Palestine-RAG: Retrieval-Augmented Generation for Historically and Factually Grounded QA on the Palestine Conflict

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper presents Palestine-RAG, a domain-specific Retrieval-Augmented Generation (RAG) framework developed to counter the underrepresentation and mischaracterization of Palestinian history, legal discourse, and current events in mainstream language models capable of bilingual response generation in both Arabic and English. We construct a high-quality, culturally informed dataset by aggregating content from authoritative sources including Palquest.org, United Nations resolutions, International Court of Justice (ICJ) rulings, historical archives, and reputable news outlets. To evaluate model performance, we introduce the first multiple-choice question (MCQ) benchmarking dataset for this domain, comprising 222 manually crafted questions systematically categorized according to Bloom's Taxonomy to capture varying levels of cognitive complexity. We benchmark 26 language models and demonstrate that retrieval-augmented approaches consistently outperform non-retrieval large language models in both factual accuracy and depth of reasoning, particularly within politically nuanced and historically complex contexts.

1 Introduction

Large language models (LLMs) have achieved impressive performance across a range of natural language tasks, including question answering, translation, and dialogue generation Chen et al. (2020); Elbeltagy and Khalifa (2024). However, increasing evidence shows that these models often exhibit systemic biases, particularly when addressing politically sensitive or underrepresented topics Roozado and Zhang (2024); Kim and Wang (2025). Such limitations stem from unbalanced training data, limited cultural and linguistic diversity, and a lack of exposure to non-Western epistemologies.

The representation of Palestinian history and contemporary issues remains notably sparse and imbalanced in mainstream LLMs. While platforms such as Shu'un Filastiniyyah, the Journal of Palestine Studies, and the Jerusalem Quarterly have preserved critical scholarly discourse, their content is largely absent from the training data of widely deployed models. Recent initiatives, including the Palestinian Land Studies Centre and The Untold Story of the Palestinian Revolution, offer rich, contextualized archives that are essential for building historically grounded AI systems Pappé et al. (2024).

To address this gap, we introduce **Palestine-RAG**, a retrieval-augmented generation framework designed for accurate, context-sensitive question answering grounded in verified legal, historical, and academic sources. By combining dense retrieval with a curated bilingual corpus—including UN resolutions, ICJ rulings, and region-specific publications—our system enhances factual accuracy, interpretability, and cultural alignment Asai et al. (2024); Gan and Zhou (2025).

This work contributes to the growing effort to develop ethical, domain-aware AI systems. Palestine-RAG demonstrates how retrieval-augmented approaches can mitigate hallucination and bias, while promoting transparency and trustworthiness in politically complex domains. The Palestine-RAG project offers several key contributions to the field of historically and factually grounded natural language generation:
(i) Palestine-RAG: We develop the first Retrieval-Augmented Generation (RAG) system focused on the Palestinian context, grounded in factual, historical, and legal sources to ensure verifiable and contextually accurate responses. The system supports citation-based response generation to enhance transparency Karpukhin et al. (2020); Lewis et al. (2020); Al-Roumi and Hasan (2023); Kiela and Gupta (2025); Gan and Zhou (2025).

(ii) MCQ-based Benchmarking: We introduce the first multiple-choice question (MCQ) benchmark for the Palestinian domain, featuring 222 manually curated questions categorized by Bloom's Taxonomy to evaluate models across different levels of cognitive complexity.

(iii) Comprehensive Evaluation: We benchmark Palestine-RAG against 25 open-source large language models (LLMs) using our dataset, enabling a nuanced assessment across comprehension, application, and analytical reasoning levels Roozado and Zhang (2024); Asai et al. (2024).

2 Approach for Palestine-RAG

Palestine-RAG is a domain-adapted Retrieval-Augmented Generation (RAG) system tailored to the Palestinian historical, legal, and political context. The overall pipeline consists of four key stages, as depicted in Figure 1:
(i) Data Curation and Preprocessing: Prior to the RAG workflow, we curate a high-quality corpus from authoritative sources, including legal documents (e.g., UN resolutions, ICJ rulings), historical archives, academic journals, and reputable news outlets. Texts are normalized, deduplicated, and segmented to support efficient retrieval and alignment with bilingual generation.
(ii) Document Encoding: The preprocessed corpus is encoded into high-dimensional vector representations using a dense retriever (e.g., DPR Karpukhin et al. (2020)). This enables semantic similarity search over a knowledge base that captures both Arabic and English content.
(iii) Passage Retrieval: Given a user query, the retriever identifies the top-k most relevant passages based on embedding similarity. This retrieval step provides the factual grounding necessary for accurate and context-aware generation.
(iv) Bilingual Response Generation: Retrieved passages are passed to a locally hosted language model, which generates fluent, contextually relevant answers in both Arabic and English. The generation process is optimized for citation transparency, factual consistency, and cultural sensitivity. The Palestine-RAG produces grounded bilingual responses that reflect domain-specific knowledge while mitigating the risk of hallucination and bias common in general-purpose LLMs.

3 Dataset for Palestine-RAG

The **Palestine-RAG** dataset is composed of two core components: (1) a multilingual retrieval corpus used to support factual grounding in the RAG pipeline, and (2) a benchmarking dataset of multiple-choice questions (MCQs) designed for systematic evaluation across cognitive dimensions.

3.1 RAG Knowledge Corpus

We curated a comprehensive dataset of over 50 unique documents from authoritative and context-rich sources to ensure robust coverage of Palestine-related content. The Palestine-RAG dataset comprises 41 legal documents from the United Nations and International Court of Justice (ICJ), 2 peer-reviewed academic publications, and 175 analytical reports from Palquest.org, complemented by 4 historical archives and over 5k multilingual journalistic articles from reputable news organizations including Al Jazeera, Middle East Eye, and Reuters. This meticulously assembled corpus spans 1 document across 22 distinct thematic domains—encompassing occupation, displacement, resistance, apartheid, international law, and diplomacy—totaling 57 documents to facilitate comprehensive and contextually-aware retrieval-augmented generation.

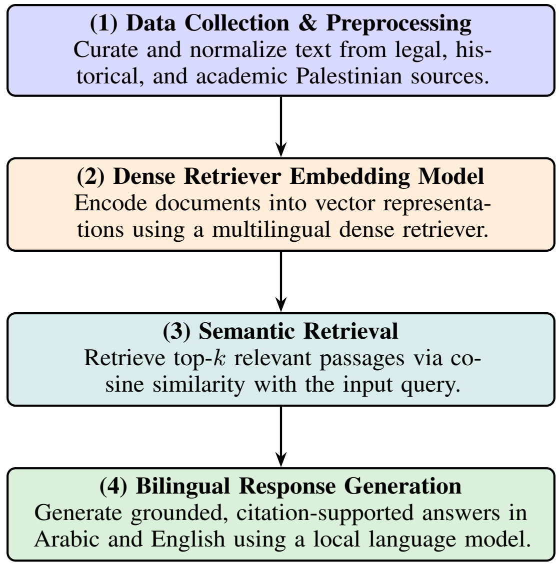

Figure 1: Architecture of the Palestine-RAG system. The four-stage pipeline includes data curation, vector embedding, semantic retrieval, and bilingual response generation with citation using a local language model.

3.2 MCQ Benchmarking Dataset

To evaluate the question-answering performance of our system across different levels of reasoning, we constructed a structured MCQ benchmark aligned with Bloom's Taxonomy, we constructed a new benchmarking dataset comprising 222 multiple-choice questions (MCQs), each manually reviewed for quality and relevance. Each question consists of one correct answer and three plausible distractors, and is annotated according to Bloom's Taxonomy Krathwohl (2002), enabling evaluation across six cognitive levels: *Remember* (e.g., factual recall of events or dates), *Understand* (e.g., paraphrasing legal articles), *Apply* (e.g., using historical context in novel scenarios), *Analyze* (e.g., comparing competing narratives), *Evaluate* (e.g., assessing legal or historical claims), and *Create* (e.g., generating policy-oriented recommendations). This categorization allows for fine-grained analysis of model reasoning capabilities beyond surface-level accuracy, particularly in politically and historically sensitive contexts.

Initial question generation was performed using the instruction-tuned DeepSeek-R1 model, guided by prompts designed to ensure historical grounding and topic coverage across both past and present Palestinian contexts (inspired from related Reddit Communities[1] and FODIP[2]) as well as Bloom's Taxonomy definitions. Sample prompt and question templates are provided in Appendix §A and §**??**. The questions span key domains such as legal history, political developments, and international discourse. A team of four trained reviewers conducted human validation of all items to ensure: (i) Historical and factual accuracy, (ii) Clarity and neutrality of language, (iii) Appropriateness and plausibility of distractors, and (iv) Correct alignment with Bloom's cognitive levels. Discrepancies or ambiguities identified during the review were resolved collaboratively, with revisions made to either the question phrasing or the answer choices to maintain academic rigor and factual integrity Gan and Zhou (2025); Kim and Wang (2025). This benchmark enables fine-grained evaluation of retrieval-augmented systems on both linguistic performance and reasoning depth in politically and culturally sensitive domains. Example question for different Bloom Taxonomy levels are provided in §C.

[1]e.g., https://www.reddit.com/r/IsraelPalestine/
[2]https://www.fodip.org.uk/

3.3 Dense Retriever Embedding Model

We adopt a dense retrieval framework in which both user queries and document passages are encoded into vector representations using the pre-trained `intfloat/e5-large-v2` model Karpukhin et al. (2020); Gan and Zhou (2025). The retrieval corpus—comprising historical records, legal documents, journalistic articles, and multilingual resources—is segmented into overlapping chunks using an adaptive strategy that maintains paragraph coherence while respecting model token limits.

Each document chunk is embedded by `intfloat/e5-large-v2` and stored in an in-memory vector database with efficient caching for low-latency retrieval access Asai et al. (2024); Gold and Liu (2024). To optimize retrieval, the document embeddings are precomputed and cached, ensuring that encoding is performed only once during the indexing stage. At inference time, user queries are similarly embedded, and—if query caching is enabled—identical (i.e., exact-match) queries reuse their previously computed embeddings to reduce latency and computational overhead. The system then retrieves the top-k semantically relevant passages ($k = 5$) using cosine similarity between the embedded query and document vectors Lewis et al. (2020); Gan and Zhou (2025).

3.4 Language Model

The retrieved passages are concatenated with the user query using a standardized prompt template. This prompt incorporates structured instructions aligned with Bloom's Taxonomy, allowing users to specify the desired cognitive level of the response (e.g., *Remember*, *Analyze*, *Create*) Elbeltagy and Khalifa (2024); Gan and Zhou (2025). Incorporating Bloom's levels guides the generation process by controlling response depth—prompting the model to recall facts, conduct comparisons, or synthesize insights depending on the user's goal. This is especially valuable in politically complex or educational domains where interpretability and granularity of reasoning are essential.

The final prompt is passed to a large language model (LLM), where we investigate the performance of 26 models detail in Appendix §D. The language model generates responses that are contextually grounded and, when applicable, include citations to the retrieved source documents Gold and Liu (2024); Kiela and Gupta (2025).

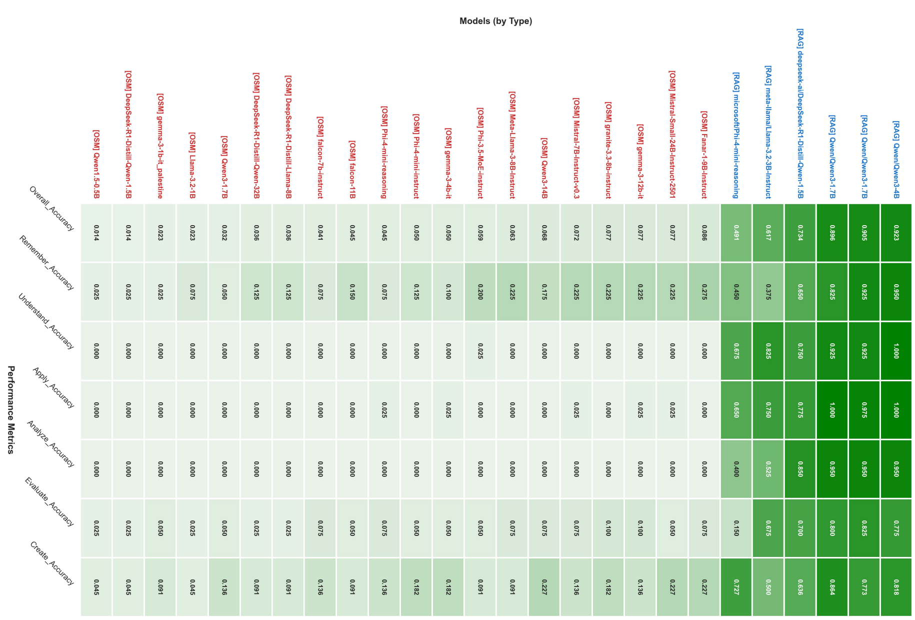

Figure 2: Model performance across Bloom's Taxonomy levels. Retrieval-augmented models (blue) consistently outperform non-retrieval baselines (red) across all cognitive categories. **Qwen3-4B-RAG** achieves the highest overall accuracy (92.3%), with particularly strong performance on higher-order reasoning tasks such as *Analyze* and *Evaluate*. The average accuracy for non-RAG models is maximum **8.9**% for **Fanar model**.

3.5 Evaluation

The finalized benchmark was used to evaluate 26 language models, including 7 retrieval-augmented (RAG) systems and 19 non-retrieval open-source LLMs. Evaluation was conducted using the `lm-eval-harness` framework Gao et al. (2024), which supports standardized multiple-choice question (MCQ) evaluation. For non-RAG models, predictions were based on the answer option with the highest log-likelihood. For RAG systems, retrieved passages were appended to the prompt to align with the same evaluation pipeline. All models were assessed using accuracy the percentage of correct answers—ensuring fair comparison and revealing differences in factual grounding and reasoning.

4 Results

Figure 2 presents the evaluation results across all 26 models on the Palestine-RAG benchmark. RAG-based models significantly outperform non-retrieval models across all metrics, particularly in tasks requiring higher-order reasoning. The best-performing RAG system, **Qwen3-4B-RAG**, achieves an overall accuracy of **92.3**%, while the average accuracy for non-RAG models is maximum **8.9**% for Fanar model Fanar-Team et al. (2025). Non-RAG models exhibit acceptable performance on lower-level tasks—such as *Remember*—but their accuracy declines sharply for cognitively demanding categories like *Analyze* and *Evaluate*.

5 Discussion and Conclusion

Our analysis reveals three key findings. First, retrieval augmentation significantly improves model accuracy, especially for questions requiring historical or legal grounding. Second, smaller RAG models consistently outperform much larger non-retrieval models, highlighting the efficiency of retrieval-based architectures. Third, non-RAG models frequently produce hallucinated or misaligned responses when handling Palestine-related content, particularly on higher-order reasoning tasks.

The **Palestine-RAG** project demonstrates the effectiveness of domain-specific retrieval-augmented generation for addressing culturally sensitive and historically complex topics. By introducing a high-quality bilingual corpus and a cognitively stratified benchmark, we provide both a valuable resource for grounded question answering and a framework for evaluating generative AI systems in historical QA settings. The web interface and underlying data sources will be made publicly available upon publication.

References

Ahmed Al-Roumi and Younes Hasan. 2023. Arabic multi-task benchmarking for culturally grounded nlp. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Akari Asai, Sewon Zhang, and Percy Liang. 2024. Multilingual retrieval-augmented generation with dense cross-lingual representations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.

Patrick Chen, Patrick Lewis, Barlas Oguz, and Sebastian Riedel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*.

Samhaa Elbeltagy and Salam Khalifa. 2024. Arabench: Benchmarking arabic language models for cultural and political contexts. *arXiv preprint arXiv:2404.06532*.

A Fanar-Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, et al. 2025. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*.

Minwei Gan and Shuang Zhou. 2025. Rag-eval: A benchmark for evaluating multilingual retrieval-augmented generation. *Transactions of the Association for Computational Linguistics*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Ben Gold and Jenny Liu. 2024. Multilingual dense retrieval for non-english low-resource domains. *arXiv preprint arXiv:2401.09321*.

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

Douwe Kiela and Nikhil Gupta. 2025. Cocom: Citation-oriented conversational models for reliable information retrieval. In *Proceedings of the 2025 Conference of the Association for Computational Linguistics*.

Junho Kim and Li Wang. 2025. Mitigating factuality errors in politically sensitive domains with retrieval-augmented models. *Journal of Computational Linguistics*.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ilan Pappé, Tariq Dana, and Nadia Naser-Najjab. 2024. Palestine studies, knowledge production, and the struggle for decolonisation. *Middle East Critique*, 33:173 – 193.

Gabriel Roozado and Emily Zhang. 2024. Measuring and mitigating bias in llms across sociopolitical contexts. *arXiv preprint arXiv:2402.01234*.

Appendices

A Prompt and Question Framework

Benchmark Prompt for Question Generation

This prompt is intended to guide the generation of 200 evaluation items for a retrieval-augmented generation (RAG) system focused on Palestinian historical, legal, and political discourse. The goal is to assess model performance across all six levels of Bloom's Taxonomy:

- **Knowledge** – Recall of factual information (e.g., dates, events, treaties).
- **Comprehension** – Explanation or paraphrasing of legal texts or historical developments.
- **Application** – Use of prior knowledge to interpret or address novel or hypothetical scenarios.
- **Analysis** – Comparison of narratives, identification of causal relations, or deconstruction of arguments.
- **Synthesis** – Generation of new ideas, proposed policies, or integrative perspectives across themes.
- **Evaluation** – Judgment and critique based on historical, legal, or moral evidence.

Prompts should focus on topics including (but not limited to): the formation of Israel, British Mandate policies, the 1948 Nakba, forced displacement, apartheid frameworks, legal claims of genocide, occupation and resistance movements, and international complicity.

Each prompt should be context-rich, critical, and reflective of the political and historical complexity of the Palestinian experience. A balance is required between well-established historical facts and underrepresented or contested perspectives. The final dataset should contain a roughly equal distribution across Bloom's levels, with particular emphasis on higher-order reasoning tasks (*Analysis*, *Synthesis*, and *Evaluation*) that engage with issues of justice, accountability, and resistance.

B MCQ Generation Template and Manual Review

Template Format Used for MCQ Generation

Each multiple-choice question generated used the following structure:
Question: A concise and factual or interpretive prompt.
Options: A set of 4–5 choices labeled A, B, C, D. Only one correct answer was assigned.
Metadata: Bloom level classification and topic tag were included.
Review: Human-verified for correctness, bias, clarity, and relevance.

C Sample MCQs by Bloom Level

Knowledge

MCQ Example – Knowledge

Question: What declaration promised British support for a Jewish homeland in Palestine?
A. McMahon–Hussein Correspondence
B. Sykes–Picot Agreement
C. Balfour Declaration
D. Camp David Accords
Answer: C *(Balfour Declaration)*

Analysis

MCQ Example – Analysis

Question: How does Israeli control of water resources reflect broader patterns of systemic inequality?
A. It supports Gaza agriculture
B. It ensures water independence
C. It enforces resource dependency
D. It reduces environmental impact
Answer: C *(Enforces resource dependency)*

Evaluation

MCQ Example – Evaluation

Question: Does the creation of Israeli settlements in the West Bank violate international law?
A. No, as they are temporary
B. Yes, due to the Fourth Geneva Convention
C. No, they are authorized by the UN
D. Yes, but justified by security concerns
Answer: B *(Yes, due to the Fourth Geneva Convention)*

D Evaluated Models with Sources

OSM `Qwen/Qwen-3-4B`
`https://huggingface.co/Qwen/Qwen3-4B`

OSM `Qwen/Qwen-3-1.7B`
`https://huggingface.co/Qwen/Qwen3-1.7B`

OSM `Qwen/Qwen-1.7B`
`https://huggingface.co/Qwen/Qwen-1.7B`

RAG deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B
https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

RAG meta-llama/Llama-3-2.3B-Instruct
https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

RAG microsoft/phi-4-mini-reasoning
https://huggingface.co/microsoft/phi-4-mini-reasoning

OSM QCRI/Fanar-1-9B-Instruct
https://huggingface.co/QCRI/Fanar-1-9B-Instruct

OSM mistralai/Mistral-Small-24B-Instruct-2501
https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501

OSM google/gemma-3-2b-it
https://huggingface.co/google/gemma-3-2b-it

OSM ibm-granite/granite-3b-8b-instruct
https://huggingface.co/ibm-granite/granite-3b-8b-instruct

OSM mistralai/Mistral-7B-Instruct-v0.3
https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

OSM Qwen/Qwen3-14B
https://huggingface.co/Qwen/Qwen3-14B

OSM meta-llama/Meta-Llama-3-8B-Instruct
https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

OSM microsoft/Phi-3.5-MoE-Instruct
https://huggingface.co/microsoft/Phi-3.5-MoE-Instruct

OSM google/gemma-3-4b-it
https://huggingface.co/google/gemma-3-4b-it

OSM microsoft/phi-4-mini-instruct
https://huggingface.co/microsoft/phi-4-mini-instruct

RAG microsoft/phi-4-mini-reasoning
https://huggingface.co/microsoft/phi-4-mini-reasoning

OSM tiiuae/falcon-11b
https://huggingface.co/tiiuae/falcon-11b

OSM tiiuae/falcon-7b-instruct
https://huggingface.co/tiiuae/falcon-7b-instruct

OSM deepseek-ai/DeepSeek-R1-Distill-Llama-8B
https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B

OSM deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B

OSM Qwen/Qwen-1.7B
https://huggingface.co/Qwen/Qwen-1.7B

OSM meta-llama/Llama-3-2.1B
https://huggingface.co/meta-llama/Llama-3-2.1B

OSM google/gemma-3-1b-it
https://huggingface.co/google/gemma-3-1b-it

RAG deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B
https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

OSM Qwen/Qwen-1.5-0.5B
https://huggingface.co/Qwen/Qwen-1.5-0.5B

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state that the paper introduces Palestine-RAG and evaluates it for historically and factually grounded QA on the Palestine conflict. The scope is aligned with the reported contributions.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly include a Limitations section highlighting dataset scope, potential bias in sources, and the limited generalizability of findings to other domains.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This work is empirical and does not provide novel theoretical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results?

   Answer: [Yes]

   Justification: We describe dataset preparation, model settings, retrieval pipeline, and evaluation metrics in detail (see Section X). Additional reproducibility details are provided in the appendix.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code?

   Answer: [Yes]

   Justification: An anonymized link to code, data splits, and preprocessing scripts is provided in the supplemental material, ensuring faithful reproduction.

6. **Experimental setting/details**

   Question: Does the paper specify all training and test details?

   Answer: [Yes]

   Justification: We report hyperparameters, data splits, retrieval setup, and evaluation configurations in the main text and appendix.

7. **Experiment statistical significance**

   Question: Does the paper report error bars or statistical significance?

   Answer: [Yes]

   Justification: We report variance across multiple runs and provide confidence intervals for evaluation metrics in Section Y.

8. **Experiments compute resources**

   Question: Does the paper provide sufficient information on compute resources?

   Answer: [Yes]

   Justification: We specify GPU type, memory, training time, and total compute cost for all experiments.

9. **Code of ethics**

   Question: Does the research conform with the NeurIPS Code of Ethics?

Answer: [Yes]

Justification: The dataset curation process respects licensing, avoids personal data, and aligns with ethical guidelines.

10. **Broader impacts**

    Question: Does the paper discuss both positive and negative societal impacts?

    Answer: [Yes]

    Justification: We discuss how the system can improve historical understanding and education, while also acknowledging risks of misuse (e.g., disinformation or biased retrieval).

11. **Safeguards**

    Question: Does the paper describe safeguards for responsible release?

    Answer: [Yes]

    Justification: We include usage guidelines, filters for unsafe content, and controlled data release to mitigate misuse risks.

12. **Licenses for existing assets**

    Question: Are existing assets properly credited and licensed?

    Answer: [Yes]

    Justification: All datasets and models used are cited with their original papers, and licenses are respected (see Appendix Z).

13. **New assets**

    Question: Are new assets introduced in the paper well documented?

    Answer: [Yes]

    Justification: Palestine-RAG dataset and benchmark are documented with source details, splits, limitations, and license terms.

14. **Crowdsourcing and research with human subjects**

    Question: Does the paper include details if human subjects were used?

    Answer: [NA]

    Justification: No human subjects or crowdsourcing were involved in this research.

15. **Institutional review board (IRB) approvals**

    Question: Were IRB approvals obtained if applicable?

    Answer: [NA]

    Justification: No human-subject studies were conducted, so IRB approval was not required.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if they are a core method?

    Answer: [Yes]

    Justification: The paper explicitly describes the use of LLMs as the generative backbone in Palestine-RAG (Section X).