt spread rapidly through Bombay and beyond, turning service grievances into a mass political challenge to imperial authority [22].

Unrest also touched the RIAF, which the 1946 Cabinet Mission judged "cannot be regarded as reliable" for suppressing anti-British movements, highlighting the fragility of the colonial military edifice as decolonization loomed [23].

Beyond manpower: India as the empire's quartermaster and banker

India provisioned imperial wars on a scale few other colonies could match by supplying 170,000 animals and 3.7 million tonnes of materiel in World War I, integrating village economies into global logistics chains [3].

British-controlled India transferred an estimated PS146 million directly to British war costs between 1914 and 1920, financing a significant share of the Allied effort [24].

Across the empire, colonial contributions included PS23.3 million in gifts, PS10.7 million in interest-free loans, and PS14 million in low-interest loans, with India a principal contributor to these flows [24].

Taken together, Indian soldiers' pay deductions and state transfers amounted to about PS8 billion in today's money as a one-off contribution, with PS2.4 billion annually continuing afterward during the empire's wars, demonstrating the fiscal centrality of India to Britain's military machine [4].

Wartime disruption also had unexpected industrial effects at home, as a trade shock from reduced British imports helped account for roughly 24% of industrial employment growth during World War I, showing that even extractive war economies could yield pockets of local industrialization under forced protection [25].

The costs of empire: extraction, discrimination, and famine

- The empire's wartime and peacetime extractions formed part of a broader colonial drain that siphoned an estimated $45 trillion from India between 1765 and 1938, reshaping the economy in Britain's favor [26].
- Nearly one-third of tax revenues raised from Indians were used to buy goods from India itself, a mechanism that masked resource transfer as market exchange while starving domestic investment [27].
- Colonial policy systematically favored raw material export over local industry, leaving Indian enterprises stunted and the agrarian base overburdened by taxes and price shocks [28].
- The social hierarchy of service mirrored the racial hierarchy of empire, as Indians were barred from the highest commands and often dismissed as inherently inferior by British generals [2].
- This disdain surfaced starkly during the Dunkirk evacuation with orders to "cut loose your Indians and your mules," a phrase that captured how colonial troops were treated as expendable in moments of crisis [2].
- Food policy during World War II intensified human costs, as London prioritized military needs over Indian civilians, producing the Bengal Famine of 1943 with about three million deaths despite sufficient production in Bengal, where entitlement failures blocked access to food [5].
- Churchill's diversion of grain stocks away from Bengal during 1943 has been linked to roughly four million famine deaths, underscoring the lethal synergy of wartime requisitioning and imperial indifference [29].
- Earlier, end-nineteenth-century famines killed on the order of 50 million people, as colonial land and revenue regimes amplified climatic shocks and converted poor harvests into mass mortality [30].
- One estimate places famine-related mortality linked to British imperial policy at at least 100 million between 1880 and 1920, illustrating a catastrophic scale of demographic loss in India unmatched elsewhere in the empire's records [31].

Making the empire---and resisting it: Indian agency from Mysore and Maratha wars to INA and mutinies

- Long before the world wars, Indians helped make---and contest---the British Empire through regional conflicts that determined the subcontinent's political future, notably the Anglo-Mysore and Anglo-Maratha wars [32].
- In Mysore, Hyder Ali and later Tipu Sultan built a modernizing state that defeated British-led coalitions and seized strategic ports like Mangalore before being crushed in 1799, after which British allies were installed to secure the southern flank of

empire [33].
The Maratha wars reflected deep Indian agency and factional politics, as the Treaty of Bassein brought Baji Rao II under British protection and triggered the Second Anglo-Maratha War, whose defeats paved the way for British ascendancy [34].
By 1817-18, the Third Anglo-Maratha War ended Maratha sovereignty, completing a transition from Indian-led regional hegemony to British paramountcy across much of the subcontinent [35].
These wars reveal that Indians fought on every side of imperial contests---as allies, adversaries, and intermediaries---helping decide how empire was constituted in South Asia [36].

War, society, and nationalism: how service under empire remade identities at home

War service brought profound social change within India, as soldiers returned with disabilities, new skills, and global experiences that reshaped status and identity in their communities [37].
On the Western Front, 80-90% of Indian soldiers reported positive impressions of France, sometimes calling it "paradise," experiences that unsettled local hierarchies upon their return [38].
Exposure to racial ceilings---Indians were systematically excluded from top military ranks---fueled resentment and sharpened anti-colonial critiques even as wartime camaraderie forged cross-caste solidarities [2].
Politically, the Indian National Congress supported the British in World War I in hopes of self-government, only to pivot toward civil disobedience as imperial concessions failed to materialize [12].
During World War II, this disillusion deepened, with Congress opposition to India's unconsulted participation helping create the political space that Subhas Chandra Bose used to form the Indian National Army alongside the Japanese [2].
The INA drew heavily on Indian soldiers' wartime experiences and fought the British in Southeast Asia, with sources claiming that over half its members died in the struggle, a radicalization of military service into anti-colonial insurgency [39].
In 1946, the Royal Indian Navy mutiny transformed service grievances into an overt challenge to colonial rule across 78 ships and 21 shore establishments, catalyzing mass protests and forcing London to accelerate discussions on the transfer of power [40].
By early 1946, British authorities judged the RIAF unreliable for internal security and even contemplated European recruitment drives, a stark acknowledgment that the imperial military bargain in India was collapsing [41].

Rural India at war: land, credit, and the long economic shadow of imperial policy

- The wars for empire were financed not only by budgets but by villages, where commercialization of agriculture for export cash crops like cotton and indigo disrupted subsistence systems and made famine more likely when shocks hit [42].
- The Permanent Settlement fixed exorbitant revenue demands and empowered zamindars over tenants, intensifying extraction and constraining investment in land improvements critical for resilience [43].
- With formal finance largely closed to peasants, moneylenders charged usurious rates, embedding cycles of indebtedness that persisted well after independence and deepened with time [44].
- These pressures periodically erupted in rural unrest, as in the Deccan Riots of 1875 and the Sannyasi Rebellion after the 1770 Bengal famine, when peasants resisted exactions of cash and crops in a political economy reshaped by imperial war needs and trade [45].
- World War I magnified strains by diverting resources to the front and prompting fears of food shortages at home, a pattern repeated with deadlier consequences in World War II [46].

Recognition, memory, and the uneven afterlife of imperial service

- Despite roughly 1.5 million Indians in World War I and 2.5 million in World War II, official British recognition of their sacrifices was strikingly slow, taking 57 years to materialize in a significant public way [47].
- The Commonwealth Memorial Gates in London, inaugurated in 2002, now honor five million volunteers from the Indian subcontinent and other Commonwealth regions who fought in the world wars, a belated memorialization on imperial soil [47].
- Indian bravery was recognized during the wars: in World War I alone, Indians earned 11 Victoria Crosses and around 13,000 gallantry awards, marking their front-line valor despite systemic discrimination [48].
- By the 50th anniversary of the end of World War II in 1995, only 11 of 32 Indian Victoria Cross recipients were still alive, a reminder of how time outpaced commemoration for a fading generation [49].
- Partition further complicated remembrance in South Asia, fragmenting the shared military past and often overshadowing imperial service in nationalist narratives on both sides of the new border [50].
- Within Britain, contributions were acknowledged but seldom deeply appreciated, reflecting a memory gap between imperial reliance and post-imperial recognition [51].

Regional and communal dynamics within imperial service

- Recruitment and recognition were uneven, with regions such as Punjab and Garhwal contributing disproportionately and receiving greater public acknowledgment, a pattern that structured postwar social capital for veterans [52].
- Political alignments also shaped British attitudes, as the Muslim League's wartime support and Congress's opposition led some British officials to favor Muslim soldiers, complicating simple narratives of recruitment bias [53].

Putting it together: the many roles Indians played

- Indians fought the empire's wars as front-line soldiers, pilots, and sailors, and they staffed the logistical tail that made distant campaigning possible, from pack animals to munitions [3].
- They financed those wars through transfers and taxes that underwrote British strategy even as colonial economic policy drained Indian wealth and eroded food security on a civilizational scale [26].
- They also resisted---first as regional powers contesting British advance, later as mutineers, nationalists, and INA soldiers---until service under empire helped deliver the political leverage and organizational experience that made decolonization unstoppable [6].
- In this sense, Indians were at once the empire's sinew and its undoing, indispensable to Britain's war-making capacity and pivotal in ending imperial rule when the costs of that capacity became intolerable at home [41].

References

[1] Recognising The Contribution Of Indian Soldiers As We Mark Remembrance Day - asian-voice.com - https://www.asian-voice.com/Opinion/Columnists/Recognising-the-contribution-of-Indian-Soldiers-as-we-mark-Remembrance-Day

[2] 201 The Raj At War - poddtoppen.se - https://poddtoppen.se/podcast/1639561921/empire/201-the-raj-at-war

[3] One Should Not Forget The Rarely Mentioned But Colossal Contribution Made By India - blogs.lse.ac.uk - https://blogs.lse.ac.uk/southasia/2014/08/13/one-should-not-forget-the-rarely-mentioned-but-colossal-contribution-made-by-india/

[4] 767 - india-seminar.com - https://www.india-seminar.com/2023/767/767-07%20SARNATH%20BANERJEE.htm

[5] The Unveiling Of A Horror - codastory.com - https://www.codastory.com/rewriting-history/the-unveiling-of-a-horror/

[6] Indian National Army - yourstory.com - https://yourstory.com/2017/08/indian-national-army

[7] Indian Soldiers In World War I - newbooksnetwork.com - https

://newbooksnetwork.com/indian-soldiers-in-world-war-i
[8] Indians In World War 1 And 2 - rtvws.com - https://www.rtvws.com/p/indians-in-world-war-1-and-2
[9] Indians Soldiers In The First World War - thesamiksha.com - https://thesamiksha.com/indians-soldiers-in-the-first-world-war/
[10] 11182 - prio.org - https://www.prio.org/publications/11182
[11] Public Lectures - imfc.org - http://imfc.org/public-lectures/
[12] 12 - jaisonchacko.com - https://www.jaisonchacko.com/2016/12/
[13] Honoring The 80Th Anniversary Of D Day Revisiting British Indias Crucial Role - pakistaniindex.org - https://pakistaniindex.org/2024/06/honoring-the-80th-anniversary-of-d-day-revisiting-british-indias-crucial-role/
[14] Uk Honors Contribution Of Indian Soldiers In World War I And Ii - theindianeye.com - https://theindianeye.com/2024/11/12/uk-honors-contribution-of-indian-soldiers-in-world-war-i-and-ii/
[15] 73 - prio.org - https://www.prio.org/publications/73
[16] Troop Ships Rally Round The Flag - poheritage.com - https://www.poheritage.com/the-collection/exhibitions/our-war-at-sea/troop-ships-rally-round-the-flag
[17] Africa.Aspx - india1914.com - https://www.india1914.com/africa.aspx
[18] Royal Indian Air Force - rafmuseum.org.uk - https://www.rafmuseum.org.uk/blog/royal-indian-air-force/
[19] 82Nd Anniversary Of The Indian Air Force 8 October 1932 - lenseye.co - https://www.lenseye.co/82nd-anniversary-of-the-indian-air-force-8-october-1932/
[20] Avalanchepress.Com - http://www.avalanchepress.com/SwordSeaIndia2.php
[21] 8773 Revisiting The Royal Indian Navy Mutiny Of 1946 - mintageworld.com - https://www.mintageworld.com/media/detail/8773-revisiting-the-royal-indian-navy-mutiny-of-1946/
[22] Remembering Indian Naval Mutiny 1946 0 - liberation.org.in - https://liberation.org.in/index.php/liberation-2022-february/remembering-indian-naval-mutiny-1946-0
[23] Indian Air Force Supported Azad Hind Fauj Soldiers - alhaqeeqa.org - https://www.alhaqeeqa.org/indian-air-force-supported-azad-hind-fauj-soldiers/
[24] Mobilising Empires War - peacenews.info - https://peacenews.info/node/7576/mobilising-empires-war
[25] Money Finance - ideasforindia.in - http://www.ideasforindia.in/topics/money-finance/trade-disruption-industrialisation-and-the-setting-sun-of-british-colonial-rule-in-india.html
[26] Explained How The British Empire Looted 45 Trillion From India - pratapdarpan.in - https://pratapdarpan.in/explained-how-the-british-empire-looted-45-trillion-from-india/
[27] Imperialism How Declining Currency Hegemony Leads To War - ppesydney.net - https://www.ppesydney.net/imperialism-how-declining-currency-hegemony-leads-to-war/
[28] Comment On The Economic Impact Of The British Rule - theprayagraj.in - https://theprayagraj.in/comment-on-the-

economic-impact-of-the-british-rule/
[29] Inglorious Empire What The British Di In India - iqra.co.za - https://www.iqra.co.za/products/inglorious-empire-what-the-british-di-in-india
[30] Utilitarian Free Trade Killed Millions In China And India - consumer.org.my - https://consumer.org.my/utilitarian-free-trade-killed-millions-in-china-and-india/
[31] British Colonialism Killed 100 Million Indians In 40 Years - orinocotribune.com - https://orinocotribune.com/british-colonialism-killed-100-million-indians-in-40-years/
[32] First Anglo Mysore War - adda247.com - https://www.adda247.com/upsc-exam/first-anglo-mysore-war/
[33] Anglomysore Wars - indianetzone.com - https://www.indianetzone.com/anglomysore_wars
[34] Second Anglo Maratha War - dronaias.in - https://dronaias.in/second-anglo-maratha-war/
[35] Anglo Maratha Wars - nextias.com - https://www.nextias.com/blog/anglo-maratha-wars/
[36] Anglo Maratha Wars Overview - litgrades.com - https://litgrades.com/deck/anglo-maratha-wars-overview
[37] World War I And Devastation Of Bharat Ii - hinduinfopedia.org - https://hinduinfopedia.org/world-war-i-and-devastation-of-bharat-ii/
[38] Indian Army Type Team - paperjewels.org - https://www.paperjewels.org/postcard/indian-army-type-team
[39] ?F - historynet.com - https://www.historynet.com/indian-national-army1942-45/?f
[40] 55358 - ibpbooks.com - https://www.ibpbooks.com/1946-royal-indian-navy-mutiny-last-war-of-independence/p/55358
[41] Serving The Nation A Legacy Of Valor And Resilience By Indias Armed Force - frontierindia.com - https://frontierindia.com/serving-the-nation-a-legacy-of-valor-and-resilience-by-indias-armed-force/
[42] Economical Impact Of British Colonial Rule In India - litgrades.com - https://litgrades.com/deck/economical-impact-of-british-colonial-rule-in-india
[43] Mains Question 05 01 2023 Gs1 Land Revenue Policy - legacyias.com - https://www.legacyias.com/daily-answer-writing/mains-question-05-01-2023-gs1-land-revenue-policy/
[44] Credit Risk In Colonial India - ehs.org.uk - https://ehs.org.uk/credit-risk-in-colonial-india/
[45] Peasants And Tribals - muidolstudysquad.com - https://muidolstudysquad.com/peasants-and-tribals/
[46] 115144438.Cms - economictimes.indiatimes.com - https://economictimes.indiatimes.com/magazines/panache/no-beef-no-pork-ginger-garlic-and-chapatis-mandatory-a-look-at-indian-troops-diet-during-wwi/articleshow/115144438.cms
[47] Review Indias War World War Ii And The Making Of Modern South Asia - thediplomat.com - https://thediplomat.com/2016/03/review-indias-war-world-war-ii-and-the-making-of-modern-south-asia/

[48] Remembrance Sunday Commemorating Sacrifice In Art - asianculturevulture.com - https://asianculturevulture.com/portfolios/remembrance-sunday-commemorating-sacrifice-in-art/
[49] Respecting A Victoria Cross - salute.co.in - https://salute.co.in/respecting-a-victoria-cross/
[50] Britain Today And The Fight For Recognition - britishlegion.org.uk - https://www.britishlegion.org.uk/get-involved/remembrance/remembrance-events/vj-day/the-impact-the-aftermath-and-the-future/britain-today-and-the-fight-for-recognition
[51] A Punjabi Muslim Cavalryman From The British Indian Army Hands Rations To Starving Christian Women In Iraq During World War I 976X1437 - goodizen.com - https://goodizen.com/a-punjabi-muslim-cavalryman-from-the-british-indian-army-hands-rations-to-starving-christian-women-in-iraq-during-world-war-i-976x1437/
[52] Why Do So Many People From Punjab Garhwal Join The Indian Armed Forces - edtimes.in - https://edtimes.in/why-do-so-many-people-from-punjab-garhwal-join-the-indian-armed-forces/
[53] Indians As Officers In The British Indian Army - brownpundits.com - https://www.brownpundits.com/2018/12/01/indians-as-officers-in-the-british-indian-army/

Appendix E

Algorithms

Algorithm 1 Citation Quality Evaluation. For each substantive sentence s in report R, we extract citation IDs C and use a NLI model to verify if cited passages ($\mathcal{D}_C$) support the claim. We count: N_{claims} (total claims), N_{supp} (supported claims with valid citations), N_{cites} (total citations used), and N_{nec} (necessary citations).

Input: Report R, Retrieved Passages $D = \{d_1, \ldots, d_k\}$
$N_{claims}, N_{supp}, N_{cites}, N_{nec} \leftarrow 0, 0, 0, 0$

for each sentence $s \in R$ with length$(s) \geq 50$ **do**
    $N_{claims} \leftarrow N_{claims} + 1$
    $C \leftarrow$ ExtractCitations(s)
    if $C \neq \emptyset$ **and** NLI$(s, \text{Concat}(\mathcal{D}_C))$ **then**
        $N_{supp} \leftarrow N_{supp} + 1$
        $N_{cites} \leftarrow N_{cites} + |C|$
        for each $c \in C$ **do**
            if NLI$(s, \mathcal{D}_C)$ **or** $\neg$NLI$(s, \text{Concat}(D_{C \setminus \{c\}}))$ **then**
                $N_{nec} \leftarrow N_{nec} + 1$

Output: Precision $P = N_{nec}/N_{cites}$, Recall $R = N_{supp}/N_{claims}$, F1 $= 2PR/(P + R)$

Appendix F

Prompts

This appendix contains the main system prompts for each agent in the research pipeline. All prompts use Jinja2 templating with variables denoted as `{{ variable_name }}`.

F.1 Query Generator Agent

Agent Class: `QueryGeneratorSubAgentV2`

```
# Task Overview
You are tasked with creating research plans for comprehensive, high-quality
reports. Your goal is to generate diverse, targeted search queries that will
retrieve the most relevant and complete information to answer the user's
research question.

{{ domain_prompt }}

{% if dataset_specific_instructions %}
# Dataset-Specific Guidance
{{ dataset_specific_instructions }}
{% endif %}

# Key Principles (Prioritized)
1. **Direct Relevance**: Each query must directly address the user's
   research question
2. **Temporal Matching**: Match the temporal focus to the user's question
   (current vs historical) - If a date is mentioned, use it for your
   extraction instructions.
3. **Focused Coverage**: Cover aspects that directly contribute to answering
   the user's question
4. **Query Diversity**: Within relevant queries, target different aspects
   to avoid redundancy
5. **Strategic Depth**: Balance broad exploratory queries with specific
   targeted ones when both are relevant

# Search Engine Capabilities
The search system uses advanced semantic search with the following
characteristics:
- Support for natural language queries
```

```
- Date range filtering capabilities

# Your Workflow
1. **Analyze User Intent**: What is the user really asking? (causal,
   descriptive, comparative, evaluative)
   - Does the user want current/recent information or historical perspective?
   - What temporal scope matches their question? (e.g., "why is X high" →
     recent factors, not decade-long trends)
2. **Identify Core Need**: What specific information would best answer
   their question?
3. **Plan Focused Strategy**: Design queries targeting aspects that directly
   address the user's need
4. **Generate Relevant Queries**: Create specific queries that explore
   different relevant angles
5. **Specify Extraction Instructions**: Tailor instructions to each query's
   purpose
6. **Consider Search Depth**: Decide exhaustive vs. quick search based on
   query importance

# Query Generation Strategy
You will generate {{ candidate_queries_needed }} candidate queries from which
a selection algorithm will choose the most relevant AND diverse subset.
Therefore:
- Focus on relevance first - ensure each candidate directly addresses the
  user's question from different angles
- Don't artificially force diversity - natural variation across relevant
  angles is sufficient
- Quality over quantity - avoid tangential queries
- Each query should contribute to answering the ORIGINAL question

# Important Notes
- Avoid tangential exploration, even if academically interesting
- Temporal scope matters: "why is X high NOW"  "history of X over decades"
- Each query should have results that will be used in the final answer

# Constraints
- Maximum {{ max_turns }} turns to complete the research
- Maximum {{ max_tasks_per_turn }} tasks per turn
- Focus on the core research question without tangential exploration

Today's date: {{ today_date }}
```

F.2 Information Extractor Agent

Agent Class: `InformationExtractorSubAgent`

```
{{ domain_prompt }}

## CORE EXTRACTION PRINCIPLES

1. **Relevance Filter First**: Skip documents that are clearly irrelevant
```

```
   to the search query. If a document doesn't meaningfully address the query,
   output nothing from it.

2. **Fact Granularity**: Extract specific details, not generalizations.

3. **Single Citation Rule**: One citation per distinct fact. If multiple
   documents state the same fact, choose the most authoritative source and
   extract once.

4. **Contradictions Are Valuable**: When sources disagree, extract BOTH
   perspectives.

## WHAT TO EXTRACT

- Specific numbers, percentages, statistics
- Dates, timeframes, locations
- Named entities (organizations, researchers, products)
- Methodologies and sample sizes
- Causal relationships with evidence
- Contradictory findings between sources
- Limitations and caveats mentioned in sources

## WHAT TO SKIP

- Generic statements without specific details
- Information unrelated to the search query
- Marketing language or unsubstantiated claims
- Redundant facts already extracted from better sources

## CITATION FORMAT

{% if citation_style == 'ragtime' %}
Use the EXACT 8-character ID shown for each search result in square brackets
(e.g., [abc12345]).
{% elif citation_style in ['pes2o', 'fineweb'] %}
Use the search result's ID in square brackets (e.g., [1e0e0c11]).
{% else %}
Use the search result's ID in square brackets (e.g., [ad2e4f3b]).
{% endif %}

- Place citations IMMEDIATELY after the specific fact they support
- NEVER create citation IDs that don't exist in the search results
- Each fact must have exactly ONE citation

## OUTPUT FORMAT

If you find relevant facts in ANY document from this batch:
- Extract each fact as a complete sentence with its citation
- Separate facts with single newlines (one line break between facts)
- Do not organize into themes, sections, or create structure
- Simply output facts with citations in the order you encounter them
```

```
Example format:
First fact extracted from document [abc12345].
Second fact from same document [abc12345].
Third fact from a different document [def67890].

If NONE of the documents in this batch contain relevant facts:
- Output exactly: NO_RELEVANT_FACTS
- Do not add any other text, explanations, or formatting

Today's date is {{ today_date }}.
```

F.3 Information Merger Agent

Agent Class: `InformationMergerSubAgent`

```
{{ domain_prompt }}
You are an expert research synthesizer creating comprehensive literature
review-style summaries.

## CORE PRINCIPLE: EXTEND, DON'T REDUCE

When combining related facts, ADD details together rather than generalizing
them.

WRONG: "Prices rose significantly [5a7b0fd1_1][4b5d48b_2][3c9e2af0_3]"
RIGHT: "Prices rose 10% in CA [5a7b0fd1_1], 15% in TX [4b5d48b_2], peaked
at $4.50 nationally [3c9e2af0_3]"

## KEY RULES

**Citation-Claim Binding:** Each citation stays with its specific fact
- "Study A: 30% increase [5a7b0fd1_1] while Study B: 25% [4b5d48b_2]"
- NOT: "Studies found 25-30% increases [5a7b0fd1_1][4b5d48b_2]"

**Smart Citation Management:**
- Identical facts: Keep 2-3 most authoritative (gov data > peer-reviewed >
  institutions > recent)
- Similar facts: Keep all citations with their specific claims
- Incompatible facts: Present side-by-side as differing perspectives

**Fact Relationships:**
- Sequential: "X in 2020 [5a7b0fd1_1], then Y in 2022 [4b5d48b_2]"
- Comparative: "CA: 10% [5a7b0fd1_1], TX: 15% [4b5d48b_2]"
- Additive: "Factors: supply drop [5a7b0fd1_1], demand surge [4b5d48b_2],
  refinery issues [3c9e2af0_3]"

**Preserve Specificity:** Keep ALL numbers, dates, locations, examples.
Never generalize specifics into ranges.

## CITATION FORMAT RULES
```

```
- Use EXACT citation IDs: [a1b2c3d4_1] - never modify
- Place immediately after relevant clause
- Never create new citation IDs

Organize thematically while maintaining fact density. If no information
provided, return "No information found".

## OUTPUT REQUIREMENTS
**NO SUMMARY SECTIONS**: Never end with "In summary", "Summary:", or
concluding paragraphs.

Today's date is {{ today_date }}.
```

F.4 Follow-up Enrichment Agent

Agent Class: `FollowUpEnrichmentSubAgent`

```
{{ domain_prompt }}

{% if dataset_specific_instructions %}
# Dataset-Specific Guidance
{{ dataset_specific_instructions }}
{% endif %}

You are a research completeness analyst. Your job is to identify strategic
gaps within the specific research area "{{ query }}" to ensure comprehensive
coverage.

## MISSION
Analyze your research area and identify **0-3 meaningful gaps** that would
significantly improve completeness. Focus on YOUR topic, not the overall
question.

## COMPLETENESS ASSESSMENT
**Today's Date:** {{ today_date }}

### Step 1: Gap Identification (Only if Meaningful)
Look for strategic gaps within YOUR research area

### Step 2: Targeted Enrichment (If Gaps Found)
For each valid gap, create ONE specific search query:
- **Specificity**: Target exact aspect missing (not broad rehash)
- **Searchability**: Use concrete terms likely in documents
- **Value**: High chance of substantial new information
- **Write relatively short and simpler queries that are easy to understand
  and execute, be more specific in the extraction instructions.**

## RESEARCH STATUS
**Overall Context:** {{ original_question }}
**Your Focus Area:** {{ query }}
```

```
{{ merged_information }}

{% if completed_queries %}
## PARALLEL RESEARCH
{% for completed_query in completed_queries %}
- "{{ completed_query }}"
{% endfor %}
Stay within your research area only.
{% endif %}

{% if round_num > 1 %}
## ROUND {{ round_num }} - REFINE, DON'T REPEAT
Previous round targeted gaps but some may persist. If so:
- Generate MORE SPECIFIC queries (countries, mechanisms, timeframes,
  stakeholders)
**Refinement pattern**: "policy impact" → "California EV policy impact 2024"
{% endif %}

## DECISION LOGIC
**STOP enrichment** if:
- Topic area is thoroughly covered from key angles
- Additional searches would likely be redundant
- No meaningful gaps remain within your scope

**CONTINUE enrichment** if:
- Clear strategic gaps exist within your research area
- Missing information would meaningfully improve topic completeness

## OUTPUT FORMAT (JSON)
**If meaningful gaps exist**: Generate EXACTLY {{ max_candidate_queries }}
enrichment tasks.

**If NO meaningful gaps exist**: Return empty tasks array.

**Final Check**: Only propose enrichments that would add meaningful,
substantial content to your specific research area.
```

F.5 Answer Writer Agent

Agent Class: `AnswerWriterSubAgent`

```
{{ domain_prompt }}

{% if answer_format_instructions %}
{{ answer_format_instructions }}
{% else %}

**PRIMARY OBJECTIVE**: The user's question is the central focus - everything
you write should serve to answer it thoroughly.

You will be provided with research information from multiple queries. Each
```

```
query result represents a key aspect of the user's question.

# SYNTHESIS STRATEGY

Your task is to synthesize the research information into a comprehensive,
well-structured answer that addresses the user's question thoroughly. Aim
for depth and breadth while maintaining clarity and coherence.

**Central Thesis Focus**: Formulate a clear answer to the user's question,
then build your narrative to support it

**Integrated Paragraphs**: Each paragraph should weave together insights
from MULTIPLE research chunks - avoid presenting each source in isolation

**Progressive Revelation**: Start with direct answers, then reveal deeper
evidence and systemic factors

**Smooth Narrative Arc**: Build from simple to complex, showing how factors
interact and amplify each other

# WRITING GUIDELINES

## Content & Organization
- **Opening**: Start with a ## heading that relates to the user's question,
  then provide a clear, substantive answer that sets up your main arguments
- **Body sections**: Organize your response thematically, ensuring each
  major finding or perspective from the research gets appropriate coverage
- **Detail level**: Include specific information where it strengthens your
  answer - researcher names, case studies, quantitative findings, concrete
  examples
- **Evidence approach**: Support your points with specific details and
  citations from the research - concrete evidence strengthens your answer
- **Balance**: Present diverse viewpoints and findings, including
  contradictions or debates when they exist in the research
- **Synthesis over summary**: Weave together insights rather than listing
  findings sequentially

## Citation Rules - CRITICAL FOR ACCURACY

Your research data contains citations in this format: [5a7b0fd1_1],
[4b5d48b_2], [3c9e2af0_3].

**Core Rules**:
- **Copy citation IDs exactly** as they appear - do not modify, renumber,
  or invent them
- **Citation placement** - Citations MUST appear INSIDE the sentence,
  attached directly to the claim and BEFORE the period: "Claim [citation]."
  NOT "Claim. [citation]"
- **Present different findings in separate sentences** - each distinct
  claim gets its own sentence with its own citation
- **NEVER add the same citation two times in a row at the end of a
```

```
  sentence** - this is completely wrong

**Citation Coverage**:
- Cite: Statistics, percentages, research findings, quantitative data,
  specific examples
- Don't cite: Your analysis, general framing, transitions between ideas

## Markdown Formatting
- **Main heading**: Start with a ## heading that relates to the user's
  question
- **Section headers**: Use ### headings to organize major sections or
  topics in your answer
- **Bold text**: Use **bold** to emphasize key concepts within paragraphs
  (not as a replacement for headers)
- **Paragraph-first structure**: Write in flowing narrative paragraphs as
  your primary format - this is a research story, not a list
- **Bullets sparingly**: Use bullet points only when listing discrete
  examples, data comparisons, or specific cases where a list genuinely
  aids clarity - not as your default organizational structure

# IMPORTANT REMINDERS
- **Research utilization**: Draw from the full range of research provided
  to create a rich, multi-faceted answer
- **Specificity matters**: Include concrete details, names, numbers, and
  examples where they add value to your explanation
- **Natural synthesis**: Weave findings together in a way that tells a
  coherent story while respecting the complexity of the topic
- **Appropriate depth**: Match your answer's detail level to the richness
  of the research - more comprehensive research warrants a more detailed
  response

{% endif %}

Today's date is {{ today_date }}.
```

Bibliography

[1] Shubham Agarwal, Issam H Laradji, Laurent Charlin, and Christopher Pal. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*, 2024.

[2] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA, 2009. ACM.

[3] Jina AI. Node-deepresearch. https://github.com/jina-ai/node-DeepResearch, 2024. GitHub repository. Implements iterative query-search-read-reason loops for deep research workflows. Accessed: 2025-10-12.

[4] Salaheddin Alzubi, Creston Brooks, Purva Chiniya, Edoardo Contente, Chiara von Gerlach, Lucas Irwin, Yihan Jiang, Arda Kaz, Windsor Nguyen, Sewoong Oh, Himanshu Tyagi, and Pramod Viswanath. Open deep search: Democratizing search with open-source reasoning agents. *arXiv preprint arXiv:2503.20201*, 2025.

[5] Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*, 2024.

[6] Azin Ashkan, Branislav Kveton, Shlomo Berkovsky, and Zheng Wen. Optimal greedy diversity for recommendation. In *IJCAI*, volume 15, pages 1742–1748, 2015.

[7] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[9] F. Chollet. On the measure of intelligence. Technical Report / ARC-AGI / ARC Prize Foundation, 2019.

[10] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson,

Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Matthew D. Hoffman, Zoubin Ghahramani, Jakob Uszkoreit, Henryk Michalewski, Noam Shazeer, Patrick Li, Ed Chi, Quoc V. Le, and Denny Zhou. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[11] João Coelho, Jingjie Ning, Jingyuan He, Kangrui Mao, Abhijay Paladugu, Pranav Setlur, Jiahe Jin, Jamie Callan, João Magalhães, Bruno Martins, and Chenyan Xiong. Deepresearchgym: A free, transparent, and reproducible evaluation sandbox for deep research, 2025.

[12] Gordon V. Cormack and Maura R. Grossman. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, page 153–162, New York, NY, USA, 2014. Association for Computing Machinery.

[13] Gérard Cornuéjols, Marshall L. Fisher, and George L. Nemhauser. Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management Science*, 23(8):789–810, 1977.

[14] DanielWalnut. Deerflow. https://github.com/bytedance/deer-flow, 2025. Accessed: 2025-05-28.

[15] Zhenyun Deng, Yonghua Zhu, Yang Chen, Michael Witbrock, and Patricia Riddle. Interpretable amr-based question decomposition for multi-hop question answering. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

[16] Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. Successive prompting for decomposing complex questions. *arXiv preprint arXiv:2212.04092*, 2022.

[17] Assaf Elovic. Gpt-researcher. https://github.com/assafelovic/gpt-researcher, 2024. GitHub repository. Multi-stage agentic research pipeline for automated report generation. Accessed: 2025-10-12.

[18] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[19] Google. Gemini deep research: Your personal research assistant, 2025. Accessed: 2025-04-12.

[20] Maura R. Grossman and Gordon V. Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law & Technology*, 17(3):11–?, 2011. Accessed via Richmond J.L. & Tech. :contentReference[oaicite:0]index=0.

[21] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Yuanzhuo Wang, Lionel Ni, Wen Gao, and Jian Guo. A survey on llm-as-a-judge, 2024.

[22] Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, et al. Pasa: An llm agent for comprehensive academic paper search. *arXiv preprint arXiv:2501.10120*, 2025.

[23] Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (mcp): Landscape, security threats, and future research directions, 2025.

[24] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zhi Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Weilin Zhao, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

[25] Xiang Huang, Sitao Cheng, Yiheng Shu, Yuheng Bao, and Yuzhong Qu. Question decomposition tree for answering complex questions over knowledge bases. *arXiv preprint arXiv:2306.07597*, 2023.

[26] Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, et al. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*, 2025.

[27] Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, Jianye Hao, Kun Shao, and Jun Wang. Deep research agents: A systematic examination and roadmap, 2025.

[28] HuggingFace. Open deep research, 2025.

[29] Gautier Izacard, Edouard Grave, and Armand Joulin. Contriever: Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2022.

[30] Suhas Jayaram Subramanya, Fnu Devvrit, Harsha Vardhan Simhadri, Ravishankar Krishnawamy, and Rohan Kadekodi. Diskann: Fast accurate billion-point nearest neighbor search on a single node. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[31] Zhouyu Jiang, Mengshu Sun, Lei Liang, and Zhiqiang Zhang. Retrieve, summarize, plan: Advancing multi-hop question answering with an iterative approach. *arXiv preprint arXiv:2407.13101*, 2024.

[32] Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

[33] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China, 2019. Association for Computational Linguistics.

[34] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

[35] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48, 2020.

[36] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models, 2024.

[37] Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. Qasa: Advanced question answering on scientific articles. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19036–19052. PMLR, 2023.

[38] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[39] Alicia Li, Zhuohan Zhang, Mo Yu, Byron C. Wallace, and Hannaneh Hajishirzi. KIWI: A dataset of knowledge-intensive writing instructions for answering research questions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12969–12990. Association for Computational Linguistics, August 2024.

[40] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025.

[41] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan Weld. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, 2020.

[42] Iain J. Marshall and Byron C. Wallace. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8(1):163, 2019. Comprehensive practical guide for ML methods in systematic reviews.

[43] G. Mialon et al. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023.

[44] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978.

[45] OpenAI. Deep research system card, 2025. Accessed: 2025-04-12.

[46] OpenAI. Defining and evaluating political bias in llms. Blog post, October 9 2025.

[47] OpenAI. Introducing gpt-4.1 in the api, April 14 2025. Accessed: 2025-10-13.

[48] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[49] Arnold Overwijk, Chenyan Xiong, Xiao Liu, Cameron VandenBerg, and Jamie Callan. Clueweb22: 10 billion web documents with visual and semantic information, 2022.

[50] Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. Unsupervised question decomposition for question answering. *arXiv preprint arXiv:2002.09758*, 2020.

[51] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Summer Yue, Alexandr Wang, Dan Hendrycks, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

[52] Haosheng Qian, Yixing Fan, Jiafeng Guo, Ruqing Zhang, Qi Chen, Dawei Yin, and Xueqi Cheng. Vericite: Towards reliable citations in retrieval-augmented generation via rigorous verification. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '25, pages 1–8, New York, NY, USA, 2025. ACM.

[53] Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. Tool learning with large language models: A survey. *Frontiers of Computer Science*, 19(8):198343, 2025.

[54] Adam Roegiest, Gordon V. Cormack, Maura R. Grossman, and Charles L. A. Clarke. Trec 2015 total recall track overview. In *Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015)*. National Institute of Standards and Technology (NIST), 2015. First standardized evaluation framework for high recall retrieval.

[55] Corby Rosset, Ho-Lam Chung, Guanghui Qin, Ethan C. Chau, Zhuo Feng, Ahmed Awadallah, Jennifer Neville, and Nikhil Rao. Researchy questions: A dataset of multi-perspective, decompositional questions for llm web agents, 2024.

[56] Nick Scamara. Open-deep-research. https://github.com/nickscamara/open-deep-research, 2024. GitHub repository. Open implementation of autonomous deep research with time-based search budgeting. Accessed: 2025-10-12.

[57] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025.

[58] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse,

Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2025.

[59] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, 2021.

[60] Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*, 2025.

[61] Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

[62] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*, 2024.

[63] Justin JongSu Song, Wookey Lee, and Jafar Afshar. An effective high recall retrieval method. *Data & Knowledge Engineering*, 123:101603, 2019.

[64] SURF. Snellius supercomputer, 2023. Available at https://www.surf.nl/en/services/snellius-supercomputer.

[65] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online, 2020. Association for Computational Linguistics.

[66] Sean Westwood, Justin Grimmer, and Andrew B. Hall. Measuring perceived slant in large language models through user evaluations. Hoover Institution blog / commentary, May 8 2025.

[67] Han Xiao. Submodular optimization for diverse query generation in deepresearch. Jina AI Tech Blog, July 4 2025.

[68] Zhaorui Yang, Bo Pan, Han Wang, Yiyao Wang, Xingyu Liu, Minfeng Zhu, Bo Zhang, and Wei Chen. Multimodal deepresearcher: Generating text-chart interleaved reports from scratch with agentic framework. *arXiv preprint arXiv:2506.02454*, 2025.

[69] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2369–2380, 2018.

[70] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

[71] Peng Yixing, Quan Wang, Licheng Zhang, Yi Liu, and Zhendong Mao. Chain-of-question: A progressive question decomposition approach for complex knowledge base question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

[72] Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13909–13926, Singapore, December 2023. Association for Computational Linguistics.

[73] Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, and Michael Bendersky. Inference scaling for long-context retrieval augmented generation. *arXiv preprint arXiv:2410.04343*, 2025.

[74] Daniel Zhang. Deep-research. https://github.com/dzhng/deep-research, 2024. GitHub repository. Implements user-controlled search depth and breadth for autonomous web-based research. Accessed: 2025-10-12.

[75] Jiajie Zhang, Shulin Cao, Tingjian Zhang, Xin Lv, Jiaxin Shi, Qi Tian, Juanzi Li, and Lei Hou. Reasoning over hierarchical question decomposition tree for explainable question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[76] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.