CultRAG at SemEval-2026 Task 7: Hybrid Sparse-Dense Retrieval with Entity-Centric Knowledge Bases for Cultural MCQ Answering

Aditya Singh* **Rickarya Das***
Indian Institute of Technology Kharagpur
{aditya26189, rickaryadas}@kgpian.iitkgp.ac.in

Abstract

We present a trust-weighted Retrieval-Augmented Generation (RAG; Lewis et al., 2020) system for SemEval-2026 Task 7 (BLEnD) Track 2 (Ousidhoum et al., 2026), targeting English cultural multiple-choice QA across 30 countries. Built atop Llama-3.1-8B-Instruct (Meta AI, 2024), the six-phase pipeline integrates hybrid BM25+FAISS retrieval, country-aware filtering, intent detection, tiered routing, anti-leak prompt engineering, and trust-weighted reranking. The core finding is that RAG *hurts* rather than helps: the LLM-only baseline achieves 78.6% accuracy, outperforming the full system at 78.5% (McNemar's test, $p = 0.962$). Oracle analysis reveals that only 40.7% of questions are answerable from the knowledge base, explaining why retrieval introduces more noise than signal. The sole recovery comes from anti-leak prompt filtering (Phase 4), which mitigates answer-anchoring artifacts. Code: https://github.com/CultRAG/BLEnD-CultRAG.

1 Introduction

The BLEnD benchmark (Myung et al., 2024; Ousidhoum et al., 2026) evaluates LLMs on culturally grounded knowledge across diverse world regions. Cultural knowledge—traditions, social norms, geography, cuisine, local customs—poses a distinctive challenge because LLM training corpora skew Western-centric. Track 2 of SemEval-2026 Task 7 (Ghosh et al., 2026) operationalizes this as a four-option MCQ task spanning 47,014 English questions across 30 country-language pairs, requiring cultural understanding unlikely to reside in parametric knowledge alone.

Retrieval-Augmented Generation (RAG; Lewis et al., 2020) is a natural mitigation: a curated cultural KB should fill parametric gaps for underrepresented cultures. We design a six-phase trust-weighted RAG pipeline on this hypothesis.

Our experiments largely reject this hypothesis. The LLM-only baseline (78.6%) matches or exceeds every RAG configuration, and the full system achieves 78.5%—statistically indistinguishable ($p = 0.962$). Oracle analysis shows only 40.7% of questions are answerable from the KB; the LLM's parametric knowledge exceeds the retrieval ceiling—consistent with findings that parametric memory suffices for well-represented knowledge (Mallen et al., 2023). Anti-leak filtering (Phase 4) is the only component producing meaningful recovery, indicating the primary failure mode is answer-anchoring noise, not retrieval quality.

We make four contributions: (1) a six-phase RAG pipeline with trust-weighted reranking and anti-leak filtering; (2) an oracle KB coverage analysis quantifying the theoretical retrieval ceiling; (3) a country-level decomposition revealing that RAG helps low-resource cultures but hurts high-resource ones; and (4) rigorous McNemar significance testing with effect size reporting for all ablation comparisons.

2 System Description

2.1 Knowledge Base Construction

The KB comprises 1,262 text chunks in 7 shards covering all 30 countries, sourced from Wikipedia, travel guides, news outlets, and country portals. Each chunk is assigned a **trust tier**: *High* (Wikipedia, government portals, reference works), *Mid* (reputable travel and news outlets), or *Low* (forums, blogs, unverified sources). Trust tiers drive a multiplicative reranking weight:

$$w_\tau(d) = \begin{cases} 1.0 & \text{if } \tau(d) = \textit{high} \\ 0.6 & \text{if } \tau(d) = \textit{mid} \\ 0.3 & \text{if } \tau(d) = \textit{low} \end{cases} \quad (1)$$

where $\tau(d)$ denotes the trust tier of chunk d (applied in Phase 5, Section 2.5).

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3137–3142

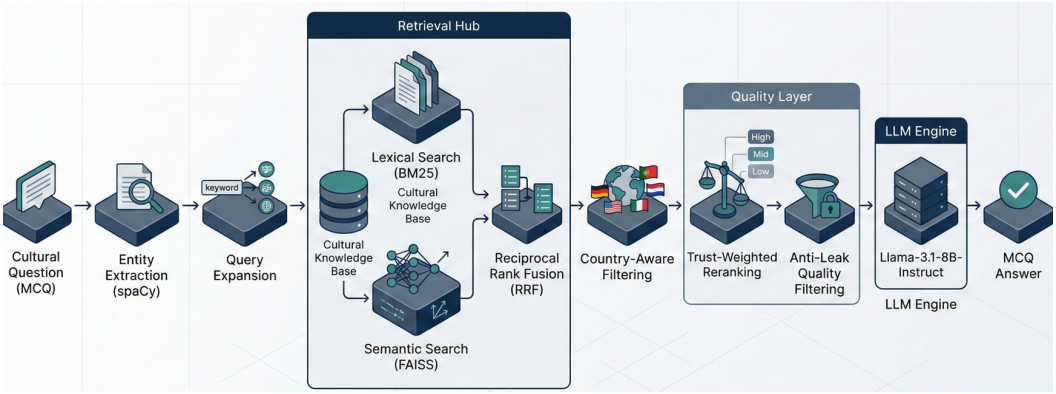

Figure 1: Architecture of the multi-phase cultural RAG pipeline. The system integrates entity extraction, hybrid retrieval (BM25 + FAISS), Reciprocal Rank Fusion, trust-weighted reranking, and phased prompt engineering to generate culturally-grounded answers.

2.2 Entity Extraction

We extract entities using spaCy's (Explosion, 2023) *en_core_web_sm* NER pipeline (∼1.5 entities per question on average: person names, locations, organizations, cultural artifacts). Let $E(q) = \{e_1, \ldots, e_m\}$ be the entities from question q. The expanded query is:

$$q' = q \oplus e_1 \oplus \cdots \oplus e_m \tag{2}$$

where $\oplus$ denotes string concatenation, improving recall for chunks mentioning the same entities in variant surface forms.

2.3 Hybrid Retrieval

Let D be the full set of KB chunks, q' the expanded query, and $D_c \subset D$ the subset of chunks tagged with country c. Retrieval operates over D_c.

Lexical Retrieval (BM25). BM25 (Robertson and Zaragoza, 2009) scores each chunk $d \in D_c$:

$$\text{BM25}(q', d) = \sum_{t \in q' \cap d} \text{IDF}(t) \times \frac{f(t,d)(k_1+1)}{f(t,d) + k_1(1 - b + b\frac{|d|}{\text{avgdl}})} \tag{3}$$

where $f(t,d)$ is term frequency, $|d|$ is document length, avgdl is the mean document length over D_c, and:

$$\text{IDF}(t) = \log \frac{|D_c| - n(t) + 0.5}{n(t) + 0.5} \tag{4}$$

with $n(t)$ denoting the number of chunks containing t. We set $k_1 = 1.5, b = 0.75$.

Semantic Retrieval (FAISS). FAISS (Johnson et al., 2021) provides dense semantic retrieval using the *all-MiniLM-L6-v2* sentence encoder (Reimers and Gurevych, 2019) ϕ:

$$\text{Dense}(q', d) = \phi(q')^\top \phi(d) \tag{5}$$

Reciprocal Rank Fusion. The two ranked lists are fused (Cormack et al., 2009):

$$S_{RRF}(d) = \sum_{r \in \{R_{BM25}, R_{Dense}\}} \frac{1}{k + \text{rank}(d, r)} \tag{6}$$

with smoothing constant $k = 60$. The retrieval output is $\mathcal{R}(q') = \text{top-3 } \text{argmax}_d S_{RRF}(d)$. Country filtering restricts candidates to D_c via the question ID prefix (e.g., `ja-JP_0042` → Japan), eliminating cross-country contamination.

2.4 Intent Detection

A keyword heuristic classifies questions into 16 intent types (food_drink, sports, government_politics, geography_places, etc.) by matching against question and option text, assigning non-"others" labels to 95.6% of questions (food_drink 46.6%, sports 44.0%, education 4.1%, geography 0.6%, others 4.4%). Despite this coverage, intent labels have zero downstream effect: Phases 2 and 3 produce identical or worse accuracy (Section 2.5), indicating BLEnD does not benefit from intent-conditioned prompting at this granularity.

2.5 Phased Prompt Engineering

The pipeline uses six cumulative prompt engineering phases:

Phase 1 (Country Filter): Restricts retrieved chunks to the target country. Results are identical to unfiltered RAG because entity-expanded queries already steer BM25+FAISS toward country-relevant

chunks.

Phase 2 (Intent Detection): Injects intent metadata into the prompt. Despite covering 95.6% of questions (Section 2.4), accuracy is unchanged (77.9%), suggesting the model cannot exploit coarse keyword-derived intent.

Phase 3 (Tiered Routing): Routes questions to intent-specific prompt templates (context-heavy for high-confidence intents, knowledge-light otherwise). Accuracy drops to 77.6%—the worst configuration—indicating poorly calibrated templates add noise rather than structure.

Phase 4 (Anti-Leak Quality Filtering): The sole impactful component. The model is instructed to ignore passage structure, formatting cues, and answer-letter mentions in KB text:

$$P_{final} = \text{Concat}(\text{Prompt}, G(\text{Context})) \quad (7)$$

The same documents are retrieved; the model reasons independently of surface cues, recovering 1.0pp over Phase 3 to 78.5%.

Phase 5 (Trust-Weighted Reranking): Reorders documents by trust tier before prompt insertion:

$$S_{\text{final}}(d) = S_{RRF}(d) \times w_{\text{tier}}(d), \quad w_{\text{tier}} \in \{1.0, 0.6, 0.3\} \quad (8)$$

Predictions are identical to Phase 4, suggesting document order has negligible impact at $k = 3$.

Phase 6 (Full System): Combines all phases. The final prediction $\hat{y}$ uses constrained decoding:

$$\hat{y} = \arg \max_{a \in \{A,B,C,D\}} P(a \mid P_{final}, \theta) \quad (9)$$

where θ denotes the Llama-3.1-8B-Instruct (Meta AI, 2024) parameters. Predictions match Phases 4–5 (36,928 correct), confirming anti-leak filtering is the only component adding value beyond baseline.

3 Experimental Setup

BLEnD Track 2 comprises $N = 47{,}014$ English MCQs (four options each) across 30 country-language pairs with approximately uniform gold labels (A=24%, B=26%, C=26%, D=24%). All experiments use Llama-3.1-8B-Instruct (Meta AI, 2024) on Kaggle GPUs with constrained decoding over $\{A, B, C, D\}$.

Accuracy is the primary metric:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = y_i^*] \quad (10)$$

where y_i^* is the gold answer and $\hat{y}_i$ the prediction.

Wilson score confidence intervals:

$$\frac{\hat{p} + \frac{z_{\alpha/2}^2}{2N} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N} + \frac{z_{\alpha/2}^2}{4N^2}}}{1 + \frac{z_{\alpha/2}^2}{N}} \quad (11)$$

with $\hat{p} = \text{Acc}$ and $z_{\alpha/2} = 1.96$ for 95% intervals.

Cohen's h measures effect size:

$$h = 2\arcsin\sqrt{p_1} - 2\arcsin\sqrt{p_2} \quad (12)$$

where p_1, p_2 are the two accuracies being compared.

4 Experimental Results

4.1 Ablation Study

Table 1 summarizes results. The baseline (78.6%) is the best configuration; unfiltered RAG degrades by 0.6pp ($p < 0.0001$). Ph1–Ph2 produce no change (77.9%). Ph3 worsens to 77.6% from poorly calibrated templates. Ph4 is the sole recovery (+1.0pp over Ph3); Ph5–Ph6 add nothing beyond Ph4.

Configuration	Correct	Acc (%)	95% CI	Δ Base
Baseline (LLM only)	36,932	78.6	[78.2, 78.9]	—
RAG Basic (unfiltered)	36,639	77.9	[77.6, 78.3]	−0.6
+ Country Filter (Ph1)	36,639	77.9	[77.6, 78.3]	−0.6
+ Intent Detection (Ph2)	36,639	77.9	[77.6, 78.3]	−0.6
+ Tiered Routing (Ph3)	36,465	77.6	[77.2, 77.9]	−1.0
+ Quality Signals (Ph4)	36,928	78.5	[78.2, 78.9]	−0.0
+ Trust Reranking (Ph5)	36,928	78.5	[78.2, 78.9]	−0.0
Full System (Ph6)†	36,928	78.5	[78.2, 78.9]	−0.0

Table 1: Ablation study results across eight configurations for 47,014 questions. †Official submission.

Figure 2 visualizes the progression: accuracy drops upon adding RAG, stays flat through Phases 1–2, dips at Phase 3, then recovers at Phase 4 to near-baseline.

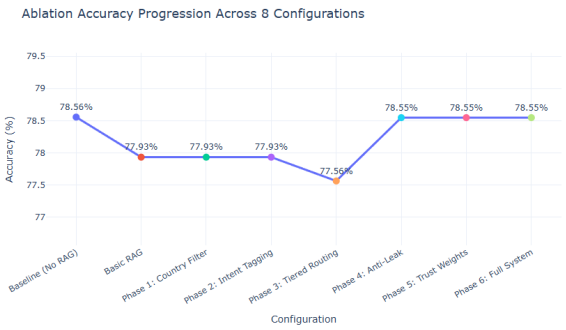

Figure 2: Ablation accuracy progression across eight cumulative pipeline configurations. The baseline (no RAG) achieves the highest accuracy; anti-leak filtering (Phase 4) is the only component that recovers performance.

4.2 Statistical Significance

We use McNemar's test for paired binary outcomes:

$$z = \frac{(|b - c| - 1)}{\sqrt{b + c}} \tag{13}$$

where b and c are discordant pair counts (Table 2). Baseline vs. RAG Basic ($p < 0.0001$) and RAG Basic vs. Full System ($p < 0.0001$) are both significant, but all Cohen's $h \leq 0.015$—detectable at $N = 47{,}014$ yet practically negligible. Baseline vs. Full System is not significant ($p = 0.962$, $h = 0.000$): the full pipeline is indistinguishable from doing nothing.

Comparison	Δ Acc	p-value	Sig	Cohen's h	Effect
Base vs. RAG Basic	−0.6	< 0.0001	Yes	0.015	Small
Base vs. Full System	−0.0	0.962	No	0.000	Small
RAG vs. Full System	+0.6	< 0.0001	Yes	0.015	Small

Table 2: McNemar's paired significance tests across 47,014 questions.

5 Analysis

5.1 Why RAG Hurts: KB Coverage Ceiling

We define the **oracle coverage** for a question q_i as:

$$\text{Cov}(q_i) = \mathbf{1}\left[y_i^* \in \bigcup_{d \in D_c} \text{text}(d)\right] \tag{14}$$

where y_i^* is the gold answer letter and D_c is the country-specific KB shard. The aggregate coverage ceiling is $\bar{C} = \frac{1}{N}\sum_i \text{Cov}(q_i)$. Only 19,148 questions (40.7%) satisfy $\text{Cov}(q_i) = 1$. The *retrieval ceiling gap* quantifies the futility of retrieval in aggregate:

$$\Delta_{\text{ceil}} = \text{Acc}_{\text{param}} - \bar{C} = 0.786 - 0.407 = +0.379 \tag{15}$$

Coverage varies substantially: Bulgaria (69%), Ecuador (68%), Ethiopia (65%) vs. Great Britain (21%), North Korea (26%), Algeria (26%). Since baseline accuracy (78.6%) exceeds this ceiling, retrieval can only add noise for 59.3% of questions. Great Britain exemplifies the problem: 92.4% baseline vs. 21% coverage, costing 3.0pp under the full system (89.7%). Conversely, Ethiopia (65% coverage, 59.0% baseline) has room for retrieval; the full system gains 4.1pp (63.0%). Figure 3 visualizes this per-country split: retrieval cannot be justified when its coverage ceiling lies below parametric performance.

5.2 RAG Backfire Decomposition

Per-question outcomes form a 2×2 contingency table: a = both correct, b = baseline-only correct, c = system-only correct, d = both wrong; net change $\Delta = c - b$. Baseline vs. RAG Basic: $c = 1{,}810$, $b = 2{,}103$, $\Delta = -293$. Baseline vs. Full System: $c = 1{,}977$, $b = 1{,}981$, $\Delta = -4$ ($p = 0.962$). RAG Basic vs. Full System: Phase 6 recovers 1,479 broken questions while introducing 1,190 new errors (+289 net). Anti-leak filtering mitigates anchoring artifacts without improving retrieval itself.

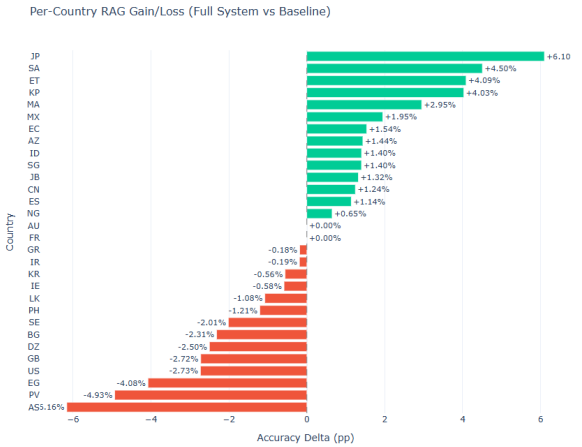

Figure 3: Per-country RAG gain/loss: accuracy delta between the full system (Phase 6) and the LLM-only baseline. Green bars indicate countries where RAG helps; red bars where it hurts.

5.3 Country-Level Analysis

Fourteen countries show improvement: Japan (+6.1pp), Saudi Arabia (+4.5pp), Ethiopia (+4.1pp), North Korea (+4.0pp), Morocco (+2.9pp), Mexico (+1.9pp), Ecuador (+1.5pp), Singapore (+1.4pp), Azerbaijan (+1.4pp), Indonesia (+1.4pp), West Java (+1.3pp), China (+1.2pp), Spain (+1.1pp), and Nigeria (+0.6pp). These are predominantly countries underrepresented in English-language pretraining.

Fourteen countries show degradation: American Samoa (−6.2pp), Basque Country (−4.9pp), Egypt (−4.1pp), Great Britain (−3.0pp), United States (−2.7pp), Algeria (−2.5pp), Bulgaria (−2.3pp), Sweden (−2.0pp), Philippines (−1.2pp), Sri Lanka (−1.1pp), South Korea (−0.6pp), Ireland (−0.6pp), Iran (−0.2pp), and Greece (−0.2pp). These are high-resource countries where parametric knowledge dominates (GB, US), or countries with sparse KB entries where context actively misleads (Basque Country, American Samoa). Fig-

ure 4 confirms the pattern: high-baseline countries cluster below the diagonal (RAG hurts); low-baseline countries lie above it (RAG helps). RAG should be applied selectively—conditioned on confidence or coverage—not uniformly.

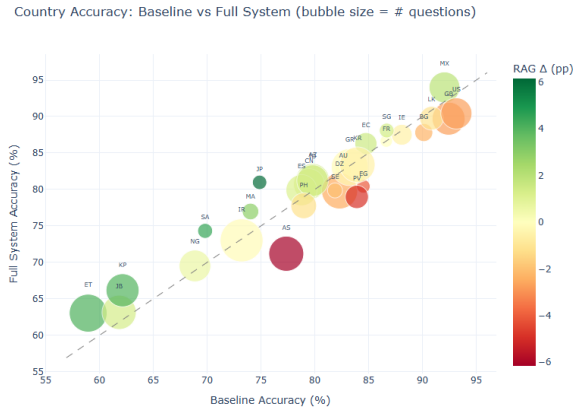

Figure 4: Baseline vs. full system accuracy per country. Bubble size encodes question count; color encodes the RAG delta. Points above the diagonal benefit from RAG.

5.4 Answer Distribution Bias

The baseline over-predicts A (32%) and under-predicts C (22%) and D (21%)—a positional bias common in instruction-tuned LLMs (Wang et al., 2022). Phase 4 partially corrects this (A reduced to ∼30%), though dedicated debiasing (answer shuffling or calibration) would be more effective.

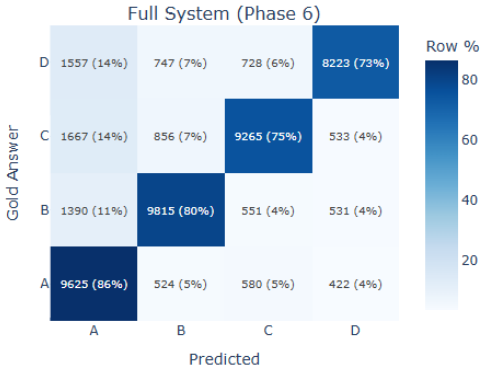

Figure 5: Confusion matrix of full system predictions versus gold labels (47,014 questions). Off-diagonal mass concentrates in the A-predicted row.

5.5 Limitation: Intent-Conditioned Prompting

The keyword classifier covers 95.6% of questions yet yields zero accuracy gain (Section 2.4). The bottleneck is prompt-template design: neither intent metadata injection (Ph2) nor intent-specific routing (Ph3) helps—Ph3 actively hurts. Intent-aware retrieval with a fine-tuned classifier, rather than prompt-only injection, is needed.

6 Official Shared-Task Result

Our submission (CULTRAG, team *aditya26189*) scored **77.55%** accuracy across 29 of 30 BLEnD Track 2 locales, placing 12$^{\text{th}}$ of 18 systems. Locale accuracy ranges from `es-MX` (94.00%) to `am-ET` (63.05%)—a 31-point spread mirroring the coverage gap (Section 5.1). The top five (`es-MX`, `en-US`, `en-GB`, `ta-LK`, `zh-SG`) all exceed 88% and are high-resource or English-native locales. The bottom five (`am-ET`, `su-JB`, `ko-KP`, `ha-NG`, `as-AS`) are low-resource locales where both parametric and retrieved knowledge are sparse.

7 Conclusion

We presented a six-phase trust-weighted RAG pipeline for BLEnD Track 2 atop Llama-3.1-8B-Instruct. The central result is negative: RAG does not help ($p = 0.962$) because parametric accuracy (78.6%) exceeds the KB coverage ceiling ($\bar{C} = 0.407$). Anti-leak filtering is the sole recovery mechanism, gaining 1.0pp by suppressing answer-anchoring artifacts.

Country-level analysis shows RAG benefits low-resource cultures (Japan +6.1pp, Ethiopia +4.1pp) while degrading high-resource ones (American Samoa −6.2pp, Great Britain −3.0pp). Four directions emerge: (1) confidence-conditioned selective retrieval; (2) targeted KB expansion for low-coverage countries; (3) intent-aware retrieval with learned classifiers; and (4) cross-encoder reranking to reduce noise injection.

Ethics Statement

Key ethical considerations: (1) **Systemic Bias**: Wikipedia and WikiVoyage reflect Western-centric perspectives that may misrepresent low-resource cultures. (2) **Cultural Stereotyping**: retrieval may anchor on over-simplified descriptions of traditions. (3) **Potential Misuse**: users might over-rely on the system for cultural advice. Mitigations include **trust-tier weighting** (prioritizing authoritative sources), **anti-leak filtering** (preventing blind context following), and transparent reporting of the RAG Paradox. No personally identifiable information was used.

Acknowledgments

We thank the organizers of SemEval-2026 Task 7 (BLEnD) for curating a culturally diverse benchmark, the Kaggle community for computational infrastructure, and the open-source community for the tools and frameworks underlying this work.

References

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.

Explosion. 2023. spACy: Industrial-strength natural language processing in Python. https://spacy.io.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors. 2026. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, United States.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

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.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Meta AI. 2024. Llama 3 model card. https://llama.meta.com/llama3.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Nedjma Ousidhoum, Junho Myung, Carla Perez-Almendros, Jiho Jin, Amr Keleg, Meriem Beloucif, Yi Zhou, Rodrigo Agerri, Vladimir Araujo, Naomi Baes, James Barry, Joanne Boisson, Nancy F. Chen, Christine de Kock, Aleksandra Edwards, Joseba Fernandez de Landa, Mohamed Fazli Imam, Huda Hakami, Shu-Kai Hsieh, and 11 others. 2026. SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. 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 3982–3992.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. SimLM: Pre-training with representation bottleneck for document retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3585–3598.