Justification Retrieval with LLMs, Retrieval-Augmented Generation, and Hybrid Labels

Georgios Arampatzis*
Vasileios Perifanis
Avi Arampatzis

Democritus University of Thrace, Dept. of Electrical and Computer Engr., Xanthi, Greece

{geoaramp,vperifan,avi}@ee.duth.gr

Abstract

This paper presents a hybrid justification labeling framework for retrieval-augmented generation (RAG), focusing exclusively on the Relevance Judgment (RJ) subtask of the TREC 2025 RAG Track. The proposed approach integrates open-weight large language models (LLMs) with traditional retrieval signals and confidence calibration mechanisms. Specifically, we combine deterministic pre-labeling using **Qwen2.5-3B-Instruct** and **StableLM 2-1 6B-Chat** with multi-stage confidence normalization and lexical-overlap heuristics. This design enables small open-weight models to approximate the reasoning behavior of larger proprietary systems while remaining transparent and fully reproducible.

We describe the end-to-end pipeline used for both automatic and semi-manual relevance judgment runs, analyze their validation consistency, and examine the impact of calibration parameters on justification quality and coverage. Empirical results indicate that hybrid confidence blending improves mid-range justification reliability and reduces variance across topics. All runs were validated using the official evaluation infrastructure and correspond solely to submissions for the RAG Relevance Judgment task.

Keywords: RAG Relevance Judgment; justification retrieval; relevance judgment; hybrid annotation; open-weight large language models; confidence calibration

1 Introduction

Retrieval-Augmented Generation (RAG) systems have gained significant attention as a paradigm that combines information retrieval with generative reasoning to enhance factual accuracy and explainability. Within this context, the **TREC 2025 RAG Track** includes a dedicated *Relevance Judgment (RJ)* subtask, which focuses on the evaluation of justification retrieval — that is, the identification and grading of supporting passages that justify an answer or stance associated with a given query. This task is inherently challenging, as it requires both lexical and semantic alignment between queries and evidence passages, along with fine-grained relevance and justification-level judgments. In this work, we focus exclusively on the Relevance Judgment (RJ) subtask, which is the only task of the TREC 2025 RAG Track for which we submitted official runs.

Traditional information retrieval benchmarks (e.g., [1, 2]) have relied on binary or graded relevance labels, often derived from manual annotation. However, justification retrieval extends this paradigm by demanding explanatory relevance, where evidence must support specific claims or answer components rather than mere topical similarity. This substantially increases the cognitive load for annotators and makes large-scale manual labeling expensive and inconsistent [3].

To address these challenges, we present **DUTH-RJ**, a hybrid system developed at the Democritus University of Thrace for the TREC 2025 RAG Relevance Judgment task. Our approach integrates automatic pre-labeling and manual verification in a unified pipeline that ensures both scalability and quality control. The system uses top-k retrieved passages from upstream retrievers as input, prepares a structured labeling sheet, and supports incremental manual review. Each retrieved document is graded on a 0–4 justification scale, reflecting increasing degrees of answer completeness and evidential quality.

The hybrid process enables rapid annotation while maintaining interpretability and alignment with the objectives of the RJ subtask. Sanity checks and distributional analyses ensure that each query receives consistent coverage across the top-ranked passages, producing a balanced set of 2,100 labeled query–passage pairs over 105 topics.

*Corresponding author.

This work contributes a reproducible and scalable approach for justification-level relevance assessment within the TREC 2025 RAG framework. The remainder of this paper is organized as follows. Section 2 reviews prior research on relevance judgment, justification retrieval, and hybrid annotation methodologies. Section 3 presents the datasets, models, and end-to-end methodological design used in both automatic and semi-manual submissions for the TREC 2025 RAG Relevance Judgment task. Section 4 presents the official NIST evaluation results, accompanied by a detailed analysis of calibration effects, confidence-blending strategies, and model-dependent variation in justification reliability. Finally, Section 5 summarizes the main findings and outlines directions for future work.

2 Related Work

Research on retrieval-augmented generation (RAG) and justification retrieval builds upon long-standing developments in information retrieval (IR) evaluation and neural retrieval architectures. Early frameworks such as the Probabilistic Relevance Model and BM25 established a foundation for ranking based on term weighting and document frequency [1]. These approaches, operationalized through large-scale benchmarks like TREC [3], demonstrated the value of standardized evaluation and manual relevance judgments in advancing retrieval performance.

With the emergence of deep learning, representation-based retrieval models began to dominate large-scale retrieval tasks. Transformer-based architectures such as Dense Passage Retrieval (DPR) and ColBERT enabled semantic matching at scale, improving upon purely lexical models. The TREC Deep Learning Track [2] formalized this transition by incorporating large, pre-trained neural retrievers evaluated over MS MARCO-style datasets, thus bridging traditional IR evaluation with neural relevance modeling.

The introduction of Retrieval-Augmented Generation (RAG) extended these ideas beyond retrieval into reasoning and synthesis. RAG frameworks combine retrievers with sequence-to-sequence generators, allowing retrieved evidence to inform generative output [4]. These systems have been widely adopted in open-domain question answering, summarization, and dialogue, yet their evaluation remains complex: retrieved evidence must not only be relevant, but also *justificatory*—that is, supporting specific claims within generated text. This has led to a growing interest in justification-level labeling, as formalized in recent shared tasks such as TREC RAG (2024–2025).

Parallel research has explored semi-automatic and hybrid annotation pipelines to address the limitations of purely manual relevance judgments. Semi-supervised labeling approaches leverage model-based pre-labeling followed by human verification to increase both scalability and consistency. This paradigm has been effective in the context of document relevance, claim verification, and evidence retrieval, suggesting a promising path for justification-level datasets. The hybrid approach implemented in DUTH-RJ follows this line of research by combining automatic pre-labeling with manual review, while preserving the interpretability and fine-grained control characteristic of human annotation.

Beyond prior studies in justification and retrieval-augmented generation, our work builds upon the DUTH team's broader experience in multilingual NLP and hybrid evaluation pipelines. In previous shared-task participations, including SemEval–2023 [5], EXIST–2025 [6], JOKER–2025 [7], and SimpleText–2025 [8], we developed modular transformer-based systems for affective analysis, bias detection, and text simplification. These efforts provided methodological insights—particularly regarding soft-label calibration and interpretability—that directly informed the design of the hybrid justification labeling framework presented here.

Overall, our work contributes to this growing body of research by operationalizing a scalable justification retrieval pipeline under the TREC RAG 2025 framework, combining the reproducibility of standardized IR benchmarks with the depth of semantic evaluation.

3 System Description

3.1 Dataset

Our experiments are conducted using the corpus and topics released for the **TREC 2025 Retrieval-Augmented Generation (RAG) Track**, organized by the U.S. National Institute of Standards and Technology (NIST). The RAG Track builds upon the *MS MARCO Document v2.1* collection, segmented into passages that serve as atomic retrieval units. Each segment contains metadata fields including `docid`, `url`, `title`, `segment`, and character offsets, allowing fine-grained evidence retrieval.

The **Justification Retrieval (RJ) Subtask** focuses on identifying passages that provide factual or argumentative justification for an information-seeking query. The official test topics consist of **105 narratives**, each requiring the retrieval of supporting evidence rather than purely relevant content. The queries were provided in JSONL format, following the schema {`qid, text`}, and reflect diverse social, legal, scientific, and policy-oriented questions.

Each query was paired with up to 100 candidate segments retrieved by the BM25 baseline implemented via Anserini [9]. From this pool, the top 20 passages per topic were selected for labeling, resulting in a total of **2,100 query–passage pairs**. Each passage was annotated using a five-point justification scale: **0** for non-relevant, **1** for marginal relevance, **2** for partial justification, **3** for strong justification, and **4** for full justification of the query intent.

An optional confidence value (0–1) was also assigned to capture annotator certainty.

Labeling followed a **hybrid human–machine approach**: an initial pre-labeling stage was conducted automatically using a large language model, followed by targeted human review to correct or confirm uncertain predictions. This hybrid workflow produced consistency comparable to human annotation while significantly reducing manual effort, similar to prior semi-automatic labeling pipelines [10, 11]. The resulting annotations were exported in TREC run format (`qid, Q0, docid, rank, score, runid`) for evaluation under official track metrics, including NDCG@10 and Justification Accuracy.

Finally, integrity checks were performed to ensure that all 105 queries were represented by exactly 20 labeled passages. The final dataset (`r_output_trec_rag_2025.tsv`) contained **2100 entries** and was verified using the `rj_pipeline.py check` utility to confirm completeness and label validity. This configuration ensures full compatibility with the TREC RAG 2025 infrastructure and supports reproducible evaluation across retrieval and generation subtasks.

3.2 Models

The submitted systems explore multiple configurations of open-weight language models and hybrid calibration strategies within the TREC 2025 RAG Relevance Judgment (RJ) subtask. Our goal was to systematically compare small- to medium-scale instruction-tuned LLMs under identical retrieval and evaluation settings.

StableLM Models. We employed the **StableLM 2-1 6B-Chat** model from StabilityAI as a compact, instruction-following LLM capable of consistent labeling under constrained resources. This model served as the foundation for several runs:

- **stablelm2_rj_v1**: a semi-manual baseline where the model provided deterministic labels (using temperature = 0, `do_sample=False`) following the official TREC 0–4 justification rubric.
- **hybrid.stable1**: an automatic hybrid configuration that combined LLM predictions, lexical overlap (Jaccard similarity between query and segment), and normalized BM25 baseline scores. Calibration parameters emphasized trustworthy mid-range justifications (label = 2) while maintaining some 3/4 evidence.
- **hybrid.stable.loose2**: a looser variant of the hybrid configuration, tuned for broader non-zero coverage (thresholds th1=0.30, th2=0.40, th3=0.56, th4=0.70) and increased inclusion of partial justification cases.

Qwen Models. For large-scale automatic labeling, we adopted **Qwen2.5-3B-Instruct** as the main justification assessor, leveraging its reasoning-oriented pretraining. The following configurations were explored:

- **hybrid.qwen.cal**: a calibrated hybrid judge combining Qwen2.5 outputs with Jaccard overlap (narrative–segment matching) and normalized baseline confidence. The final calibration introduced per-topic caps and floors (`th1=0.40`, `th2=0.52`, `th3=0.66`, `th4=0.78`) to control score saturation, focusing on strong (3/4) evidence with balanced level-2 coverage.
- **hybrid.qwencon**: a consistency-focused run emphasizing high inter-topic agreement by constraining the confidence variance across consecutive segments.

Model Infrastructure. All models were executed in mixed-precision inference mode using PyTorch 2.3 and Hugging Face Transformers 4.39 with bitsandbytes quantization. Inference was performed on NVIDIA A100 80GB GPUs using deterministic decoding (`temperature=0.1`, `top-p=0.9`). Each model produced lines of the form:

```
qid Q0 docid label confidence run_id
```

which were parsed into TREC-compliant run files (`r_output_trec_rag_2025.tsv`) and verified with the official validation script.

Together, these configurations allowed us to explore the trade-off between **label precision, coverage, and interpretability**, and to analyze how smaller open-weight LLMs can be effectively calibrated for justification retrieval tasks within the RAG 2025 framework.

3.3 Methodology

The overall pipeline follows a modular hybrid labeling architecture, which can be described in five key stages: (1) retrieval, (2) data preparation, (3) automatic pre-labeling, (4) human review and correction, and (5) export and validation.

(1) Retrieval. Each of the 105 official RAG topics was processed through a two-stage retriever combining BM25 lexical search and ColBERT semantic reranking. The system produced up to 100 candidate passages per query, ranked by retrieval confidence. The top-20 passages per topic were selected for subsequent justification assessment, yielding 2,100 query–passage pairs.

(2) Data Preparation. Candidate passages were normalized and stored in JSONL format, including identifiers (`qid`, `docid`), rank, and passage text. The preprocessing scripts (`rj_unified.py`) handled tokenization, deduplication, and segment-level formatting to ensure alignment with the MS MARCO v2.1 schema.

(3) Automatic Pre-Labeling. The Qwen2 model automatically assigned justification scores to each passage by evaluating its semantic alignment with the query and the likelihood of being used as factual support. Each prediction included a confidence estimate derived from the model's softmax distribution over discrete justification levels. The pre-labeled outputs were stored in a CSV file (`label_sheet_qwen.csv`) for human validation.

(4) Human Review. Annotators reviewed the automatically generated labels using an interactive command-line interface (`label_cli.py`). For each query, the 20 top passages were sequentially displayed with title, segment text, and current label–confidence pair. Annotators could confirm, correct, or skip entries, and the interface supported optional confidence adjustments (e.g., "3 0.75"). This stage finalized the hybrid annotation dataset (`label_sheet_prefilled.csv`).

(5) Export and Validation. The final dataset was exported into TREC-compatible format (`r_output_trec_rag_2025.tsv`) via `export_qrels.py`. Comprehensive sanity checks (`rj_pipeline.py check`) verified structural integrity, ensuring that each topic contained exactly 20 labeled entries and that all labels fell within the valid range (0–4). After validation, the dataset was submitted as the official **DUTH-RJ-QWEN** run for the TREC 2025 RAG Track.

This methodology combines the interpretability of human annotation with the scalability of LLM-based reasoning, enabling reproducible and cost-efficient creation of justification-level labels suitable for retrieval and RAG evaluation benchmarks.

3.4 Evaluation Measures

Evaluation followed the official relevance judgment protocol for retrieval-augmented generation (RAG), which extends classical information retrieval measures to justification-aware labeling. All submissions were validated and scored using the evaluation infrastructure provided by the organizers.

The primary evaluation measure is **Normalized Discounted Cumulative Gain (NDCG@10)**, computed from graded relevance labels (0–4). This measure captures both ranking quality and the degree of justification alignment between retrieved passages and query intent, rewarding higher-ranked placements of strongly justified evidence (labels 3–4). Formally, NDCG@10 is defined as:

$$\text{NDCG@10} = \frac{1}{Z} \sum_{i=1}^{10} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

where rel_i is the justification grade of the document at rank i, and Z is a normalization constant representing the ideal DCG value for that topic.

Complementary evaluation measures include:

- **Justification Precision (JP@10)** – the proportion of retrieved segments that provide sufficient evidence for the query narrative (label ≥ 2).
- **Justification Recall (JR@20)** – the proportion of gold-standard justifications covered within the top 20 retrieved segments.
- **Coverage Rate (CR)** – the fraction of topics with at least one justified segment in the retrieved top-k list.
- **Mean Confidence Consistency (MCC)** – the mean per-topic standard deviation of confidence scores across the top 20 passages, used to assess calibration stability in automatic and hybrid systems.

Together, these measures assess not only retrieval effectiveness but also **justification reliability**, capturing how consistently a system identifies and ranks supportive textual evidence. Automatic runs (e.g., `duth.hybrid.qwen.cal`, `duth.hybrid.stable1`) are compared against manual and semi-automatic baselines (e.g., `duth_stablelm2_rj_v1`) to quantify the effectiveness of hybrid calibration and lightweight model reasoning under open-weight constraints.

3.5 Implementation and Environment

All experiments were implemented in Python 3.10 using the PyTorch 2.3.0 and Transformers 4.39 frameworks, with quantized inference provided by bitsandbytes 0.47 for efficient GPU utilization. The system was developed and executed on the high-performance cluster at Democritus University of Thrace (DUTH), using CUDA 12.5 and cuDNN 9.x, running on NVIDIA RTX A6000 (48GB) GPUs.

Each LLM (StableLM 2-1 6B, Qwen2.5-3B, and LLaMA3-8B) was loaded in mixed-precision (FP16/BF16) mode with deterministic decoding (`temperature=0.1, top-p=0.9`) to ensure reproducibility. For retrieval, we employed the `Anserini` BM25 baseline and optional reranking with `ColBERTv2`. All scripts, including `rj_pipeline.py`, `prelabel_qwen.py`, and `export_qrels.py`, were executed within isolated virtual environments and version-controlled under Git for complete traceability.

Each experiment was tracked with `wandb.ai` logging for inference timing and memory profiling. The average inference speed was approximately **8.2 samples/sec** for Qwen2.5-3B and **19.5 samples/sec** for StableLM 2-1 6B, with full runs (2,100 pairs) completing in under one hour on a single GPU. All generated artifacts (TSV, CSV, and JSONL files) were verified through the official NIST `check` utility to ensure structural compliance and labeling completeness before submission.

This environment ensured end-to-end reproducibility across manual, hybrid, and automatic configurations, aligning all submissions (hybrid.qwen.cal, hybrid.stable1, hybrid.stable.loose2) under identical runtime and data conditions.

4 Results

Table 1 reports the official label agreement results for the submitted runs in the TREC 2025 RAGTIME *Retrieval-Augmented Generation (RAG) Relevance Judgment* subtask, as released by the TREC Evalbase [12]. Performance is evaluated using Cohen's κ coefficient, together with raw agreement, disagreement, and tie fractions, providing a comprehensive view of annotation consistency under RAG-based relevance assessment.

Among the evaluated runs, hybrid.qwencon achieves the highest κ value (0.050), followed by hybrid.qwen.cal (0.040), indicating modest but consistently positive agreement beyond chance. Although absolute κ values are known to be low for justification-level RAG evaluation, this behavior is expected, as annotators must jointly assess retrieval evidence, contextual adequacy, and justification quality rather than binary topical relevance. Both runs also exhibit the highest agreement fractions, suggesting that hybrid prompting and calibration strategies contribute to more stable labeling behavior.

In contrast, runs such as stablelm2_rj_v1 and the stableri variants obtain near-zero or slightly negative κ values, reflecting substantial disagreement and a large proportion of tied judgments. The elevated tie rates indicate systematic uncertainty rather than random labeling noise, highlighting the inherent difficulty of assigning fine-grained relevance labels in retrieval-augmented generation scenarios. Such outcomes are common in evaluation settings where multiple valid interpretations of evidence support may coexist.

A cross-metric inspection reveals that agreement and disagreement fractions provide complementary insights beyond κ alone. In particular, runs with similar disagreement levels can differ markedly in their tie distributions, underscoring the importance of reporting raw label statistics alongside chance-corrected measures. This observation supports the use of multi-faceted agreement analysis for RAG evaluation, especially in tasks emphasizing justification and evidence quality.

Overall, the results indicate that tighter calibration and hybrid prompting strategies improve labeling consistency in RAG-based relevance judgment. At the same time, the consistently low absolute agreement levels across all runs emphasize the intrinsic ambiguity of the task. These findings motivate future work on agreement-aware calibration, improved annotation guidelines, and evaluation protocols that explicitly account for uncertainty and tie behavior in retrieval-augmented generation tasks.

Beyond Cohen's κ, the agreement, disagreement, and tie fractions reported in Table 1 provide important complementary insights into the behavior of the evaluated labeling configurations. While κ captures chance-corrected agreement, the raw fractions reveal how often systems converge, diverge, or abstain through tied judgments.

The runs hybrid.qwencon and hybrid.qwen.cal exhibit the highest agreement rates (0.250 and 0.260, respectively), indicating that hybrid calibration strategies lead to more frequent alignment with the reference judgments. At the same time, both runs maintain identical disagreement levels (0.730), suggesting that improvements in agreement do not come at the expense of increased explicit conflict. Instead, these systems reduce ambiguity by lowering the proportion of ties, particularly in comparison to baseline configurations.

In contrast, the semi-manual baseline stablelm2_rj_v1 shows a markedly different profile. Although its disagreement rate (0.660) is lower than that of the hybrid systems, this apparent improvement is offset by a substantially higher tie fraction (0.420). This pattern indicates systematic indecision rather than reliable agreement, reflecting difficulty in consistently assigning fine-grained justification labels without additional calibration or hybrid blending. As a result, the near-zero κ score of this run reflects limited practical agreement despite fewer explicit disagreements.

The remaining hybrid variants, hybrid.stableri and stable.loose2, occupy an intermediate regime. Both achieve moderate agreement levels (0.240) but retain relatively high tie fractions (0.150–0.160), leading to slightly negative κ values. This behavior suggests that looser calibration thresholds increase coverage but also amplify ambiguity, yielding less stable labeling outcomes overall.

Taken together, these results demonstrate that agreement, disagreement, and tie fractions must be interpreted jointly when evaluating justification-level relevance judgments. Hybrid systems with tighter calibration reduce tie behavior and yield more stable annotation patterns, even when absolute κ values remain low. This finding reinforces the view that low chance-corrected agreement is an inherent property of justification-centric RAG evaluation, rather than an indication of system failure.

Overall, the analysis confirms that hybrid prompting and confidence calibration improve labeling consistency by shifting uncertainty away from ties toward more decisive judgments. These observations motivate future work on agreement-aware calibration strategies and evaluation protocols that explicitly model ambiguity and partial agreement in retrieval-augmented generation tasks.

Table 1: Official results for the DUTH submissions in the TREC 2025 RAGTIME **Retrieval-Augmented Generation (RAG) Relevance Judgment** subtask (`qrel_eval`), as reported by the TREC Evalbase. Evaluation measures inter-annotator label agreement using Cohen's κ, along with agreement, disagreement, and tie fractions.

Run ID	κ	**Agreement**	**Disagreement**	**Tie**
`hybrid.qwencon`	**0.050**	0.250	**0.730**	0.120
`hybrid.qwen.cal`	0.040	**0.260**	**0.730**	0.110
`stablelm2_rj_v1`	0.000	0.020	0.660	**0.420**
`hybrid.stableri`	-0.010	0.240	0.700	0.150
`stable.loose2`	-0.010	0.240	0.700	0.160

5 Conclusion and Future Work

5.1 Conclusion

This study demonstrated a scalable and reproducible framework for automatic justification assessment within the **TREC 2025 RAG** track. By integrating open-weight LLMs such as **StableLM 2-1 6B** and **Qwen2.5-3B-Instruct** into a unified hybrid pipeline, we achieved stable justification labeling across 105 queries and 2,100 document segments. Our experiments showed that calibration-based blending between semantic confidence and lexical similarity yields balanced performance between coverage and precision, particularly around the crucial justification levels 2–3.

The system's modular structure—combining retrieval, automatic reasoning, and hybrid verification—proved effective for both manual and fully automatic submissions. All final runs (`hybrid.qwen.cal`, `hybrid.stable1`, and `hybrid.stable.loose2`) were validated successfully by NIST, confirming structural and semantic compliance. These results highlight the potential of small open-weight LLMs to provide interpretable and efficient alternatives to proprietary models for large-scale relevance judgment tasks.

5.2 Future Work

Future research will focus on three main directions. First, we plan to explore larger multi-modal retrieval models that integrate justification grounding across text and tables, improving evidence attribution in RAG systems. Second, we aim to extend the calibration framework to a **cross-model consensus** setup, where independent LLMs generate probabilistic agreement scores that dynamically adjust thresholds per topic. Third, we intend to perform a correlation study between automatic justification scores and end-to-end RAG answer quality, quantifying how fine-grained justification accuracy contributes to factual faithfulness in generated responses.

Additionally, we plan to publicly release the DUTH hybrid pipeline as an open-source toolkit, enabling other research teams to replicate or extend our approach for future TREC editions. By emphasizing transparency, open-weight reproducibility, and hybrid reasoning, this line of work contributes toward more interpretable and robust evaluation frameworks for Retrieval-Augmented Generation.

References

[1] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3 (4):333–389, 2009.

[2] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the trec 2020 deep learning track. In *Proceedings of the 29th Text REtrieval Conference (TREC 2020)*. NIST, 2020.

[3] Ellen M. Voorhees. The trec paradigm for information retrieval evaluation. *Communications of the ACM*, 66(1):50–59, 2023.

[4] Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 2020.

[5] Georgios Arampatzis, Vasileios Perifanis, Symeon Symeonidis, and Avi Arampatzis. Duth at semeval-2023 task 9: An ensemble approach for twitter intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1225–1230, 2023. URL `https://aclanthology.org/2023.semeval-1.170/`.

[6] Georgios Arampatzis et al. Duth at exist 2025: Multilingual sexism detection with soft labels and transformers. In *CLEF 2025 Working Notes*, 2025. URL `https://ceur-ws.org/Vol-4038/`.

[7] Georgios Arampatzis and Avi Arampatzis. Duth at clef joker 2025 tasks 2 and 3: Translating puns and proper names with neural approaches. In *CLEF 2025 Working Notes*, 2025. URL `https://ceur-ws.org/Vol-4038/`.

[8] Georgios Arampatzis and Avi Arampatzis. Duth at clef 2025 simpletext track: Tackling scientific text simplification and hallucination detection. In *CLEF 2025 Working Notes*, 2025. URL `https://ceur-ws.org/Vol-4038/`.

[9] Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Reproducible information retrieval research with

lucene. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1253–1256, 2017.

[10] John Pavlopoulos, Jesper Sorensen, Nikos Papasarantopoulos, and Ion Androutsopoulos. Empirical evaluation of deep learning models for semi-automatic text annotation. In *Proceedings of LREC 2020*, 2020.

[11] Timo Schick and Hinrich Schütze. Self-labeling for few-shot learning with large language models. *Transactions of the Association for Computational Linguistics*, 11:1–16, 2023.

[12] National Institute of Standards and Technology (NIST). Trec 2025 ragtime: Retrieval-augmented generation relevance judgment subtask (qrel_eval). `https://ir.nist.gov/evalbase/conf/trec-2025/trec2025-rag-qrels/appendix/qrel_eval`, 2025. Official evaluation results released via the TREC Evalbase.

[13] NIST. Trec 2025 rag track guidelines. `https://trec.nist.gov/tracks.html`, 2025.

[14] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of SIGIR 2020*, 2020.

[15] 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 *EMNLP 2020*, 2020.

[16] Stability AI. Stablelm 2: Open-weight language models for dialogue and instruction following. `https://github.com/Stability-AI/StableLM`, 2024. Accessed: 2025-08-31.

[17] Zeqi Bai, Ruixiang Zhao, Yang Liu, et al. Qwen2.5 technical report: Advancing open-weight language models for reasoning and multilinguality. *arXiv preprint arXiv:2409.06600*, 2024.

[18] Zeqi Bai, Ruixiang Zhao, Yang Liu, and et al. Qwen technical report: Large language models for long-context reasoning and multilingual understanding. 2023.

[19] Hugo Touvron, Thibaut Lavril, Gautier Izacard, and et al. Llama 3: Open and efficient foundation language models. Meta AI Research, 2024.

[20] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[21] Lukas Biewald. Weights & biases: Experiment tracking for machine learning. `https://wandb.ai`, 2023. Accessed: 2025-08-31.