WueRAG at RAGTIME 2025: Retrieval, Fusion, and Citation for Grounded Report Generation

Julia Wunderle, **Julian Schubert**, **Joachim Baumeister**, **Andreas Hotho**

Center for Artificial Intelligence and Data Science (CAIDAS)
Julius-Maximilians-Universität Würzburg (JMU)
lastname@informatik.uni-wuerzburg.de

denkbares GmbH, Würzburg
firstname.lastname@denkbares.com

Abstract

We present WueRAG, a retrieval-augmented generation pipeline for the TREC 2025 RAGTIME English report generation task. Our approach combines query reformulation, remote candidate retrieval, and local per-topic reranking using a hybrid dense and lexical fusion. To ensure citation accuracy, grounding is enforced at two levels: first, through a generation phase that requires bracketed citations for factual claims and second, via a postprocessing filter that removes any remaining unverified sentences. On the official evaluation WueRAG achieved the highest F1 score (0.421) for the english subtask, indicating that combining multi-stage retrieval with explicit grounding constraints can effectively balance attribution accuracy and report quality.

1 Introduction

Large language models (LLMs) excel at fluent text generation but are limited by a fixed training corpus (Ji et al., 2023). They can not access recent events and do not provide verifiable sources for generated statements. This makes hallucination, generating plausible sounding but factually incorrect content, a persistent concern. This limitation is especially critical in the news domain, where accuracy and verifiability are non-negotiable. Although finetuning can incorporate new content, the high computational cost and need for constant retraining make it impractical for such dynamic data.

Retrieval-Augmented Generation (RAG) offers a more scalable approach (Lewis et al., 2020). By querying an external document collection at inference time, RAG grounds model responses in current evidence without expensive weight updates. We apply this setting to the TREC 2025 RAGTIME challenge, which focuses on English report generation from a news corpus. RAGTIME requires systems to synthesize concise reports that address a specific problem statement and user background while ensuring that every claim is supported by a verifiable citation.

We present **WueRAG**, a two-stage RAG pipeline that retrieves candidate documents via the RAGTIME API, reranks passages through a fusion of dense and sparse retrieval, and generates cited reports with GPT-4o (OpenAI et al., 2024). WueRAG ranked first among all participating teams in the English subtask, with an overall F1 score of 0.421, achieving strong citation fidelity (sentence support = 0.734) with moderate topic coverage (nugget coverage = 0.325).

2 Task Setting

RAGTIME is a TREC 2025 shared task focused on citation-grounded report generation from news corpora (Lawrie et al., 2025). We participate in the monolingual English subtask.

Data The dataset consists of approximately 1.01 million English documents sourced from the CommonCrawl news service, crawled between August 2021 and July 2024 (Lawrie et al., 2025). RAGTIME topics are structured as complex, professional report requests. Each topic is defined by four fields: a `title`, a `background` describing the persona of the requesting person or context, a `problem statement` specifying the report objective (often requiring historical timelines, stakeholder analysis, or geographic comparisons), and a character `limit`.

Evaluation The organizers report three metrics to evaluate system performance. `Nugget coverage` measures the recall of key information identified by human annotators, while `sentence support` assesses grounding by verifying whether generated claims are supported by their referenced documents. The `F1 score` serves as the harmonic mean of nugget coverage and sentence support, balancing topical breadth against citation accuracy. The evaluation methodology scales according to report length: short reports of 2000 characters undergo manual assessment by human experts, whereas long reports of 10 000 characters are evaluated automatically via Auto-ARGUE (Walden et al., 2025), an LLM-based framework using Llama 3 70B. At the time of the official release, 16 main-task topics had been evaluated by the organizers, providing the scores used throughout this paper (Lawrie et al., 2025).

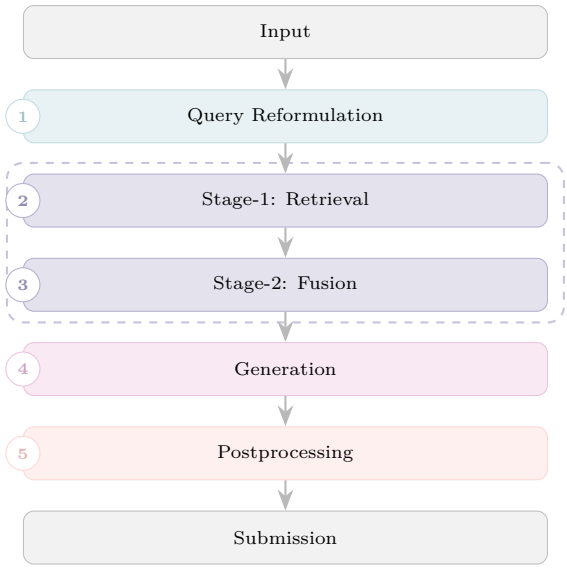

Figure 1: WueRAG pipeline. The topic prompt is first reformulated into a concise retrieval query; stage-1 fetches documents via the RAGTIME API; stage-2 builds a per-topic in-memory index and fuses dense and BM25 rankings via RRF; the top passages are passed to GPT-4o for citation-constrained generation; postprocessing discards uncited sentences before final submission.

3 System

WueRAG follows a straightforward RAG design: retrieve relevant documents, rerank them, and generate a cited report. The overall pipeline is illustrated in Figure 1. We build on LlamaIndex (Liu, 2022), an open-source framework that provides composable abstractions for the mentioned design components, allowing us to assemble the full pipeline easily, from well-tested components.

Input For our WueRAG input, we concatenate the `background`, `problem_statement`, and `limit` fields into a single prompt. We exclude the `title` field, as it provides redundant information already captured in the other sections.

Query Reformulation

Prompt 1: Query Reformulation

```
You are reformulating user requests into search
queries for retrieval. Rules:
- Return ONLY a search query, not instructions or
answers.
- Be concise but preserve key details (topics,
timeframe, actors).
- Do NOT include phrases like 'create a report',
'explain', or 'should include'.
- Output a single line, no quotes.
User request: {query}
Search Query:
```

To improve retrieval effectiveness, we use GPT-4o to extract concise search queries from the initial user request, as raw prompts often contain instructional overhead that can degrade search precision. If reformulation fails, the system falls back to the original prompt. Importantly, the reformulated query is used *only* for retrieval; the original full topic string is passed as the query to the generation step, preserving all background and context for the LLM.

Two-Stage Retrieval In the `stage-1 retrieval`, we retrieve a broad initial candidate set of k=30 documents from the RAGTIME API. However, since raw documents often mix relevant and irrelevant content, passing them directly to generation may introduce noise and weaken citation precision. Therefore, in `stage-2` we re-index the documents locally and rerank them at a passage level. We run two complementary retrievers over this temporary index: (1) a `dense retriever` that encodes passages into a shared embedding space via all-MiniLM-L6-v2 (SentenceTransformers), and (2) `BM25` (Robertson and Zaragoza, 2009), which scores passages by exact keyword overlap to capture named entities and rare terms. Their rankings are merged by `reciprocal rank fusion` (RRF) (Cormack et al., 2009), which combines ranked lists by summing the reciprocals of each passage's rank positions, without requiring score normalisation across retrievers, producing a final set of k=10 passages. In short: stage 1 handles recall; stage 2 handles precision.

Generation and Citation Policy The `CitationQueryEngine` is a LlamaIndex component that automatically numbers the retrieved passages as sources, injects them into the prompt, and maps bracket citations in the output back to their originating nodes. It selects the top-5 passages from the fused ranking and utilizes Prompt 2 for generation. Citation grounding is enforced at two levels: the generation prompt requires a bracket citation after every factual statement, and postprocessing removes any sentence that lacks a resolved citation marker. This two-level policy directly targets the sentence support metric used in RAGTIME evaluation. Character-budget compliance is additionally checked by a hard assertion before results are written to disk.

Prompt 2: Generation

```
Please answer using **only** the provided sources.
Every factual statement **must** be followed by
its source citation, placed at the **end of the
sentence**.
Use the source number in square brackets
(e.g., "Water freezes at 0 degrees C [2].").
Do not include any uncited facts and keep your
```

```
answer as short as possible not surpassing
{char_limit} characters.
Sources: {context_str}
Query: {query_str}
Answer:
```

Postprocessing The generated text is split into sentences using a punctuation-based boundary detector. Each sentence is checked for bracket citation markers (`[n]`); sentences without at least one valid citation are discarded. For retained sentences, citation indices are resolved to source document identifiers, bracket markers are stripped, and a citation dict mapping source ID to relevance score is stored alongside the cleaned text.

4 Evaluation

This section summarizes the official scores provided by the organizers and analyzes the behavior of our system, highlighting strengths and weaknesses.

4.1 Official Aggregate Results

Among the participants, our submission `WueRAG_2025_08_22` achieved the highest F1 score of 0.421 for the English subtask. The strong sentence support (0.734) indicates that the claims generated by our system are well grounded in the retrieved evidence. However, notably lower nugget coverage (0.325) suggests that the retrieval stage does not yet capture all relevant facts required for the task.

Table 1: Metrics for `WueRAG_2025_08_22`. High F1 and sentence support scores indicate well-grounded claims, while lower nugget coverage suggests retrieval limitations.

Metric	Value
F1 score	0.421
nugget coverage	0.325
sentence support	0.734

4.2 Topic Analysis

For the final submission of the English subtask, our system cited 938 sentences across 122 topics, averaging 7.67 sentences per topic. Notably, three topics produced no output, likely because all generated sentences were discarded by the postprocessing filter due to invalid citations.

4.3 Topic-Level Behavior

Figure 2 shows the per-topic distribution across the 16 topics graded by the organizers. Sentence support is consistently high with little variance, confirming that citation grounding is reliable across topics. Nugget coverage, in contrast, fluctuates notably which suggests that the bottleneck is retrieval

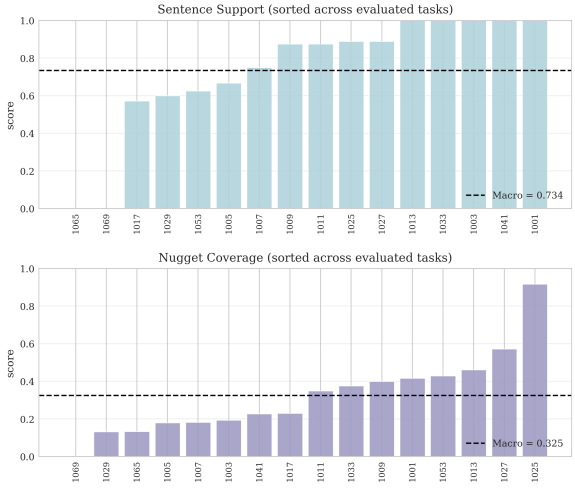

Figure 2: Per-topic metrics on the evaluated subset (top to bottom): sentence support, and nugget coverage. Support metrics are generally high, while nugget coverage shows larger spread, indicating that evidence is often well grounded but not always sufficiently broad.

breadth rather than generation quality. This pattern suggests three causes: (1) topic difficulty varies strongly, especially for long multi-facet prompts; (2) single-query retrieval can miss relevant facets, reducing coverage; and (3) strict postprocessing protects faithfulness but can amplify recall errors by removing weakly grounded content instead of partially covering a facet.

The impact of topic difficulty is further analyzed in Figure 3, relating the F1 score to input and output length. We observe a moderate negative correlation with problem-statement length ($r = -0.50$), indicating that longer and likely more multi-facet requests are harder for this retrieval setup. Conversely, answer length shows a stronger positive correlation with the F1 score ($r = 0.64$), suggesting that when enough relevant evidence is retrieved, the model can use the character budget to cover more required nuggets.

5 Conclusion

LLMs remain limited by fixed training data and weak source traceability, which is especially problematic for time-sensitive news-domain report generation. WueRAG addresses this with a two-stage, API-first pipeline that separates retrieval breadth from synthesis precision: remote candidate retrieval, local dense and sparse reranking, and citation-constrained generation. In the TREC 2025 RAGTIME English subtask, this design ranked first, reaching an F1 score of 0.421. The evaluation shows a clear metric profile: strong grounding (0.734) but lower nugget coverage (0.325). Topic-level analysis further indicates that broad multi-facet prompts are

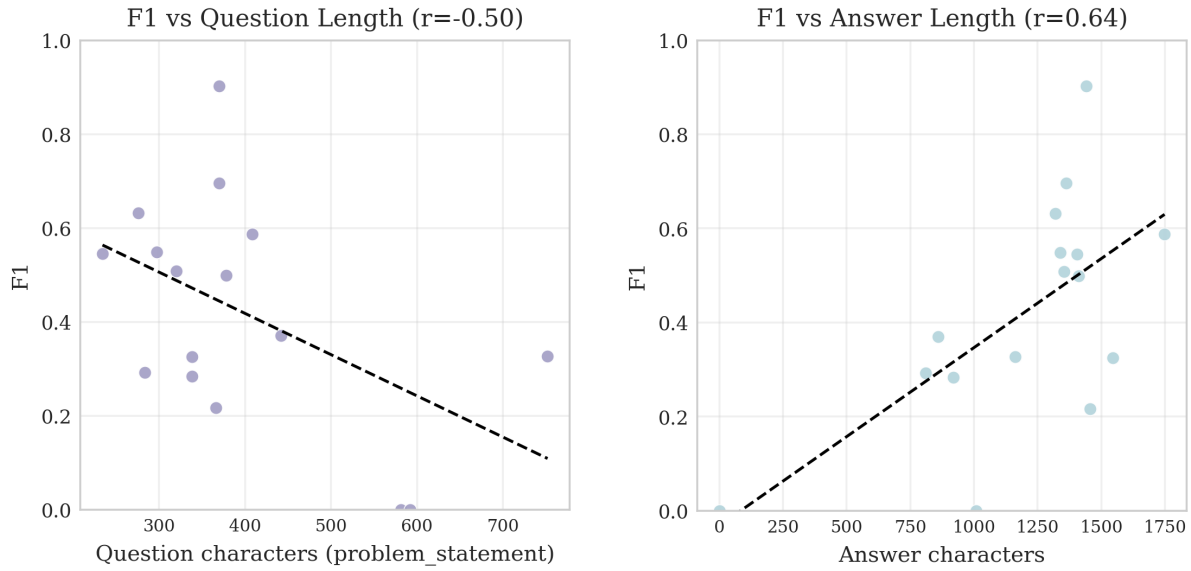

Figure 3: Length effects on F1 score (left: versus problem-statement length; right: versus answer length). In the evaluated subset, longer answers tend to correlate with higher F1 score, while longer problem statements correlate with lower F1 score.

the main challenge case, while narrower prompts are handled more robustly. Overall, these results position WueRAG as a practical baseline for citation-grounded report generation under strict length budgets, with recall breadth as the central remaining bottleneck.

6 Future Work

The primary limitation of our system lies in the retrieval stage. In particular, nugget coverage decreases for broad multi-faceted prompts, indicating that the retriever fails to surface all relevant evidence even when downstream citation-support metrics remain stable.

One promising direction is to augment the retrieval pipeline with a structured knowledge graph, similar to GraphRAG (Edge et al., 2024). By explicitly modeling entities such as countries, persons, and thematic categories, the system could traverse relational structures to retrieve semantically connected evidence that is not directly co-located in the document space.

Additionally, query strategies could incorporate domain knowledge to guide graph traversal. Iterative reasoning over entity relationships may progressively refine the candidate node set, focusing retrieval on the most relevant subgraph (Schubert et al., 2024). Complementary self-reflective retrieval policies could further improve coverage by allowing the system to detect incomplete evidence and trigger additional retrieval iterations before finalizing generation (Asai et al.).

Finally, deployment flexibility could be improved by leveraging compact local language models (Yang et al., 2025; Olmo et al., 2025; Pfister et al., 2025). Smaller models reduce latency and mitigate data-governance concerns in sensitive environments. This enables the application of WueRAG beyond the news domain to high-stakes fields such as medicine, where patient records must remain within institutional infrastructure. In this context, the system could be reliably used to extract and cite patient information from unstructured clinical notes, ensuring that all findings are backed by verifiable evidence.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, An-

drea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Luca Soldaini, Eugene Yang, and Andrew Yates. 2025. Overview of the TREC 2025 RAGTIME track. In *Proceedings of the Thirty-Fourth Text REtrieval Conference (TREC)*.

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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Jerry Liu. 2022. LlamaIndex. https://github.com/run-llama/llama_index.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. Olmo 3.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie

Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. Llämmlein: Transparent, compact and competitive german-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246.

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

Julian Schubert, Markus Krug, and Joachim Baumeister. 2024. Knowledge retrieval with llms using context-specific intent and slot classification. In *Proceedings of the LWDA 2024 Workshops: Lernen, Wissen, Daten, Analysen*. Workshop on Information Retrieval (FGIR) at LWDA 2024.

SentenceTransformers. all-minilm-l6-v2.

William Walden, Marc Mason, Orion Weller, Laura Dietz, John Conroy, Neil Molino, Hannah Recknor, Bryan Li, Gabrielle Kaili-May Liu, Yu Hou, et al. 2025. Auto-argue: LLM-based report generation evaluation. *arXiv preprint arXiv:2509.26184*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.