HistRAG: Historical Retrieval-Augmented Generation and Cognitive-Aware Benchmarking of Historical Understanding for Generative AI

Anonymous ACL submission

Abstract

We present **HistRAG**, a benchmark for evaluating large language models (LLMs) on historically grounded reasoning. Historical understanding requires temporal grounding and the ability to connect events, actors, and locations across long narratives, capabilities not fully captured by general-purpose benchmarks. HistRAG is built from 13,233 structured events extracted from three book-length historical texts (84.4% of events originate from the Western-focused *Story of Civilization*). From a curated subset of 1,007 events we generate **4,991 multiple-choice questions** aligned with all six levels of Bloom's Taxonomy, from factual recall (*Remember*) to complex synthesis (*Create*). A human annotation study on 100 sampled higher-order questions confirms that AI-assigned answer keys are 95% accurate and Bloom-level labels reach κ=0.88 against expert consensus. We evaluate **17 instruction-tuned LLMs** in both *closed-book* and *retrieval-augmented generation* (RAG) settings. Under this evaluation setup, we observe three consistent patterns: (1) *Remember*-level questions yield the lowest accuracy despite requiring only factual recall, an accuracy inversion driven by temporal precision demands and interpreted conservatively, since higher-order levels are evaluated on an easier event subset (Section 4); (2) under our specific baseline dense-retrieval configuration (`all-MiniLM-L6-v2`, top-3 passages), retrieval augmentation improves lower cognitive levels but can *hurt* higher-order reasoning, likely via context noise; and (3) temporal grounding is a consistent weak point across all 17 models.

1 Introduction

The digitization of historical archives has created an enormous body of machine-readable text that remains largely under-utilized for NLP evaluation (Fafalios et al., 2023; Hawkins, 2021). At the same time, LLMs are increasingly deployed for historical research and education, yet the community lacks standardized protocols for measuring how well these systems actually understand history.

Historical reasoning imposes distinct demands on LLMs: precise temporal grounding, long-range dependency tracking across narrative text, and cultural context not uniformly represented in training corpora, where Western bias remains a concern (Kandpal et al., 2023; Keleg and Magdy, 2023). General-purpose benchmarks such as MMLU (Hendrycks et al., 2021) evaluate decontextualized short questions and do not differentiate cognitive depth. Dedicated historical resources are either small-scale, narrowly typed, or limited to factual recall (Rovera et al., 2019; Hervieux et al., 2024; Hauser et al., 2024; Qiu et al., 2025).

We address these gaps with HistRAG (Figure 1), an end-to-end resource comprising a large structured event dataset, a cognitively graded benchmark, and a standardized evaluation protocol.

Our contributions are:

1. **Event dataset.** 13,233 structured historical events from three book-length texts (Western, Islamic, and comparative religious history), stored in a unified schema. We note a cultural imbalance: 84.4% of events originate from the Western-focused *Story of Civilization*; coverage of other regions is uneven.

2. **Benchmark.** 4,991 multiple-choice questions over 1,007 events, aligned with all six Bloom's Taxonomy levels (Remember through Create). Higher-order questions are LLM-generated and validated by a human annotation study confirming answer-key accuracy (95.0%) and Bloom-level reliability (κ=0.875) against expert consensus.

3. **Evaluation protocol.** Standardized closed-book and RAG evaluation of 17 instruction-tuned LLMs. The RAG retriever is selected from an ablation over

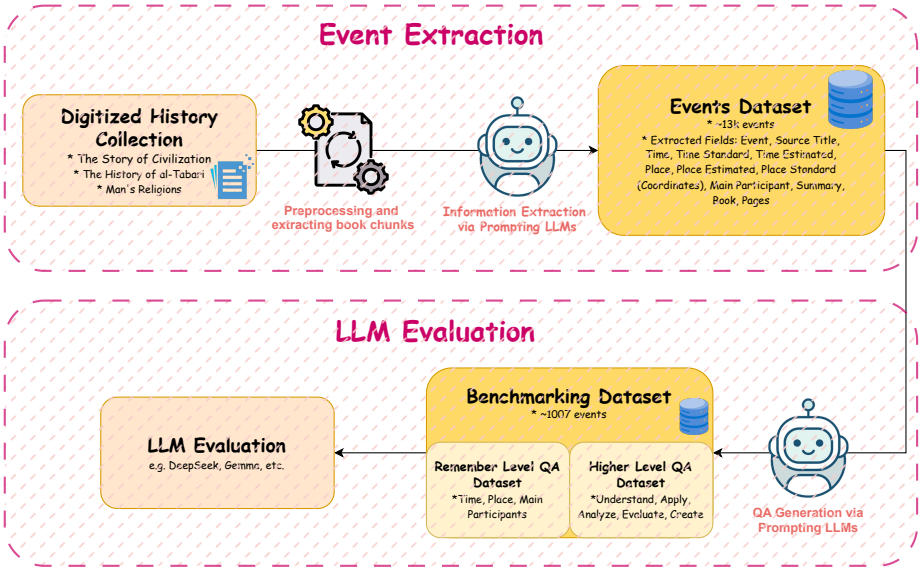


Figure 1: HistRAG pipeline. **Top:** structured events are extracted from digitized book-length texts via LLM prompting. **Bottom:** the event collection is used to generate multiple-choice questions spanning all six Bloom's Taxonomy levels and to evaluate LLMs in both closed-book and retrieval-augmented settings.

five candidates (BM25, off-the-shelf and domain-fine-tuned `all-MiniLM-L6-v2`, `multilingual-e5-large`, `Qwen3-Embedding-0.6B`) and two hybrid fusions; we adopt hybrid **BM25 + MiniLM-L6-v2** combined via Reciprocal Rank Fusion (top-3 passages), the best-performing configuration.

2 Related Work

Historical event extraction. Event extraction from historical text has been studied with GCN-based detection (Nguyen and Grishman, 2018), transformer entity extraction (Du et al., 2021), and domain-specific annotation schemes (Sprugnoli and Tonelli, 2019; Lai et al., 2021), as well as event-and-role identification in historical war memoirs (Rovera et al., 2019). Framing extraction as QA proved effective in historical settings (Liu et al., 2020; Borenstein et al., 2023). Related efforts annotate literary narratives for event detection (Sims et al., 2019) or auto-generate large question sets over books (Moskvichev and Mai, 2023). These efforts remain limited by small datasets, fixed event taxonomies, or clean and pre-segmented text; LLMs instead allow scalable, open-taxonomy extraction at book length, including from noisy, OCR-derived, and non-English sources.

Historical and temporal benchmarks. General-purpose suites such as MMLU (Hendrycks et al., 2021) include history questions but treat them as decontextualized factual recall. History-focused benchmarks narrow this: HiST-LLM (Hauser et al., 2024) covers historical societies from the Seshat Databank but focuses on factual recall without cognitive-level differentiation, and HistBench (Qiu et al., 2025) supports multimodal QA but is limited in scale. A parallel line targets temporal reasoning: TimeQA (Chen et al., 2021), TempReason (Tan et al., 2023), and MenatQA (Wei et al., 2023) probe time-sensitive QA and temporal sensitivity, while TRAM (Wang and Zhao, 2024) and TimeBench (Chu et al., 2024) aggregate ordering, duration, and arithmetic tasks. These benchmarks either build on already-structured data or test temporal skills over short, decontextualized items, and none combine narrative event extraction with Bloom's Taxonomy grading. HistRAG addresses this by extracting grounded events from raw historical narrative and separating factual recall from higher-order historical reasoning under closed-book and RAG evaluation.

Table 1: Source text statistics and event counts.

Source	Pages	Words	Chars	Events
Story of Civilization	9,570	4.24M	24.7M	11,176
History of al-Tabari	6,166	1.63M	8.1M	1,570
Man's Religions	768	0.30M	1.8M	487
Total	**16,504**	**6.17M**	**34.6M**	**13,233**

Table 2: Per-field extraction quality: annotator accuracy and inter-annotator agreement. † Fields where near-zero or negative κ is due to class imbalance, not annotation disagreement.

	Time	Time Est.	Time Std.	Place	Place Est.	Place Std.	Participants	Pages	Avg.
Ann. 1	90%	96%	98%	94%	100%	88%	96%	100%	95.3%
Ann. 2	88%	96%	88%	94%	98%	88%	98%	94%	93.0%
Avg.	89%	96%	93%	94%	99%	88%	97%	97%	94.1%
Agree. (%)	90.0	96.0	86.0	96.0	98.0	92.0	98.0	94.0	93.8
κ	.490	.479	$-.036^{\dagger}$	.645	$.000^{\dagger}$	.621	.658	$.000^{\dagger}$	.357

3 Dataset Construction

3.1 Source Texts

We draw on three book-length works (Table 1). Note that 84.4% of extracted events originate from *The Story of Civilization*, which has a strong Western and European narrative emphasis; geographic scope is therefore uneven, a limitation we discuss in Section 7.[1]

The Story of Civilization (Will & Ariel Durant, 1935–1975, 11 vols.) traces global history from prehistory through the Napoleonic era, with a narrative emphasis on European and Western developments (Durant, 1942).

The History of al-Tabari (*Tarikh al-Rusul wa al-Muluk*, completed 915 CE, 11 vols.) is an annalistic chronicle covering ancient empires, prophetic traditions, and Islamic history through the early Abbasid period (al Tabari and Rosenthal, 1988). We work directly with the Arabic original to avoid translation artefacts.

Man's Religions (John B. Noss, rev. ed. c. 1980s) is a single-volume comparative survey of global belief systems spanning primitive religions, Hinduism, Buddhism, East Asian religions, and Abrahamic traditions (Noss, 1956).

3.2 Preprocessing and Event Extraction

We digitized or obtained digital versions of all three sources (Durant, 2016; al Tabari, 1967), applying OCR correction, structural segmentation, and chunk-level annotation (title + page range) as preprocessing. Each chunk (1,500–3,000 words) is then passed to **GPT-4** (Achiam et al., 2023) (32k context) for structured event extraction, producing JSON records conforming to the schema in Table 4 (Appendix A).

Two extraction challenges required special handling. *Missing temporal or spatial data* was mitigated by supplying the 10 most recently extracted events as context; the model infers missing values from parametric knowledge and marks them with a Boolean `estimated` flag. Among the 1,007 benchmark events, 9.4% have estimated time values and 7.6% have estimated place values; these fields are source-grounded in the remainder. Because inferred values rely on GPT-4's parametric knowledge rather than the text itself, they may contain hallucinations. *Standardization* of time (to Gregorian year) and place (to lat/lon coordinates) was handled by dedicated prompt fields. The full extraction prompt is in Appendix B.

3.3 Quality Validation

Two independent annotators manually verified 50 randomly sampled events. Table 2 reports per-field accuracy and Cohen's κ. Overall accuracy is **94.1%**; overall $\kappa = 0.357$.

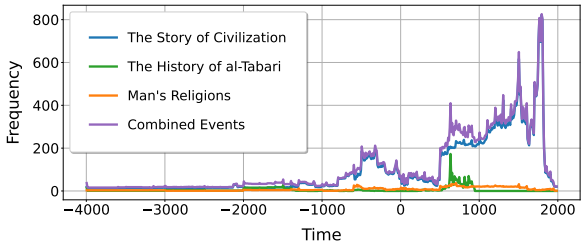
The sub-unity κ is primarily an artefact of class imbalance: when nearly all values in a field are correct, expected chance agreement approaches observed agreement, collapsing κ toward zero or below regardless of actual annotation quality (Gwet, 2008). Fields with genuine class variation (*Main Participants*: κ=0.658; *Place Std.*: κ=0.621; *Time*: κ=0.490) all reach moderate-to-substantial agreement. Negative values for *Time Std.* and near-zero values for *Pages* and *Place Est.* reflect fields where annotators agreed almost unanimously on a single class. Expanding validation to a larger, multi-annotator sample is left for future work.

4 The HistRAG Benchmark

4.1 Event Selection and Geographic Coverage

From 13,233 events we selected a representative subset of **1,007 events** balancing geographic regions, historical periods, and thematic domains, keeping the size feasible for manual quality checks. Figure 2 shows the temporal and geographic distributions of both the full dataset and the benchmark subset.

[1] No copyrighted text is reproduced; all content is paraphrased and attributed, consistent with fair-use/fair-dealing norms for non-commercial scholarly research.



(a) Temporal distribution of all 13,233 extracted events by source. Event density peaks in the last 1,500 years, reflecting the narrative scope of the source texts.

(b) Full dataset geographic distribution (13,233 events).

(c) Benchmark subset geographic distribution (1,007 events).

Figure 2: Dataset distributions. **(a)** Temporal distribution by source text. **(b)–(c)** Geographic distributions; color gradient: blue = older events, red = more recent; intensity reflects event density. The benchmark subset preserves the geographic spread of the full dataset. Of the 13,233 events, 84.4% originate from *The Story of Civilization*, which has a Western and European narrative emphasis; this concentration is a noted limitation of the current dataset.

4.2 Bloom's Taxonomy Design

Figure 3 illustrates the six cognitive levels we use. Each level targets a qualitatively different reasoning demand:

- **Remember**: Recall exact dates, place names, and participants.
- **Understand**: Explain the significance or context of an event.
- **Apply**: Use historical knowledge in a new or analogical context.
- **Analyze**: Identify causal relationships, effects, and patterns.
- **Evaluate**: Assess decisions, compare sources, or judge outcomes.
- **Create**: Synthesize counterfactuals or alternative interpretations.

4.3 Question Generation

Remember level. For each of the 1,007 events, GPT-4o Mini (OpenAI, 2024) generated three multiple-choice questions targeting *time*, *place*, and *main participants*, yielding **3,021 questions**. Context was enriched by supplying the five events immediately preceding and following the target in the source text.

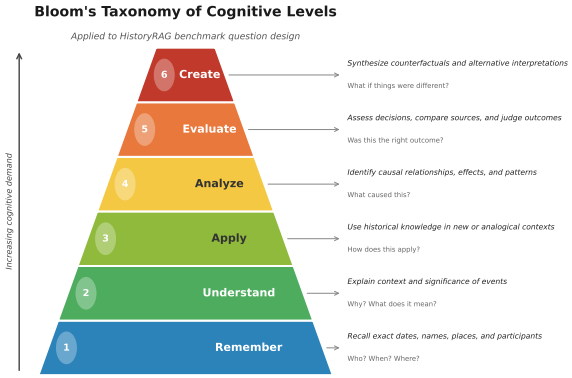


Figure 3: Bloom's Taxonomy pyramid used in HistRAG. Each band (bottom to top) represents an increasing level of cognitive demand, from factual *Remember* to creative *Create*. Right-hand annotations show the key reasoning demand and guiding question for each level.

Higher-order levels. Only events answered correctly by *all* evaluated models at Remember level were retained for higher-order benchmarking. This filtered set of **394 events** (39.1%) received five additional questions (one per level), initially generated by GPT-4o Mini (OpenAI, 2024) and refined by DeepSeek-V3 (DeepSeek-AI et al., 2025) to increase cognitive depth. This yields **1,970** higher-order questions.

Total: 4,991 questions across 6 Bloom levels from 1,007 events. The question generation prompts are provided in Appendix B.

Note on level comparability. Remember is evaluated on all 1,007 events, while higher-order levels cover only the 394-event subset for which *all* models succeeded at Remember. This subset is, by construction, drawn from events LLMs find easier to recall, so comparisons of accuracy *across* levels should be interpreted with caution. The design choice makes the observed Remember-is-hardest result *conservative*: even on a harder full set, higher-order levels would be evaluated on easier events. A matched-set cross-level comparison is left for future work.

4.4 Question Quality Validation

To establish the reliability of the higher-order questions, we conducted a human annotation study on a randomly sampled subset of 100 questions drawn from the 1,970 higher-order benchmark items.

Annotation protocol. Two domain-knowledgeable annotators independently evaluated each question *without* access to the originally assigned answer or Bloom level, eliminating anchoring bias. For each question, annotators recorded: **(i)** the correct answer choice (0–3), or flagged the item as *Question incorrect* if no single defensible answer could be identified; **(ii)** the Bloom's Taxonomy level (*Understand/Apply/Analyze/Evaluate/Create*); and **(iii)** the source consulted (source book alone, or additional reference). Annotators worked from the primary source texts cited in each question's metadata and were permitted to consult standard reference works (Wikipedia, Britannica, World History Encyclopedia) when the source pages alone were insufficient. After independent annotation, disagreements were adjudicated by a third annotator to produce a single consensus gold label per item. Annotators participated on a voluntary basis and received no compensation.

Inter-annotator agreement. Table 3 reports percent agreement and Cohen's κ for both dimensions. Answer agreement reached 89.0% ($\kappa = 0.854$) and level agreement reached 89.0% ($\kappa_{unw} = 0.862$; $\kappa_{lin} = 0.879$, where linear weights penalize adjacent-level confusions less than distant ones, consistent with Bloom's ordinal hierarchy). Both values fall in the *almost perfect* range ($\kappa > 0.80$) of the Landis–Koch scale (Landis and Koch, 1977). Of the 11 answer disagreements, two items were flagged as *Question incorrect* by one annotator (with written justifications citing ambiguous source passages), and nine reflected genuine uncertainty between adjacent answer choices. Of the 11 level disagreements, five involved the Analyze–Evaluate boundary, consistent with the known difficulty of separating these adjacent cognitive levels (Krathwohl, 2002).

Benchmark validity. We additionally compared the consensus gold labels against the AI-assigned answer keys and Bloom levels produced during benchmark construction (Table 3). AI-assigned answer keys matched human consensus on 95.0% of questions ($\kappa = 0.933$), confirming high factual accuracy of the automatically generated correct-answer labels. AI-assigned Bloom levels agreed with consensus on 90.0% of questions ($\kappa_{unw} = 0.875$; $\kappa_{lin} = 0.912$). The primary source of level error was the *Analyze* category, where the AI over-predicted *Evaluate* in five of ten level errors (76.0% per-level accuracy); all other levels achieved $\geq 89.5\%$ accuracy. Taken together, these results indicate that the benchmark's answer keys and cognitive-level labels are reliable within the validated sample, with agreement statistics in line with human-annotated NLP benchmarks.

Table 3: Question quality validation results on 100 sampled higher-order questions. **Top half**: inter-annotator agreement (Human 1 vs. Human 2). **Bottom half**: benchmark validity (consensus gold label vs. AI-assigned label). κ_{unw}: unweighted Cohen's κ; κ_{lin}: linear-weighted κ (Bloom's ordinal hierarchy; not applicable to the nominal Answer variable).

Comparison	Dimension	% Agree	κ_{unw}	κ_{lin}
Human 1 vs. Human 2	Answer	89.0	0.854	N/A
Human 1 vs. Human 2	Level	89.0	0.862	0.879
Gold vs. AI-assigned	Answer	95.0	0.933	N/A
Gold vs. AI-assigned	Level	90.0	0.875	0.912

5 Experimental Setup

Evaluation settings. We evaluate all models in two conditions: **(i) Closed-book**, where the model sees only the question and answer choices, and **(ii) Retrieval-Augmented Generation (RAG)**, where the top-3 passages retrieved from the source corpus are prepended to the question. To select the retriever used in the RAG condition, we conducted a preliminary ablation comparing five candidate systems on a held-out retrieval-evaluation split: a sparse lexical retriever (**BM25** (Robertson and Zaragoza, 2009)), three dense bi-encoders (*all-MiniLM-L6-v2* (Wang et al., 2020) both off-the-shelf and domain-fine-tuned, *multilingual-e5-large* (Wang et al., 2024), and *Qwen3-Embedding-0.6B* (Qwen Team, 2025)), and two hybrid configurations that fuse BM25 with each dense model via **Reciprocal Rank Fusion** (Cormack et al., 2009) (RRF, k=60). We compared retrievers on Recall@3 and downstream answer accuracy; the hybrid *BM25 + MiniLM-L6-v2 (RRF)* system consistently outperformed every individual retriever by combining the lexical precision of BM25 with the semantic generalisation of the dense encoder. All RAG experiments reported in this section therefore use this hybrid configuration as the retriever, with the Sentence-Transformers (Reimers and Gurevych, 2019) implementation of MiniLM (v5.1.1) and a FAISS (Johnson et al., 2021) inner-product index over the dense embeddings. Experiments were run with `sentence-transformers`

5.1.1, `transformers` 5.5.4, and `torch` 2.10.0.

Models. We evaluate 17 instruction-tuned LLMs spanning multiple families and scales (0.5B–24B parameters): *Mistral-Small-24B-Instruct-2501* (Mistral AI, 2025) and *Mistral-7B-Instruct-v0.3* (Jiang et al., 2023); *Fanar-1-9B-Instruct* (Abbas et al., 2025); *Qwen3-14B* and *Qwen3-1.7B* (Yang et al., 2025), *Qwen2.5-7B-Instruct* (Qwen Team, 2024b), and *Qwen1.5-0.5B* (Qwen Team, 2024a); *Gemma-3-12B-IT* (Gemma Team, 2025); *Granite-3.3-8B-Instruct* (IBM Granite Team, 2025a) and *Granite-4.0-Micro* (IBM Granite Team, 2025b); *Llama-3-8B-Instruct* and *Llama-3.2-1B* (Grattafiori et al., 2024); *Phi-4-Mini-Reasoning* (Xu et al., 2025); *DeepSeek-R1-Distill-Llama-8B* and *DeepSeek-R1-Distill-Qwen-1.5B* (DeepSeek-AI, 2025); and *Falcon-11B* and *Falcon-7B-Instruct* (Almazrouei et al., 2023).

Metric. We report **accuracy** (fraction of correct multiple-choice answers) per Bloom level and overall, comparing closed-book against the RAG condition. The predicted answer is extracted by parsing each model's generated text to identify the chosen option (0–3).

6 Results

Figure 4 summarizes our main findings. Panel (a) shows mean accuracy per Bloom level across all 17 models; panel (b) shows the RAG improvement (Δ percentage points) per model and level. Full numerical results are in Appendix E.

Finding 1: Remember yields the lowest observed accuracy. Despite requiring only factual recall (the lowest cognitive demand in Bloom's hierarchy), Remember achieves the lowest mean accuracy in both settings (CB: 35.7%, RAG: 44.1%) while Evaluate and Analyze reach 69.5% and 60.6% (CB) and 65.6% and 59.3% (RAG). Top models widen the gap further: on Mistral-Small-24B, Remember accuracy is 42.93% (CB) and 49.02% (RAG), versus >77% on every higher-order level. Remember questions demand verbatim-precise temporal or geographic facts, leaving no room for inference or partial credit, with time sub-questions the primary bottleneck. This pattern is consistent with findings that LLMs struggle with long-tail factual knowledge (Kandpal et al., 2023). *Interpretation note.* Because Remember is evaluated on all 1,007 events while higher-order levels use only the 394-event subset on which *all* models already succeeded at Remember, the higher-order accuracy figures reflect an inherently easier subset. This means the accuracy inversion reported here is a *conservative lower bound* on the gap, not an overestimate; the full discussion appears in Section 4.3.

Finding 2: Hybrid RAG helps lower cognitive levels but not higher ones. Panel (b) shows a consistent pattern under our hybrid BM25+MiniLM RRF retriever: augmentation strongly improves Remember and Understand (median $\Delta \approx$+8.1 pp and +14.5 pp respectively) but provides no aggregate gain at Apply ($\Delta \approx 0$ pp) and small losses at Analyze, Evaluate, and Create (-1.3, -3.9, -2.6 pp on average). The losses are concentrated in specific models: Qwen3-14B drops from 75.63% to 67.51% on Analyze (-8.12 pp) and from 55.33% to 51.27% on Create (-4.06 pp), and DeepSeek-R1-Distill-Llama-8B collapses from 57.61% to 30.71% on Analyze (-26.9 pp) as its long-chain reasoning is derailed by the retrieved context. We hypothesize that retrieved passages can introduce *context noise*: a topically related but causally mismatched passage can mislead models performing multi-step causal reasoning. Even though the hybrid retriever was chosen for its superior recall, the present results suggest that higher-order historical reasoning benefits less from additional context than from better filtering or reranking, an avenue we leave to future work.

Finding 3: Temporal grounding is a structural weakness. Temporal sub-questions within Remember level show the largest and most consistent gap: all 17 models fail at rates far above their errors on place or participant questions, and this gap is only partially closed by RAG. Even when a passage containing the relevant date is retrieved, models frequently fail to extract or correctly apply the temporal information.

Finding 4: Model scale and family both matter. Mistral-Small-24B achieves the best closed-book overall accuracy (64.06%) and the best RAG accuracy (66.14%). In closed-book, Fanar-1-9B (61.08%) and Mistral-7B (60.22%) follow; under RAG the mid-tier reshuffles so that Mistral-7B (63.78%) overtakes Fanar-1-9B (62.35%) for second place, with Qwen2.5-7B-Instruct (61.81%) close behind. Model family matters as much as scale: Mistral-7B outperforms Qwen3-14B on Create (72.59% vs. 51.27% under RAG), and Fanar-

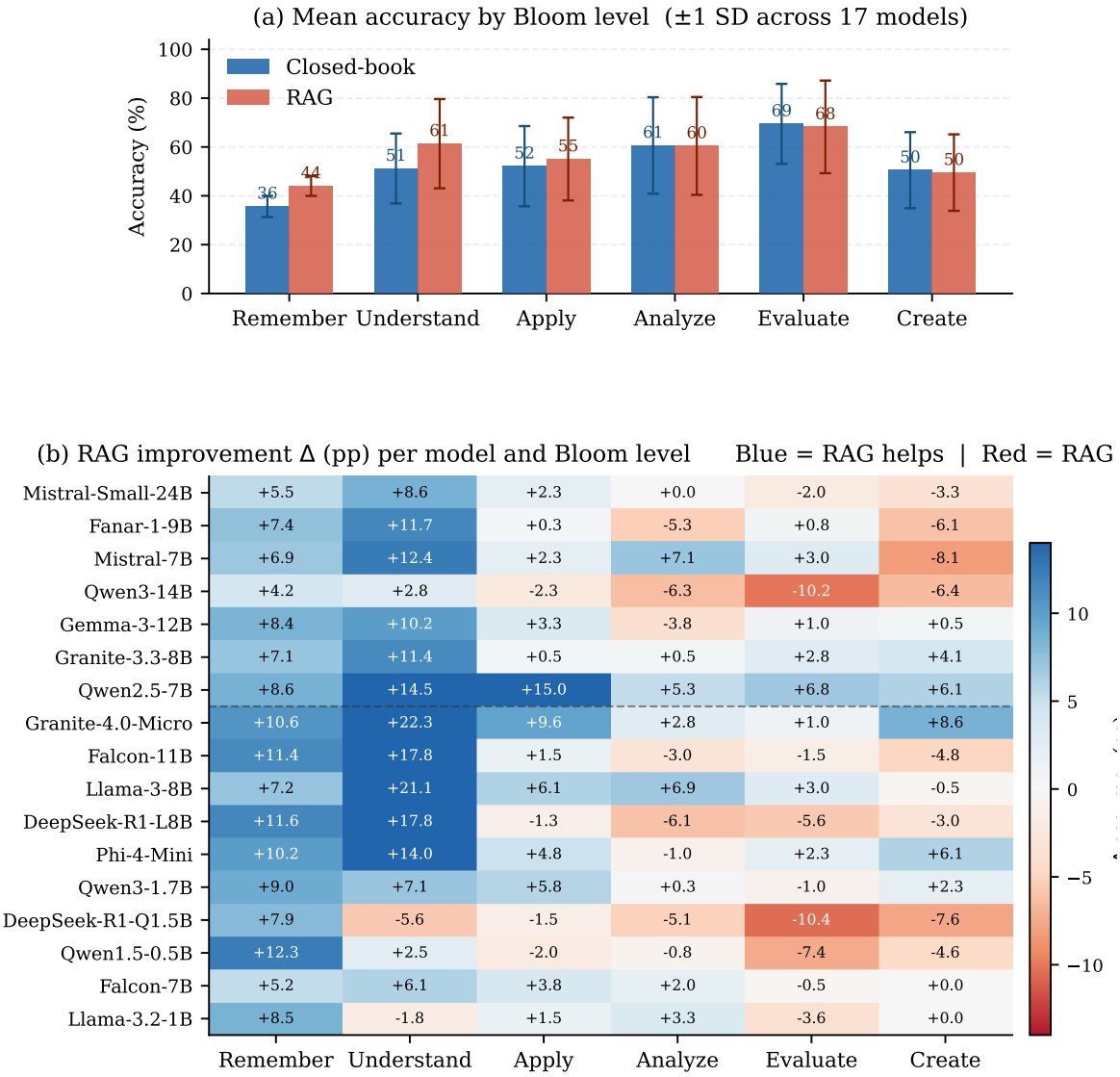


Figure 4: **(a)** Mean accuracy (±std across 17 models) per Bloom level for closed-book and RAG settings. **(b)** Per-model RAG improvement over closed-book (green = RAG helps, red = RAG hurts). Models are ranked by closed-book overall accuracy (highest at top). The dashed line separates the top-7 from the bottom-10 models.

1-9B (a domain-relevant Arabic–English model) ranks among the top three in both settings despite its smaller size, suggesting domain alignment advantages. Small models (<2B params) generally fall below 40% overall, with Llama-3.2-1B and Qwen1.5-0.5B near 30%.

Finding 5: Error patterns differ by cognitive level. At *Remember*, errors are predominantly temporal (century confusion, dynasty misattribution, event conflation). At *Apply* and *Understand*, over-generalization is common (applying context from one civilization to an incompatible one). At *Analyze*, models misidentify plausible but incorrect causal chains. At *Evaluate* and *Create*, errors are less severe (models select defensible but suboptimal answers), suggesting that general reasoning capacity broadly available in large models partially compensates for gaps in historical knowledge.

7 Conclusion

We introduced HistRAG, a structured event dataset, cognitively graded benchmark, and evaluation protocol for historical reasoning with and without retrieval augmentation. Across 17 LLMs, three consistent findings emerge: factual recall of temporal information is harder than higher-order reasoning (though levels differ in event coverage; see Section 4.3); under our hybrid BM25+MiniLM RRF

retriever selected after an ablation over five candidates augmentation improves lower cognitive levels but is neutral or slightly harmful at higher ones; and temporal grounding remains a structural limitation of current LLMs. A human annotation study on 100 sampled higher-order questions confirmed that answer keys and Bloom level assignments are reliable ($\kappa > 0.85$ for inter-annotator agreement; 95% and 90% accuracy for AI-assigned answers and levels respectively against human consensus), though validation coverage remains partial. Important caveats include source concentration (84.4% Western-centric) and the use of a single retriever without ablations. HistRAG provides a foundation for studying how LLMs acquire and apply historical knowledge, and for developing retrieval strategies that better support complex historical inference.

Limitations

Source concentration. 84.4% of extracted events originate from *The Story of Civilization*, resulting in stronger European coverage relative to sub-Saharan Africa, the Americas, and South/Southeast Asia. Expanding to additional primary sources from under-represented regions is a natural direction for future work.

Validation coverage. Both validation studies, namely event extraction (50 events) and question quality (100 items; Section 4.4), sample fractions of the full dataset. Agreement and accuracy statistics are strong within these sampled scopes; broader coverage is left for future work. A small fraction of benchmark events (9.4% time, 7.6% place) carry GPT-4-inferred values that may contain estimation errors.

Cross-level comparability. Remember is evaluated on all 1,007 events while higher-order levels use the easier 394-event subset; accuracy comparisons across levels should be interpreted accordingly (Section 4).

Ethical Considerations

We disclose that **Claude**, **GPT-4o**, and **Gemini** were used to assist in rewriting and improving manuscript drafts across revision rounds. All scientific content, experimental design, and conclusions are the authors' own responsibility.

The annotation study (Section 4.4) involved no external participants; no institutional ethics review was required.

Historical sources reflect incomplete records, cultural biases, and contested interpretations. Extracted events and questions may reproduce over-representation of certain regions or narratives; users should not treat benchmark outputs as a comprehensive or neutral account of history. LLM-assisted extraction may introduce extraction errors or misleading paraphrases; bibliographic and page-level references support verification against original sources. The dataset is released under the Creative Commons Attribution–NonCommercial 4.0 International license (CC BY-NC 4.0).

References

Ummar Abbas and 1 others. 2025. Fanar: An arabic-centric multimodal generative AI platform. *arXiv preprint arXiv:2501.13944*.

Josh Achiam and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

A.J.M.J. al Tabari and F. Rosenthal. 1988. *The History of al-Tabari*. Bibliotheca Persica. State University of New York Press.

Ibn Jarir al Tabari. 1967. Tarikh al-rusul wa al-muluk (the history of al-tabari). https://www.ghbook.ir/index.php?option=com_dbook&task=viewbook&book_id=9678&lang=fa. Accessed online.

Ebtesam Almazrouei and 1 others. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Nadav Borenstein, Natália da Silva Perez, and Isabelle Augenstein. 2023. Multilingual event extraction from historical newspaper adverts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10304–10325. Association for Computational Linguistics.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. *Preprint*, arXiv:2311.17667.

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. ACM.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

DeepSeek-AI and 1 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Xinya Du, Alexander Rush, and Claire Cardie. 2021. GRIT: Generative role-filler transformers for document-level event entity extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 634–644. Association for Computational Linguistics.

W. Durant. 1942. *The Story of Civilization*. Simon and Schuster.

Will Durant. 2016. The story of civilization (complete). https://archive.org/embed/TheStoryOfCivilizationcomplete. Accessed: 2016-12-22.

P. Fafalios, Yannis Marketakis, A. Axaridou, Yannis Tzitzikas, and M. Doerr. 2023. A workflow model for holistic data management and semantic interoperability in quantitative archival research. *Digital Scholarship in the Humanities*.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Aaron Grattafiori and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kilem L. Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Jakob Hauser, Daniel Kondor, Jenny Reddish, Majid Benam, Enrico Cioni, Federica Villa, James S. Bennett, Daniel Hoyer, Pieter Francois, Peter Turchin, and R. Maria del Rio-Chanona. 2024. Large language models' expert-level global history knowledge benchmark (hist-llm). In *Advances in Neural Information Processing Systems*, volume 37, pages 32336–32369. Curran Associates, Inc.

Ashleigh Hawkins. 2021. Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web. *Archival Science*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Natalie Hervieux, Peiran Yao, Susan Brown, and Denilson Barbosa. 2024. Language resources from prominent born-digital humanities texts are still needed in the age of llms. *NLP4DH*.

IBM Granite Team. 2025a. Granite-3.3-8B-instruct. https://huggingface.co/ibm-granite/granite-3.3-8b-instruct.

IBM Granite Team. 2025b. Granite-4.0-micro-instruct. https://huggingface.co/ibm-granite/granite-4.0-micro-instruct.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

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

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. *Preprint*, arXiv:2211.08411.

Amr Keleg and Walid Magdy. 2023. Dlama: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. *Preprint*, arXiv:2306.05076.

David R. Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4):212–218.

Viet Dac Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. 2021. Event extraction from historical texts: A new dataset for black rebellions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2390–2400. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651. Association for Computational Linguistics.

Mistral AI. 2025. Mistral small 3. https://mistral.ai/news/mistral-small-3.

Arsenii Moskvichev and Ky-Vinh Mai. 2023. NarrativeXL: A large-scale dataset for long-term memory models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15058–15072, Singapore. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI'18. AAAI Press.

John B. Noss. 1956. *Man's Religions*. The Macmillan Co.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Jiahao Qiu and 1 others. 2025. On path to multimodal historical reasoning: Histbench and histagent. *Preprint*, arXiv:2505.20246.

Qwen Team. 2024a. Qwen1.5. https://qwenlm.github.io/blog/qwen1.5/.

Qwen Team. 2024b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Qwen Team. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation model integration. *arXiv preprint arXiv:2506.05176*.

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. Association for Computational Linguistics.

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

Marco Rovera, F. Nanni, and Simone Paolo Ponzetto. 2019. Providing advanced access to historical war memoirs through the identification of events, participants and roles. *arXiv.org*.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Rachele Sprugnoli and Sara Tonelli. 2019. Novel event detection and classification for historical texts. *Computational Linguistics*, 45(2):229–265.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14820–14835, Toronto, Canada. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Yuqing Wang and Yun Zhao. 2024. Tram: Benchmarking temporal reasoning for large language models. *Preprint*, arXiv:2310.00835.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. MenatQA: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1434–1447, Singapore. Association for Computational Linguistics.

Haoran Xu and 1 others. 2025. Phi-4-mini-reasoning: Exploring the limits of small reasoning language models in math. *arXiv preprint arXiv:2504.21233*.

An Yang and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Event Representation Schema

Table 4: Unified schema for extracted historical events.

Field	Description
Event	Short title or description
Source title	Title as it appears in the original text
Time (original)	Temporal description from the source
Time (Gregorian)	Normalized year; negative = BCE
Time estimated	`true` if inferred; `false` if explicit
Place	Geographical location name
Place estimated	`true` if inferred; `false` if stated
Coordinates	Standardized latitude and longitude
Main participants	Key individuals or groups involved
Summary	Concise model-generated summary
Source reference	Source name and volume
Page range	Start and end page in the source

B Extraction and Generation Prompts

B.1 Event Extraction Prompt (GPT-4, 32k context)

Each extraction call receives three components: **(1) System instruction**: act as a historical event extractor; return a JSON array following the schema in Table 4; infer missing time or place from training knowledge and set the `estimated` flag. **(2) Recent events**: the 10 most recently extracted events as context for temporal/geographic continuity. **(3) Target passage**: current chunk with title and page range.

```
Recent events (last 10):
{recent_events_json}
Passage: {chunk_title}
(pp. {start}-{end})
Text: {chunk_text}
Extract all historical events. Return a
JSON array. Each event must include all
fields from the schema.
```

B.2 Remember-Level Question Generation

```
You are a historical quiz generator.
For the event below, generate
3 multiple-choice questions: one
about TIME, one about PLACE, and one
about MAIN PARTICIPANTS.
Event: {event_json}
Context (5 events before and after):
{context_json}
For each question: clear stem, 4
choices (A-D), correct answer indicated,
plausible distractors. Return JSON.
```

B.3 Higher-Order Question Generation (GPT-4o Mini + DeepSeek-V3 refinement)

```
You are an expert assessment designer.
Generate one multiple-choice question at
Bloom's Taxonomy level {LEVEL} for the
event below.
Event: {event_json}
Context (5 events before and after):
{context_json}
Level definition: {level_definition}
Requirements: (i) genuinely requires
{LEVEL}-level thinking; (ii) 4 choices;
(iii) unambiguous correct answer;
(iv) plausible distractors; (v) not
answerable by simple recall. Return
JSON.
```

After initial generation, DeepSeek-V3 was prompted to critique and revise each question for greater cognitive depth.

C Temporal Distribution by Volume

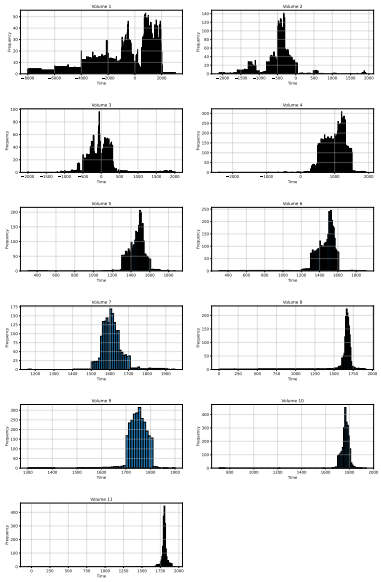


Figure 5: Temporal distribution of events extracted from *The Story of Civilization* by volume. Later volumes concentrate increasingly on the post-1200 CE period.

D Question Quality Annotation Guideline

This appendix reproduces the instructions given to annotators for the question quality validation study described in Section 4.4.

Overview

Annotators validated 100 questions from the HistRAG higher-order benchmark. Questions were generated automatically by GPT-4o Mini and refined by DeepSeek-V3. For each question, annotators recorded exactly three things:

1. **Answer:** Which of the four choices is correct (or that the question is flawed).
2. **Level:** Which Bloom's Taxonomy level the question belongs to.
3. **Source:** Whether the answer was found from the book alone, or required additional sources.

Important: Both the correct answer and the assigned Bloom's level were hidden from annotators. Annotators had to arrive at both independently. This eliminates anchoring bias: seeing a pre-marked answer causes annotators to unconsciously seek confirming evidence rather than evaluating all options equally.

Source Books

Each question includes a **Book** code and a **Pages** range. Annotators always started there. If the source page was not sufficient, annotators could use Wikipedia, Britannica, World History Encyclopedia, or Google Scholar.

Book code	Book
HistOfCiv	*The Story of Civilization* (Durant, 11 vols.)
Tabari	*History of al-Tabari* (Tabari, 11 vols.)
Man's Religions	*Man's Religions* (Noss)

Recording the Answer

Annotators read the question and all four choices carefully, consulted the source, then selected one of five options:

Option	When to select it
0	Choice 0 is the correct answer.
1	Choice 1 is the correct answer.
2	Choice 2 is the correct answer.
3	Choice 3 is the correct answer.
Question incorrect	No single choice is clearly correct: the question is ambiguous, more than one option is defensible, contains a factual error, or no answer can be verified from available sources.

Whenever *Question incorrect* was selected, annotators wrote a brief note explaining why (ambiguous wording, two defensible answers, factual error in the stem, etc.).

Recording the Bloom's Taxonomy Level

After answering, annotators decided which Bloom's Taxonomy level the question belongs to:

Level	What it requires	Typical signals in the stem
Understand	Explain or interpret what happened. Answer is directly stated or closely paraphrased in the source.	*What was the purpose of…? Why did X happen? What does X mean?*
Apply	Transfer historical knowledge to a new or hypothetical situation. Requires reasoning beyond the text into an analogy or counterfactual.	*If X had not happened…? Which modern situation parallels X?*
Analyze	Identify causes, effects, or relationships not explicitly stated in one passage. Requires connecting information across the text.	*What was the relationship between X and Y? What caused…? What pattern does X illustrate?*
Evaluate	Judge a historical decision, strategy, or interpretation using a criterion. There must be one defensible best answer, not open opinion.	*How justified was X's decision? Was X's strategy effective? Which interpretation best explains…?*
Create	Design or synthesize something new from historical knowledge: a research proposal, a counterfactual, an alternative plan.	*Design a research project that…/ Construct a counterfactual where…*

The most common error: disguised recall. If a question can be answered by reading a single passage with no reasoning beyond paraphrasing, it is **Understand** regardless of how it is worded. The key question is: *Does answering require connecting information or reasoning beyond the text, or just finding a sentence in the passage?*

Key distinctions.

- **Understand vs. Analyze:** Understand = answer is stated in one passage. Analyze = answer requires identifying an unstated relationship across the text.
- **Apply vs. Analyze:** Apply = use history in a new/hypothetical situation. Analyze = dissect

the internal workings of a past event.

- **Evaluate vs. Create:** Evaluate = judge something that already exists. Create = build or design something new.

Recording the Source

Option	When to select it
Book only	Sufficient evidence was found from the source book pages listed in the question metadata alone.
Additional sources	Wikipedia, Britannica, World History Encyclopedia, Google Scholar, or any source beyond the listed book pages was needed.

Annotation Sheet Summary

Column	What to enter	Options
ID	Question identifier	1–100
Answer	Chosen correct answer	0 / 1 / 2 / 3 / Question incorrect
Level	Assigned Bloom's level	understand / apply / analyze / evaluate / create
Source	Where the answer was found	Book only / Additional sources
Answer Notes	Required if Question incorrect: brief reason	Free text
Level Notes	Optional: brief justification if choice was not obvious	Free text

Practical Notes

- **Annotate independently.** Do not discuss answers or levels with the other annotators until all have finished their assigned questions.
- **Two annotators** evaluated each question. Cohen's κ was computed on both Answer and Level. Disagreements were adjudicated by a third annotator.

E Full Numerical Results

Table 5 reports accuracy for all 17 models across all six Bloom levels in both evaluation settings.

Model	Closed-book						
	Rem.	**Und.**	**App.**	**Ana.**	**Eva.**	**Cre.**	**Overall**
Mistral-Small-24B-Instruct	**42.93**	**77.41**	77.92	**83.50**	**88.07**	74.37	**64.06**
Fanar-1-9B-Instruct	40.45	69.04	**78.17**	82.99	84.77	69.54	61.08
Mistral-7B-Instruct-v0.3	39.39	67.77	74.11	79.19	83.50	**76.40**	60.22
Qwen3-14B	40.78	63.71	62.18	75.63	85.03	55.33	56.40
Gemma-3-12B-IT	38.66	64.72	67.01	76.65	79.19	60.66	56.20
Granite-3.3-8B-Instruct	37.93	61.93	62.18	78.68	82.49	57.36	55.25
Qwen2.5-7B-Instruct	37.87	56.85	54.82	74.11	81.98	65.48	54.16
Granite-4.0-Micro	35.39	49.24	54.82	66.75	79.70	49.75	49.35
Falcon-11B	35.62	48.98	53.05	62.69	72.59	51.78	48.18
Llama-3-8B-Instruct	37.84	44.67	43.40	64.97	70.05	51.52	47.51
DeepSeek-R1-Distill-Llama-8B	35.32	46.45	52.28	57.61	69.80	49.49	46.53
Phi-4-Mini-Reasoning	32.14	53.30	48.48	64.97	74.62	40.86	45.90
Qwen3-1.7B	34.56	37.82	38.07	42.89	56.85	39.09	39.30
DeepSeek-R1-Distill-Qwen-1.5B	29.73	33.25	34.26	38.83	44.92	35.03	33.99
Qwen1.5-0.5B	29.89	31.47	32.74	28.43	43.40	28.17	31.56
Falcon-7B-Instruct	27.90	31.22	28.68	25.63	46.45	30.71	30.53
Llama-3.2-1B	29.79	32.23	24.11	26.90	37.82	22.84	29.22
Model	**RAG (top-3 retrieved passages)**						
	Rem.	**Und.**	**App.**	**Ana.**	**Eva.**	**Cre.**	**Overall**
Mistral-Small-24B-Instruct	**49.02**	**91.88**	**81.73**	84.01	86.55	72.08	**66.14**
Fanar-1-9B-Instruct	47.04	83.76	76.40	80.71	85.53	61.93	62.35
Mistral-7B-Instruct-v0.3	47.00	85.03	75.13	**85.03**	85.03	**72.59**	63.78
Qwen3-14B	46.34	66.50	56.60	67.51	77.16	51.27	55.07
Gemma-3-12B-IT	48.43	80.96	70.30	72.59	77.66	60.66	60.43
Granite-3.3-8B-Instruct	46.87	79.44	59.90	78.17	81.73	60.91	59.45
Qwen2.5-7B-Instruct	47.43	76.40	67.51	79.44	**88.83**	68.78	61.81
Granite-4.0-Micro	47.53	70.56	59.90	68.02	77.92	60.91	57.50
Falcon-11B	48.13	72.59	51.52	55.84	70.05	45.18	53.58
Llama-3-8B-Instruct	45.95	65.23	49.24	69.80	76.40	51.78	54.22
DeepSeek-R1-Distill-Llama-8B	40.75	44.92	30.46	30.71	30.96	24.37	36.52
Phi-4-Mini-Reasoning	41.64	62.69	51.27	60.41	67.01	44.42	49.40
Qwen3-1.7B	45.65	51.02	41.88	53.55	65.99	42.64	48.33
DeepSeek-R1-Distill-Qwen-1.5B	37.47	34.77	31.22	36.29	39.59	29.44	35.87
Qwen1.5-0.5B	41.18	31.73	30.20	27.16	31.98	21.83	34.88
Falcon-7B-Instruct	30.88	36.29	27.66	30.96	40.61	26.40	31.63
Llama-3.2-1B	38.20	32.74	25.13	27.66	32.23	19.04	32.78

Table 5: Full results: accuracy (%) per model, Bloom level, and evaluation setting. **Bold** = best per column within each setting block. Models sorted by closed-book overall accuracy (descending).