ArabKT: A Comprehensive Arab Knowledge Evaluation Suite for Large Language Models

Anonymous ACL submission

Abstract

The evaluation of large language models (LLMs) is crucial for understanding their capabilities, yet current methods rely heavily on manually created benchmarks that cover a small fraction of specific knowledge domains. To address this gap, we propose an automated approach that generates evaluation questions for each concept within a domain to construct a comprehensive benchmark. We demonstrate this approach through a preliminary implementation in the Arab world. First, we construct ArabKT, an Arab-world Knowledge Taxonomy derived from Wikipedia, which organizes 140,433 categories and 1.67M articles into a 15-layer tree structure. Subsequently, we developed an automated pipeline to generate 6M QAs encompassing all articles within ArabKT. Experiments reveal that: (1) LLMs demonstrate limitations in handling sensitive and region-specific topics (e.g., culture and religion), indicating a need for improved alignment and native feedback; (2) increasing model size shows no significant improvement in knowledge intensive and knowledge integration areas (e.g., cross-regional topic in Middle East). These findings provide statistical evidence and actionable guidance for improving LLMs in underexplored areas.

1 Introduction

The evaluation of large language models (LLMs) has become increasingly important (Hendrycks et al., 2021; Koto et al., 2024; Wang et al., 2024; Lin et al., 2021). Current evaluation methods mainly rely on manually created benchmarks using real-world data. For example, MMLU contains $12,554$ questions across 57 categories (Hendrycks et al., 2021). However, this represents only a tiny fraction of general knowledge. Wikipedia, in comparison, contains 1.8 million categories and 1.3 billion articles (Vrandečić and Krötzsch, 2014). This huge disparity makes it hard to fully assess models

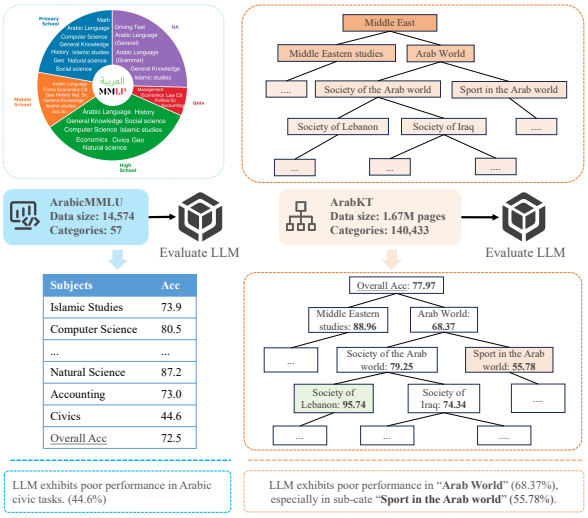


Figure 1: Overview of ArabicMMLU and ArabKT evaluation benchmarks for assessing LLMs' Arab knowledge. The numbers and accuracies within ArabicMMLU is from (Koto et al., 2024).

especially in specific knowledge domains. Limited evaluation data often misses important long-tail knowledge (Üstün et al., 2024; Kim et al., 2008) and specialized topics with which LLMs struggle.

To deal with this problem, there is a growing need to shift towards automated generation of evaluation datasets. This approach presents two main challenges: generating high-quality evaluation data and ensuring comprehensive coverage across topics. Recent advances in LLMs have rendered the automated generation of high-quality data increasingly feasible (Yang et al., 2024b; Zheng et al., 2024). For comprehensive coverage, encyclopedias or called Body of Knowledge (BOK) in professional contexts (contributors, 2024) can serve as valuable references containing wide range of knowledge in specified domains. Representative examples include SWEBOK (Bourque and Fairley, 2004) for software engineering and projects like YAGO (Suchanek et al., 2007) and WikiData (Vrandečić and Krötzsch, 2014).

In this work, we focus on knowledge about the Arab world, an area with rich linguistic and cultural diversity (Koto et al., 2024), but not extensively explored by current LLMs. In addition, we choose this domain as a representative case for knowledge in low-resource languages. Our ultimate objective is to develop an Arab BOK through various of corpus supported by substantial domain expertise. But at first step, we aim to build a prototype first by leveraging vast knowledge in Wikipedia and constructing an Arab Knowledge Taxonomy (ArabKT). To construct it, we developed a systematic approach that proceeds in three main steps: (1) leveraging Wikidata's category system to extract a comprehensive concept network related to the Arab world as the foundation for our taxonomy; (2) developing an agent-based process to rectify and enhance category definitions; (3) eliminating loop dependencies in the category network to transform the complex network into a more manageable tree structure. As a result, we build an ArabKT with 15 layers, containing $140,433$ categories and 1.67 million articles. This taxonomy covers around 77% knowledge of the Arab pre-training corpus and 84% of Arab benchmarks. It is noteworthy that the proposed framework does not incorporate specialized designs for Arab knowledge. This decision aligns with our primary objective of developing a domain-agnostic framework that can be readily adapted to various fields, exploring a possible way for knowledge in different low-resource languages.

After the construction of ArabKT, we developed an automated evaluation process with human verification, to evaluate how well LLMs understand Arab world knowledge. Specifically, language models are used to create test questions based on key information extracted from Wikipedia articles within ArabKT. To ensure a comprehensive and rigorous evaluation, we adopted a multi-perspective approach to question generation and applied automated LLMs validation for each question after generation. This process yielded 6 million question-answer pairs for evaluating various language models. As shown in Fig. 1, evaluation results on ArabKT demonstrate varying levels of accuracy across different topics. While LLMs show strengths in topics related to "Society of Lebanon", they exhibit weaknesses in "Arab World" content, particularly regarding the sub-category "Sport in the Arab World". Our evaluation reveals that LLMs consistently struggle with religiously sensitive topics and knowledge-based cognitive conflicts, indicating the need for alignment data incorporating native cultural feedback. Furthermore, although larger models demonstrate superior performance in handling complex knowledge, they show no particular advantages in knowledge-intensive and knowledge integration domains, such as regional academic and cross-cultural topics. This suggests that while model size is crucial for handling straightforward domains, data quality and coverage show higher priority in expertise-driven and interdisciplinary areas. These findings provide a foundation for comprehending model capabilities and guide future improvements.

The contributions of this work are summarized as follows: First, we introduce ArabKT, a comprehensive Arab Knowledge Taxonomy derived from Wikipedia and Wikidata. Second, we develop an automated process to generate large-scale evaluation data, producing 6 million question-answer pairs to assess LLMs' understanding of Arab world knowledge. Third, extensive experiments reveal patterns about the weakness of LLMs: further alignment and native feedback is important in sensitive and cognitive conflict areas, and data acquisition show higher priority than model scale especially on niche expertise and interdisciplinary topics. These findings provide valuable insights for understanding and improving LLMs' capabilities in specific knowledge domains.

2 Building Knowledge Taxonomy

2.1 Overview of the Workflow

Fig. 2 illustrates our workflow for building the ArabKT and evaluations based on it. Using WikiData's API (Vrandečić and Krötzsch, 2014), we collected all articles and categories related to the Arab world. We then applied a combination of rule-based filtering and LLM-based semantic understanding to remove non-Arab content and articles with content lacking valid information. This initial process resulted in a directed graph of knowledge from Arab world.

The conversion of this graph into a practical, manageable hierarchical structure presented two principal obstacles. The first challenge was missing or incorrect category definitions. Motivated by self-improve frameworks (Dhuliawala et al., 2023; Zhang et al., 2023; Weng et al., 2022), we developed a pair of agentic models that work together - one for generating definitions and another for critiquing them, allowing iterative improvements.

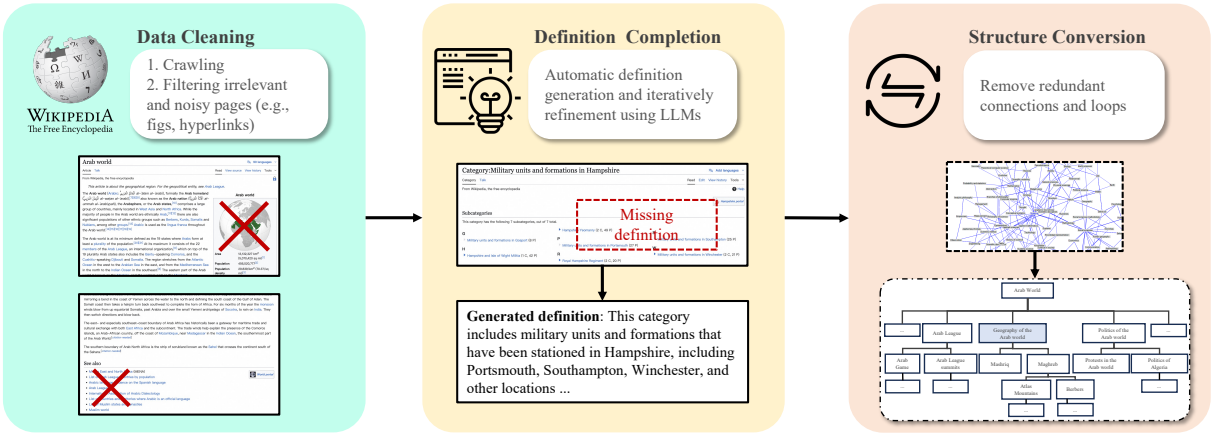


Figure 2: Construction workflow of ArabKT (Arab Knowledge Taxonomy). The process consists of three main stages: (1) Data cleaning: removing noise elements (figures, hyperlinks, references) from Wikipedia pages; (2) Definition completion: generating missing definitions for Wikipedia categories using an agentic model; and (3) Structure conversion: transforming the Wikipedia category network into a tree structure through loop elimination.

The second challenge was redundancy in the graph structure, particularly cycles. We solved this by combining depth-first search algorithms with LLM assistance to remove redundant connections, transforming the graph into a proper tree structure.

Finally, we used the ArabKT to guide question generation for evaluating existing LLMs. This evaluation process produced what we call an "Accuracy Tree", which provides a detailed analysis of different language models' capabilities across various levels and categories of knowledge.

2.2 Data Crawling and Cleaning

Based on the API provided by WikiData, we started by using "Middle East" as the entry point for queries, recursively searching for unique subcategories and their associated articles. To ensure comprehensive coverage, we retained as many categories as possible, ultimately collecting 5.4 million pages (including both categories and articles). Details of the content are in Appx. A.2.

Rule-based Data Cleaning. Based on the structural characteristics of Wikipedia article pages, we developed a set of rules to eliminate content lacking valid textual information. This process involved removing various non-essential elements, including hidden content, floating images, tables, text boxes, prompts, footer boxes, and multiple types of citations. Additionally, we targeted textual content by excluding long strings of characters such as coordinates, and mathematical formulas. Meanwhile, we remove all superscript symbols in the main context. After cleaning 5.4 million pages in total, we removed entries with empty content, resulting in a final collection of 3.7 million pages.

Heuristic-based Data Cleaning. Furthermore, we sampled 1,000 pages to identify typical characteristics of unreasonable pages. We found the following common issues: 1) Pages with specific titles, such as those containing "File", "Template", and similar terms. These pages typically lack effective textual descriptions, prompting us to filter them out whenever matching. 2) Continuous short texts, such as lists of a particular topic. These pages also lack sufficient descriptions and pose parsing challenges. We record the length of each text segment and filtered out pages where continuous short text comprised more than 50% of the content. 3) Webpage redirects. For these pages, we copied the content from the target page while retaining the original title and added redirect information in the metadata. By implementing these methods, we removed around 0.2 million pages from our dataset.

Semantic Filtering. We also implemented a two-stage filtering combining heuristic rules and LLM. First, with the assist of native experts, we extract a comprehensive keyword list comprising 448 terms across six domains: geographic regions, country names, important cities and landmarks, ethnic cultures, languages, and religions. For other low-resource corpus, we can use TF-IDF to extract candidate keywords and use LLM to filter. Yet, it will be more effective with the aid of local experts. Pages with titles containing these keywords were automatically retained. For the remaining pages, we employed an LLM to evaluate their relevance to Arab knowledge, which has a 95% consistency compared with manual annotations in validation.

Detailed methodology and evaluation metrics are provided in the Appx. A.3.

Multilingualism. Multilingualism is common in Arab knowledge and information. The same piece of information often exists in different languages. Some specific knowledge is only available in certain languages. In developing ArabKT, we selected English Wikipedia as our initial corpus due to its largest collection of Arab-related articles and widest coverage of concepts and topics. Future work will incorporate knowledge from other language Wiki articles into ArabKT.

2.3 Definition Completion

Our analysis of Wikipedia categories revealed that only 17.3% contain valid definitions. They either provide overly brief descriptions, containing irrelevant content, or lack definitions entirely. However, definitions play a crucial role for the precision and downstream application of the constructed ArabKT. On the one hand, due to the vast and complex knowledge in Arab world, without a systematic and precise definition for each concept (equivalent to ontology in BOK (Burgueño et al., 2019; Luo et al., 2021)), **misunderstandings and misclassification of certain concepts can easily occur**. Take "Madrasa" (refers to Islamic schools in Arabic) as an example - without proper definition, it might be broadly miscategorized as just a "religious school," leading to incorrect equivalencies between institutions like the 11th century Nizamiyya Madrasa in Baghdad (an academic center teaching philosophy and mathematics) and 18th century Quranic recitation schools in rural Morocco (focusing on basic religious instruction).

More importantly, for downstream application of the ArabKT, definitions will serve crucial functions: (1) **Knowledge Integration**: Definitions facilitate the integration of new knowledge (particularly Arabic-language content from external sources) by providing clear criteria for determining appropriate hierarchical placement within the taxonomy. (2) **Enhanced Retrieval Capabilities**: Definitions enable sophisticated semantic similarity computations for identifying related nodes from vast knowledge concepts and improve knowledge navigation (users can preview node definitions before deciding whether to explore subtrees). (3) **Data Synthesis**: Rewriting knowledge to inject knowledge into LLMs is a common practice (Maini et al., 2024; Yang et al., 2024b). Clear definitions ensure consistent interpretation of concepts across models, preventing biases in different models' interpretations of concepts (especially domain-specific concepts), thereby improving the quality of synthesized data.

To address this issue, we implemented a pair of agentic models for iterative definition completion. For generation process, it primarily relies on Wikipedia's own content, with web searches serving as a supplementary source when the initial generation fails or when the critic model indicates insufficient information. For the critic process, it evaluates the generated definitions using five key dimensions: Accuracy, Clarity, Non-Circularity, Scope, and Conciseness. It helps determine the reasonableness of definitions and identifies specific areas requiring improvement. The feedback is then input to the next round generation. Through an iterative process involving five rounds of generation and evaluation for each category, we successfully created $120,000$ definitions. The quality of these definitions is reflected in their average score of $4.83/5$. Full details of our method and evaluation are provided in the Appx. A.4.

2.4 Category Rectification.

Loop Removal. During our implementation, we encountered frequent loops in the knowledge paths. To address this issue, we employed depth-first search (Tarjan, 1971) to detect loops in the paths. When a loop was found, we identified cases where a sub-category appeared in previous super-categories. In these cases, we cut and removed the redundant paths to eliminate the loops. This process transformed the crawled structure into a directed acyclic graph, where each path follows a clear hierarchical order without any circular references.

Tree Conversion. We aimed to simplify nodes that had multiple super-categories to create a more human-comprehensible structure. Our simplification process involved three steps: First, for each node with multiple super-categories, we removed redundant connections where one super-category was already a parent of another super-category. For example, $\mathcal{C}$ is denoted as the super-categories of one node, we remove the $c \in \mathcal{C}$ when c is also the parent of another $c' \in \mathcal{C}$. Next, among the remaining super-categories $\hat{\mathcal{C}}$, we identified candidate categories at the deepest level using depth-first principle. Finally, when multiple candidates existed at the same level, we used an LLM to select the most appropriate one, which we termed as the golden super-category.

We maintained the connection between the node and its golden super-category, along with all subsequent connections. For other super-categories, while we removed their direct connections, we preserved copies of these relationships as hyperlink-like references. This approach maintained the tree structure while preserving important cross-references in the knowledge hierarchy.

3 Arab Knowledge Taxonomy

Following the approach in the previous section, we constructed a ArabKT for the Arab world. To evaluate it, we analyzed it from three key dimensions: statistic, coverage, and accuracy. First, in Sec. 3.1, we assessed the scale and the distribution to understand its overall structure and composition. Next, in Sec. 3.2, we compared its coverage with publicly available Arabic training- and test-sets to determine its breadth and representativeness. Finally, in Sec. 3.3, we evaluated the accuracy of the generated definitions by comparing them with expert-annotated results.

3.1 Statistics

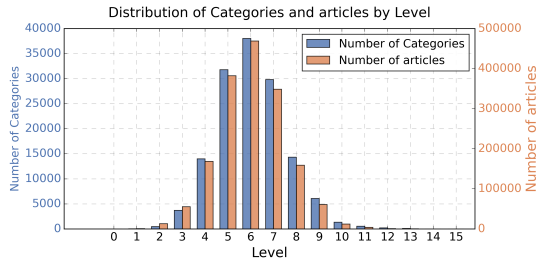


Figure 3: Hierarchical distribution of categories and articles in ArabKT taxonomy.

The ArabKT contains a hierarchical structure spanning 15 layers and encompassing 140,433 distinct categories. These categories are linked to a substantial collection of 1.67 million articles. Fig. 3 presents a detailed breakdown of how categories and articles are distributed across the hierarchical layers, alongside the distribution pattern of articles within individual categories. Notably, we observed that the middle layers (4 through 8) house 87% of all articles, establishing these layers as the ArabKT's most information-rich region.

3.2 Coverage

In this section, we evaluate the coverage of ArabKT by assessing how well ArabKT encompasses the knowledge contained in common Arabic datasets.

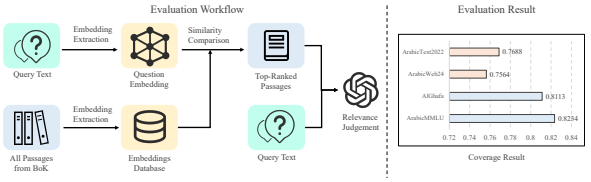


Figure 4: Coverage evaluation workflow and result statistics of the coverage of ArabKT.

Here, semantic coverage refers to that the knowledge points in ArabKT can effectively represent and explain the concepts, facts, and relationships present in the sample form Arabic datasets. Specifically, we choose two widely-used Arabic cultural evaluation datasets (AlGhafa (Almazrouei et al., 2023) and ArabicMMLU and two Arabic pre-training datasets (ArabicText2022 (BAAI et al., 2022) and ArabicWeb24 (Farhat et al., 2024)).

The evaluation workflow is shown in Fig. 4. We adopt a RAG-inspired approach (Lewis et al., 2020) for efficient retrieval and coverage assessment. More details are in App. A.8. Through this process, we can assign a $0/1$ for each chunk (paragraph) in corpus or question in benchmarks. Then the coverage score is defined as:

$$C(D) = \frac{|\{d \in D | \exists k \in K : I(d,k) = 1\}|}{|D|} \quad (1)$$

where $|D|$ denotes the total number of dataset D, and $I(d,k)$ is an indicator function for $1/0$. The results are shown in Fig. 4. The ArabKT achieves coverage rates of 76.88% and 75.64% on training corpus, while achieving 82.34% and 81.13% on evaluation datasets.

Conversely, we can also evaluate how many knowledge points are covered by the current benchmarks.

$$C_{rev}(D) = \frac{|\{c \in C_{cat} | \exists d \in D : I(d,c) = 1\}|}{|C_{cat}|} \quad (2)$$

where C_{cat} represents all category nodes in our ArabKT, and $I(d,c)$ indicates whether sample d covers category node c or its descendants. The result reveals that ArabicMMLU only covers 15.51% of the knowledge categories in our ArabKT. It indicates that ArabKT contains large number of new knowledge points to evaluate LLMs.

3.3 Precision

We evaluate the precision of the generated definition by comparing the performance between GPT-4o and human annotators following the head-to-head evaluation in Alpaca-Eval (Li et al., 2023a).

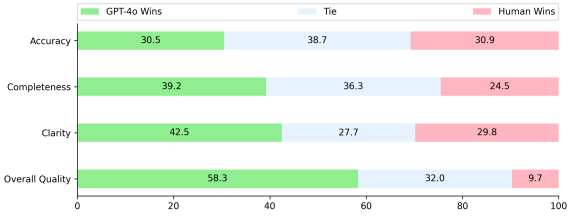


Figure 5: Comparative analysis of definition quality: LLM-generated vs. human-expert annotations. The proposed agentic approach achieves slightly better results than expert-crafted definitions.

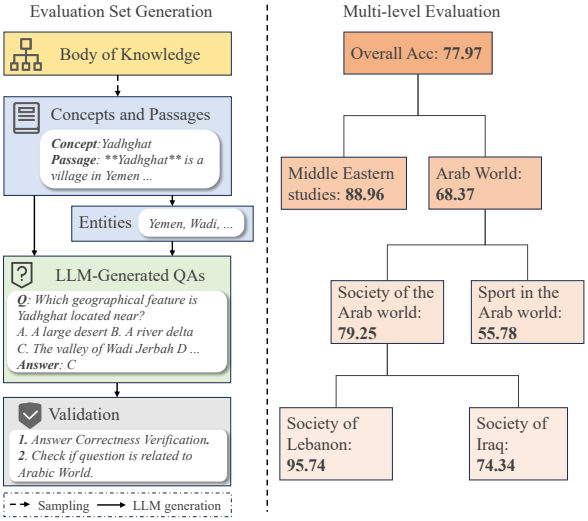


Figure 6: (**Left**) Workflow diagram of question generation leveraging ArabKT articles; (**Right**) Hierarchical visualization of accuracy metrics across taxonomic levels and categories.

The precision is assessed across four dimensions: accuracy, completeness, clarity, and overall quality. A more detailed setting is provided in App. A.4.

The evaluation results are shown in Fig 5, Our evaluation results demonstrate that GPT-4o performs comparably or superiorly to human annotators across all assessed dimensions. The model achieves near-identical accuracy scores with humans (30.5% vs. 30.9%), while showing notable advantages in completeness (39.2% vs. 24.5%) and clarity (42.5% vs. 29.8%). Most significantly, in terms of overall quality, GPT-4o substantially outperforms human annotators with 58.3% of its definitions being preferred, compared to 9.7% for human-written definitions. These findings suggest that GPT-4o can generate definitions that not only match but often exceed human-expert quality.

4 Evaluation of LLMs based on Arabic Knowledge Taxonomy

In this section, we introduce one of the prominent applications of our ArabKT, i.e., evaluating LLMs' understanding of Arab-related knowledge. We aim to answer two research questions within this section: (1) **R1: How well do current prevalent LLMs comprehend knowledge related to the Arab world?** (2) **R2: How do models of different sizes vary in their understanding of Arab knowledge?** Specifically, we first introduce the overall evaluation workflow and experiment settings (§4.1). Then we discuss the evaluation and analysis results for R1 4.2 and R2 4.3.

4.1 Evaluation Workflow

Using all articles from ArabKT, the questions are automatically generated as shown in Fig 6. Following the construction of recent knowledge-based questions and benchmarks Yang et al. (2024b); Wang et al. (2023), we adapt their prompt and process to generate multiple-choice questions. Two types of questions are considered for thoughtful coverage of given knowledge points. 1) Multiple-choice questions are directly summarized by LLM. This type of questions will consider the whole content and more deeper. 2) Entities are first extract from the articles. The questions are then generated to discuss these two selected entities. This type of questions are able to contain easily overlooked content. For each type, three questions are generated.

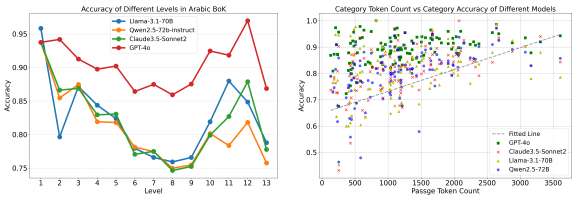
Additionally, we also validate the generated questions to avoid knowledge hallucination issue in LLMs (Huang et al., 2025). This process involves two steps: First, we check the correctness of the generated answers using the approach in (Wang et al., 2023). Specifically, we prompt the LLM to answer the questions based on the provided passages, checking if the model's predicted answers match the generated answers. Secondly, we use the LLM to determine whether the questions are related to the Arab world, filtering out irrelevant questions. After validation, 6.21M evaluation questions are gathered. Furthermore, we employed manual assessment of the question quality and results demonstrate high quality across multiple dimensions (e.g., fluency and answerability). The prompts, cost, and evaluation details are available in Appx. A.5.

Evaluation setting. We use the same prompt from (OpenAI, 2024) that first generates a chain of thoughts and then outputs the final choice. The temperature is set 0 during inference to facilitate

reproducibility of the results.

Evaluation models. For **R1**, we select two prevalent proprietary LLMs (GPT-4o (Hurst et al., 2024) and Claude-3.5-Sonnet 2 (Anthropic, 2023)) and two open-sourced LLMs (Llama-3.1-70B-Instruct (Dubey et al., 2024) and Qwen-2.5-72B-Instruct (Yang et al., 2024a)) for comparison. For **R2**, to compare the extent of knowledge acquisition across models of varying sizes, we selected the Qwen2.5 model series (Yang et al., 2024a), which have a wide range of different sizes. Finally, we choice all available sizes ranging from 3B to 72B.

4.2 LLMs performs poorly on areas with sensitive and region-specific topics



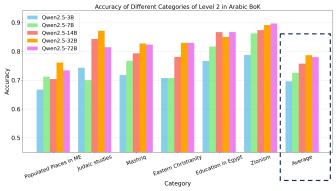
(a) (b)

Figure 7: (a) Accuracies within different levels of ArabKT on four prevalent LLMs. (b) Relationship between category accuracy and average token length of the passages within the corresponding category.

The overall accuracy scores for different models are: GPT-4, 85.7%; Llama-3.1-70B, 78.2%; Qwen2.5-72B-Instruct, 78.0%; and Claude3.5-Sonnet2, 77.3%. The accuracy on different LLMs show strong consistency across different categories. Fig. 7a shows the averaged accuracy across the layers. For example, the correlation coefficient between Qwen2.5-72B-Instruct and Claude3.5-Sonnet2 is 0.7988. As widely recognized, we observe that low accuracy occurs when the content of category is less frequent in the training corpus (Fig. 7b with more details in Appx. A.5). However, beyond frequency-based patterns, we also find LLMs' deficiencies have high correlation with special content (with over 30 related articles vs. 10 on average). **The model demonstrates limited performance when handling sensitive and region-specific topics.** For example, LLMs present markedly lower accuracy on *religiously sensitive topics demonstrate*, such as *LGBTQ in the Middle East* (0.69 average accuracy across four models) and *Political parties* (0.59). It is potentially attributable to overly conservative safety protocols implemented in LLMs' alignment tuning stage. Another distinctive type is *knowledge domains exhibiting cognitive conflicts*, including *Economy of Oman* (0.63) and *Education in Saudi Arabia* (0.48), etc. It is likely stemming from discrepancies between Arabic cultural contexts and English-dominated knowledge bases. Due to the unique regional culture of the Arab world, these cognitive conflicts are manifested in many concepts. Through ArabKT, we can pinpoint these specific knowledge points.

These observations suggest two critical strategic directions for enhancing the performance of existing LLMs in handling Arab world knowledge: (1) For sensitive topics, future model development should prioritize the incorporation of more comprehensive alignment data and systematic feedback from native cultural experts during the alignment phase (e.g., through Direct Preference Optimization (DPO) (Rafailov et al., 2023) and Reinforcement Learning from Human Feedback (RLHF) (Bai et al., 2022)). This approach would help ensure that models better reflect authentic regional perspectives and cultural sensitivities, reducing potential biases and misinterpretations that may arise from Western-centric training data. (2) For domains where cognitive conflicts exist between different cultural frameworks, increasing the representation of Arab world-specific knowledge is essential. This involves not only expanding the quantity of relevant content but also ensuring its quality and authenticity.

4.3 Large models have no superiority on knowledge intensive and knowledge integration areas



Distinguished Topics	Comparable Topics
Middle Eastern mythology	Persian people
Culture of Qatar	Middle Eastern art
Ancient Near East religion	Middle Eastern women
Arabic-language culture	Sport in the Arab world

Figure 8: (**Left**) Accuracy of Qwen2.5 series models on categories of level 2 in ArabKT. (**Right**) "Distinguished topics" showcase Qwen2.5-72B's superior performance over smaller models, whereas on "Comparable topics" gaps between Qwen2.5-72B and smaller models are minimal.

Similarly, we compared the overall accuracy of Qwen2.5 series models of various sizes (Fig 8). On average, the accuracy increases as the model size increases. Among various categories, larger models have better capability of comprehension and excel

at nuanced topics, e.g. *Middle Eastern mythology* and *Culture of Qatar*. Fig 9 illustrates this distinction with examples. For the geographical question requiring simple mineral-related knowledge recall, both Qwen2.5-7B and Qwen2.5-72B demonstrate comparable performance. However, in tasks involving complex poetry analysis and cultural interpretation, Qwen2.5-7B exhibits significant comprehension deficiencies.

Despite the advantage of larger models in categories demonstrated above, there are also some categories that Qwen2.5-72B-Instruct has no superiority than over smaller counterparts. These categories are mainly focus on areas containing knowledge intensive and knowledge integration topics. Specifically, (1) **knowledge intensive topics:** areas such as *Biblical archaeology* (with accuracy 0.41 on 72B vs 0.38 on 3B) and *Sport in the Arab world* (0.55 on 72B vs 0.57 on 3B), performance remains relatively consistent across model sizes. These topics are full of various information that requires the model to memorize. **This phenomenon can potentially be attributed to underrepresented precise terminology and specialized methodological knowledge in training data.** (2) **Knowledge integration topics:** areas requiring multilingual or cross-regional synthesis, such as *Iranian diaspora* (0.58 on 72B vs 0.56 on 3B) and *Expatriates in Syria* (0.42 on 72B vs 0.42 on 3B). These categories also show minimal variation between model sizes. **This limitation likely stems from the inherent complexity of synthesizing information across different linguistic and cultural contexts.**

These findings suggest two potentially strategic directions for future practical applications and model development:(1) In domains that heavily involve knowledge intensive and knowledge integration, simply increasing model size shows limited effectiveness. Hence, the primary focus should be directed towards enhancing the quality, diversity, and comprehensive coverage of training data. (2) For factually unambiguous domains (such as economy, geography, and quantitative information), especially nuanced topics like cultural, religious and historical, larger models consistently demonstrate clear advantages. Therefore, continued scaling of model size remains the most promising path to performance improvement. This approach is particularly effective when dealing with objective facts and content requiring comprehension, where the model's increased capacity directly translates to better information retention and processing capabilities.

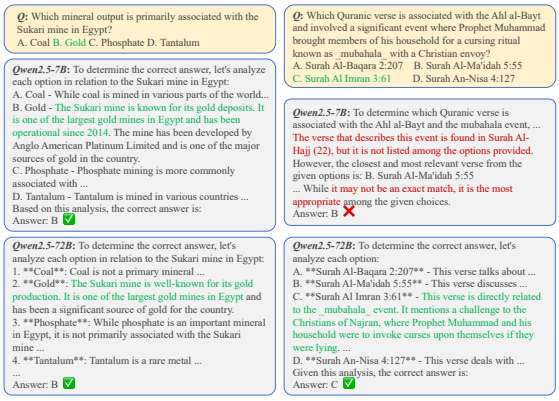


Figure 9: Response analysis of Qwen2.5-7B and Qwen2.5-72B across different question categories. (Left) Both models exhibit similar capabilities in basic knowledge recall tasks (mineral-related geography). (Right) Qwen2.5-7B shows deficiency in complex reasoning tasks (poetry analysis and cultural interpretation).

5 Conclusion and Discussions

In this work, we presented ArabKT, a comprehensive Arabic Knowledge Taxonomy derived from Wikipedia and Wikidata, along with an automated process for generating large-scale evaluation data. Through extensive experiments with 6 million generated questions, we revealed important patterns in how LLMs master the knowledge about the Arab world. Our findings demonstrate that LLMs struggle with sensitive and region-specific topics. Meanwhile, for knowledge intensive and integration topics, scaling the model have no advantage.

Several promising directions remain for future work. First, the taxonomy could be enhanced by incorporating expert knowledge to establish more professional and logical hierarchical relationships. The coverage could also be expanded by including more languages and sources beyond Wikipedia. Additionally, this knowledge taxonomy framework could be applied to various downstream tasks, such as synthetic data generation for model training, knowledge graph construction, and visualization of model reasoning paths. Notably, extending the evaluation scope to Arabic-centric Large Language Models (e.g., ACE (Huang et al., 2024), Jais (Sengupta et al., 2023), and Fanar (Team et al., 2025)) would provide valuable insights for further improving these LLMs.

6 Limitations

Our work is not without limitations. First, the reliance on Wikidata and Wikipedia as foundational resources introduces potential noise and incompleteness. Wikidata's category definitions are missing or inaccurate for approximately 83% of categories, and about 27% of category associations suffer from errors, such as cycles caused by editing mistakes. These issues, although mitigated through our agentic correction process, may still affect the quality and reliability of the Arab Knowledge Taxonomy (ArabKT). Second, the use of large language models (LLMs) for automated question generation and evaluation is subject to inherent limitations. LLMs may produce incorrect or biased questions and answers, and not all such errors can be fully detected or corrected, even with human verification. This underscores the need for continuous refinement of both knowledge sources and evaluation processes to ensure robust and accurate assessments of LLM capabilities.

References

Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. 2024. Arabicaqa: A comprehensive dataset for arabic question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2049–2059.

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, et al. 2023. Alghafa evaluation benchmark for arabic language models. In *Proceedings of ArabicNLP 2023*, pages 244–275.

Manel Aloui, Hasna Chouikhi, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024. 101 billion arabic words dataset. *arXiv preprint arXiv:2405.01590*.

Anthropic. 2023. Claude 3.5 sonnet model announcement. `https://www.anthropic.com/news/claude-3-5-sonnet`. 2025-02-10.

BAAI, AASTMT, BA, and IIAI. 2022. Arabictext-2022: Large-scale arabic text dataset. Beijing Academy of Artificial Intelligence. The world's largest open-source Arabic text dataset for pre-training language models.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Pierre Bourque and RJNICS Fairley. 2004. Swebok. *Nd: IEEE Computer society*.

Loli Burgueño, Federico Ciccozzi, Michalis Famelis, Gerti Kappel, Leen Lambers, Sebastien Mosser, Richard F Paige, Alfonso Pierantonio, Arend Rensink, Rick Salay, et al. 2019. Contents for a model-based software engineering body of knowledge. *Software and systems modeling*, 18:3193–3205.

Wikipedia contributors. 2024. Body of knowledge. Wikipedia, The Free Encyclopedia. Accessed January 2024.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

May Farhat, Said Taghadouini, Oskar Hallström, and Sonja Hajri-Gabouj. 2024. Arabicweb24: Creating a high quality arabic web-only pre-training dataset.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Project Management Institute. 2013. *A Guide to the Project Management Body of Knowledge: PMBOK(R) Guide*, 5th edition. Project Management Institute.

Kiyoung Kim, Kyungho Jeon, Hyuck Han, Shin-gyu Kim, Hyungsoo Jung, and Heon Y Yeom. 2008. Mrbench: A benchmark for mapreduce framework. In *2008 14th IEEE International Conference on Parallel and Distributed Systems*, pages 11–18. IEEE.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, et al. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. *arXiv preprint arXiv:2402.12840*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023a. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Ting Luo, Xiaolong Xue, Yongtao Tan, Yuna Wang, and Yuanxin Zhang. 2021. Exploring a body of knowledge for promoting the sustainable transition to prefabricated construction. *Engineering, Construction and Architectural Management*, 28(9):2637–2666.

Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14044–14072, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2024. simple-evals. `https://https://github.com/openai/simple-evals`. 2025-02-10.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric P. Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *CoRR*, abs/2308.16149.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Robert Tarjan. 1971. Depth-first search and linear graph algorithms. In *12th Annual Symposium on Switching and Automata Theory (swat 1971)*, pages 114–121.

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*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Jinyuan Wang, Junlong Li, and Hai Zhao. 2023. Self-prompted chain-of-thought on large language models for open-domain multi-hop reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2717–2731, Singapore. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2022. Large language models are better reasoners with self-verification. *arXiv preprint arXiv:2212.09561*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, and Tatsunori Hashimoto. 2024b. Synthetic continued pretraining. *Preprint*, arXiv:2409.07431.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. 2023. Cumulative reasoning with large language models. *arXiv preprint arXiv:2308.04371*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Appendix

A.1 Related works

Recent years have witnessed significant efforts in developing comprehensive benchmarks to evaluate large language models' capabilities. MMLU (Hendrycks et al., 2021) introduced a multitask evaluation framework covering 57 diverse subjects, revealing that even the largest models struggle to achieve expert-level performance across different domains. Similarly, specialized benchmarks like TruthfulQA (Lin et al., 2021) and StrategyQA (Geva et al., 2021) focus on specific capabilities such as truthfulness and implicit reasoning. For Arabic language evaluation specifically, a thorough review of Arabic evaluation datasets is shown in Tab. 1 most existing benchmarks (such as ArabicMMLU (Koto et al., 2024) and AlGhafa (Almazrouei et al., 2023)) primarily focus on models' ability to handle formal question-answering in Arabic, with only part of categories within the benchmark that assess the knowledge about the Arab world. The most relevant work to ours is ArabicaQA (Abdallah et al., 2024), which also generates QA pairs from Wikipedia. Yet ArabicaQA utilizes around 10,000 Wikipedia passages (less than 1% of all Arab-related articles), making its evaluation scope limited. Moreover, it lacks categorical information, making it difficult to identify specific knowledge areas where models underperform.

The concept of Body of Knowledge (BOK) has been widely adopted across various professional domains as a comprehensive framework to structure and standardize domain knowledge. Notable examples include the Software Engineering Body of Knowledge (SWEBOK) (Bourque and Fairley, 2004) maintained by IEEE Computer Society, which systematically organizes software engineering knowledge into 15 knowledge areas, and the Project Management Body of Knowledge (PMBOK) (Institute, 2013) by PMI, which has become the global standard in project management. These structured knowledge frameworks typically organize information hierarchically, with high-level categories branching into more specific topics, providing a systematic approach to knowledge representation and assessment. Inspired by these established BOK practices, our work presents a comprehensive Arabic knowledge taxonomy that systematically organizes cultural, linguistic, and domain-specific knowledge, enabling more structured and thorough evaluation of Arabic language models.

A.2 Crawled Articles and Categories

In the ArabicKT knowledge system, there are two main types of nodes: pages (Fig. 11) and categories (Fig. 10). Page nodes contain basic metadata information such as page ID (pageid), title, namespace (ns), as well as links to other language versions (langlinks), associated categories (categories), subcategories, and related pages, establishing hierarchical relationships. Category nodes, on the other hand, primarily store the specific content of pages, language information, page ID, and related pages (related pages), forming a structured knowledge organization system.

A.3 Semantic Filtering

Keywords generation The keywords were extracted through a top-down approach from Wikipedia categories, which were reviewed by

Table 1: Comparison of Different Arabic Evaluation Datasets

Dataset	Evaluation Focus	Data Source	Categories	Language	Size
ArabicMMLU (Koto et al., 2024)	Multi-task capabilities in Arabic across STEM, social science, humanities, language, and 5 other domains	Primary/Secondary school exam questions	Yes	Arabic	14,575
AlGhafa (Almazrouei et al., 2023)	Multi-tasks evaluation like sentiment analysis, reading comprehension, and factual verification	Translated existing datasets	Yes	Arabic	25,088
EXAMs (Hardalov et al., 2020)	Understanding of various subjects (e.g., physics, chemistry, history, geography)	Exam questions	Yes	Arabic	24,143
ArabicaQA (Abdallah et al., 2024)	Reading comprehension and open-domain QA capabilities	Arabic Wiki	No	Arabic	92,796
AceGPT (Huang et al., 2024)	Arabic QA answering	Quora	No	Arabic	8,000
Ours	Comprehensive knowledge evaluation of the Arab world	English Wiki	Yes	English	∼6M

academic experts specialized in Arabic literature and native Arabic speakers. Representative and distinctive keywords (such as "Lebanon") are selected from category titles and their variations (like "Lebanese") are expanded as keywords by these experts. To ensure reliable matching results, we limited our selection to expert-validated keywords rather than including loosely related terms. For the remaining potentially relevant Wikipedia titles not captured by these keywords, we employ large language models in conjunction with article content analysis to determine their relevance to the Arab world, thereby minimizing the risk of overlooking pertinent knowledge. In the future, we plan to expand our vocabulary by mining additional terms from pre-training corpora with expert assistance. The keywords for filtering are shown in Fig. 13.

LLM filtering The study utilized Large Language Models (i.e., GPT-4) to automatically identify and filter out pages unrelated to Arabic culture. The filtering prompt, illustrated in Figure 12, was developed based on a comprehensive definition of Arabic cultural relevance. This definition was synthesized from characteristics identified through a manual analysis of 1,000 randomly sampled Wikipedia articles pertaining to Arabic culture.

To validate the LLM's effectiveness, we conducted a manual analysis of 400 samples and compared them with the LLM's assessments. The results demonstrated high reliability, with an accuracy rate of 94.9% and a recall rate of 99.4%. This combined filtering approach is able to preserve nearly all Arab-related content while maintaining a low false positive rate of approximately 5% non-Arab knowledge points.

A.4 Definition Completion

Iterative refinement of definition We implemented a dual-agent framework for iterative definition refinement, consisting of a generator model for definition creation and a critique model for quality assessment, as shown in Fig. 15. The critique model evaluates generated definitions across five key dimensions: *(accuracy* (assessing the completeness and precision of category descriptions). *clarity* (evaluating the definition's precision and absence of ambiguity), *non-circularity* (ensuring avoidance of self-referential or synonymous explanations), *scope* (verifying appropriate coverage without over- or under-generalization), and *conciseness* (confirming succinct yet comprehensive expression). The specific prompts for both generator and critique models are illustrated in Figure 16 and Figure 17, respectively. Our experimental results, as demonstrated in Figure 14, indicate that this multi-round refinement approach effectively enhances definition quality through iterative improvement.

Human evaluation of definition We recruited twelve MSc students from the Department of Arab Studies in the School of International Studies, specializing in Arabic literature. We randomly selected 200 concepts lacking definitions. We divided the volunteers into two groups evenly. Specifi-

Example of a page in ArabicKT


```
{
  "pageid": 33745012,
  "title": "Category:Works about the Middle East",
  "level": 1,
  "ns": 14,
  "langlinks": {
    ...
  },

  "categories": [
    {
      "title": "Category:Middle East",
      "pageid": 743029
    },
    {
      "title": "Category:Works about regions",
      "pageid": 43478226
    }
  ],
  "subcategories": [
    {
      "title": "Category:Books about the Middle East",
      "pageid": 7960198
    },
    ...
  ],
  "pages": [
    {
      "title": "The Image Book",
      "pageid": 52748682,
      "ns": 0
    },
    ...
  ]
}
```


Figure 10: Example of an category in ArabicKT

Example of an category in ArabicKT


```
{
    "data": {
        "content": "**Élisabeth Terroux** (1759–1822)
        was a Swiss painter active in Russia.\n\n
        Terroux was born in Geneva and trained under
        Jean-François Favre. She became a\npopular
        miniature painter and travelled to Russia
        where she was active for\nCatherine II. Her
        self-portrait was shown at the Paris Exposition
        Universelle\n(1878), \"Les Portraits
        nationaux\", palais du Trocadéro.\n\n
        Terroux died in Geneva.\n\n",
        "language": "en",
        "pageid": 49714232,
        "related_pages": [
            {
                "title": "Switzerland",
                "url": "/wiki/Switzerland"
            },
            ...
    },
    "status": "success"
}
```


Figure 11: Example of a page in ArabicKT

cally, group A is responsible for generating definitions, and group B is responsible for evaluating the definitions. The evaluators were instructed to assess the definitions across four dimensions: accuracy, completeness, clarity, and overall quality (the detailed evaluation questionnaire can be found in Fig 18). While domain experts were involved in questionnaire design and keyword selection, the large-scale annotation task required substantial human resources. Therefore, we choose to employ students for annotation. We acknowledge this as a limitation of our study and will explicitly address this in the revised paper.

A.5 Details of Evaluation Workflow

The prompts used for the generation of question q_B, entity extraction, generation of question q_R, and determining if the questions are related to Arab-related knowledge are available in Fig 23, 24, 25, and 26 respectively. The generation procedure for q_R is as follows: First, we utilize LLMs to extract as many entities $(e_1, e_2, e_3, ..., e_n)$ as possible from each article. Then, for question generation, we randomly select pairs of distinct entities (always two entities per pair, $e_i, e_j (i \neq j)$) from the extracted set. The LLM is then prompted to generate questions that explore and discuss the relationships between these entity pairs. If two entities are not related, the generated question will be removed through our quality verification. These details will be included in the appendix in the subsequent sections.

We showcased four example questions generated using different concepts from our Arabic BoK in Fig 19, 20, 21, and 22 with two of q_B and two of q_R. The choice presented in bold indicates the correct choice.

The prompt we used for evaluating the model's performance on our generated test dataset is shown in Fig 27, which demands the model to first generate a chain of thoughts and then provide the answer in a specific format.

Human validation Following the methodology outlined in (Abdallah et al., 2024). We recruited six graduate students with backgrounds in Arabic literature from the School of Foreign Languages and NLP majors from the School of Computer Science. These annotators are asked to evaluate the randomly-sampled 1,000 questions across four dimensions (Fluency, Answerability, Relevance, and Non-ambiguity on a scale of 1-5), with results shown in the Tab. 2. The results demonstrate consistently high scores across all dimensions, indicating that the generated questions are generally clear and well-formed.

Evaluation cost analysis For the evaluation of 6M questions, we employed different acceleration strategies based on model types. For proprietary models (GPT-4o), we leveraged OpenAI's batch inference service via API calls, completing the evaluation in approximately 10 hours. For open-source models (Llama-3.1.70B), we utilized vLLM (Kwon

Prompt Template for Filtering the Pages that are unrelated to Arab Culture

```
You are a Wikipedia expert. Your task is to determine whether
a given Wikipedia category directly related to **OR** belongs
to `Generalized Middle East` **OR** `Broader Arab world` with
highly-related historically or geographical connections. It
could be:
1. Countries including `Afghanistan`, `Algeria`, `Comoros`,
`Cyprus`, `Djibouti`, `Morocco`, `Iraq`, `Iran`, `Pakistan`,
`Turkey`, `Tunisia`, `Syria`, `Somalia`, `Yemen`, `Sudan`,
`Libya`, `Egypt`, `Saudi Arabia`, `United Arab Emirates`,
`Qatar`, `Bahrain`, `Kuwait`, `Oman`, `Jordan`, `Lebanon`,
`Palestine`, `Israel`, `Jerusalem`, `Hebron`, `Gaza`,
`Jericho`, etc, modern or ancient.
2. Cities or Locations in above areas.
3. Peoples, Organizations or Persons in above areas.
4. Cultures, Societies, Works, Art, Science, Religions,
Educations, Histories, Geographies, Politics, Economies,
etc in above areas.
5. Other related concepts.
6. If you are not very certain because the relation is ambiguous,
output 1.

---

**Input Information:**
- **Title**: {title}
- **Definition**: {definition}
- **Subcategory Samples**:
```
{subcategories_str}
```
- **Page Samples**:
```
{pages_str}
```
---
```

Figure 12: Prompt template for filtering the pages that are unrelated to Arab Culture

Table 2: Human evaluation scores for generated questions across four quality dimensions

Criterion	Fluency	Answerability	Relevance	Non-ambiguity
Annotator 1	4.681	4.732	4.933	4.553
Annotator 2	4.907	4.627	4.831	4.673
Annotator 3	4.530	4.707	4.907	4.647
Annotator 4	4.809	4.627	4.693	4.273
Annotator 5	4.509	4.350	4.767	4.467
Annotator 6	4.827	4.461	4.929	4.585
Average	**4.711**	**4.584**	**4.843**	**4.533**

et al., 2023) as our inference framework, which required around 20 hours to process all questions.

A.6 Evaluation Results

The relation between category accuracy and average token length of category we examined the correlation between category accuracy and the average token length of passages within each category (tokenized using the Qwen2.5 tokenizer), as illustrated in Fig 7b. The analysis revealed a positive correlation between question-answering accuracy and passage token count, with Qwen2.5-72B-instruct demonstrating a correlation coefficient of 0.5141. Categories with lower performance generally corresponded to shorter passages.

The relation between category accuracy and article frequency in training corpus we conducted a quantitative analysis of article title frequency in model pre-training corpora. Specifically, we performed a comparative study by randomly sampling 100 articles from two contrasting categories: a high-performing category ("*Israelites*") and a low-performing category ("*Culture of Saudi Arabia*"). By analyzing their title frequency distribution in the Arabic101 pre-training dataset (Aloui et al., 2024), we found a stark contrast: concepts from high-performing categories appeared substantially more frequently, with an average occurrence of 13,738.4 instances, whereas concepts from low-performing categories averaged only 168.2 instances. This significant disparity in representation strongly supports our second hypothesis that models exhibit diminished performance on long-tail knowledge with limited presence in pre-training corpora.

More results Due to space limitations in the main text, we only provided the accuracy of the models for Level 2 categories. Here, we present additional results to support the findings within §4.2: Fig 28, 30, and 32 demonstrate the accuracy of GPT-4o, Claude 3.5-Sonnet2, Llama-3.1-70B, and Qwen2.5-72B within the category of Level 1, 2, and 3 respectively. Fig 29, 31, and 33 demonstrate the accuracy of Qwen2.5-3B, Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B, and Qwen2.5-72B within the category of Level 1, 2, and 3 respectively.

Due to the large number of categories in Level 2 (170) and Level 3 (595), we only present the accuracy of 20 categories of these two levels. For Level 1, the complete results of all categories are presented.

A.7 Significance of Definition Completion

Our ultimate goal is to construct a taxonomy leveraging diverse Arab world knowledge to comprehensively evaluate model capabilities. However, Arab world knowledge is vast and complex. Without a systematic knowledge framework (including hierarchical structures, definitions, and ontology (Burgueño et al., 2019; Luo et al., 2021)), misunderstandings of certain concepts can easily occur. Take "Madrasa" (refers to Islamic schools in Arabic) as an example - without proper definition, it might be broadly miscategorized as just a "religious school," leading to incorrect equivalencies between institutions like the 11th century Nizamiyya Madrasa in Baghdad (an academic center teaching philosophy and mathematics) and 18th century Quranic recita-

tion schools in rural Morocco (focusing on basic religious instruction). This oversimplification obscures their fundamental differences in advancing scientific knowledge versus religious education.

Therefore, definitions are integral to our taxonomy design. Yet definitions serve several other important functions: (1) **Knowledge Integration**: Definitions facilitate the integration of new knowledge (particularly Arabic-language content from external sources) by providing clear criteria for determining appropriate hierarchical placement within the taxonomy. (2) **Enhanced Retrieval Capabilities**: Definitions enable sophisticated semantic similarity computations for identifying related nodes and can improve knowledge navigation interpretability (users can preview node definitions before deciding whether to explore subtrees). (3) **Data Synthesis**: Rewriting knowledge to inject knowledge into LLMs is a common practice (Yang et al., 2024b; Maini et al., 2024). Having definitions helps the rewriting models understand the meaning of each concept, preventing biases in different models' interpretations of concepts (especially domain-specific concepts), thereby improving the quality of synthesized data.

A.8 Coverage Evaluation Details

We evaluate the coverage of ArabKT by assessing how well ArabKT encompasses the knowledge contained in common Arabic datasets. The exhaustive semantic matching between every dataset sample and knowledge points is computationally intensive. Therefore, we adopt a RAG-inspired approach (Lewis et al., 2020) for efficient retrieval and coverage assessment (as shown in Fig. 4). First, we encode the knowledge point within each node in ArabKT using the GTE model (Li et al., 2023b) to construct an embedding database. For each query text from the datasets, we similarly extract its embedding and retrieve the top-k relevant knowledge points based on embedding similarity. Finally, we employ LLM (i.e., GPT-4o (Hurst et al., 2024)) to determine whether any retrieved knowledge points semantically cover the query text.

Keywords for filtering Arab-related wiki pages

```
# Geographic Region-Related Vocabulary
"Middle East", "Middle Eastern", "Levant", "Mashriq", "Gulf", "Arabian Peninsula", "Fertile Crescent", "Sinai", "Mesopotamia",
"Anatolia", "Levantine", "Caspian", "Persian Gulf", "Tigris", "Euphrates", "Arabian Sea", "Red Sea", "Dead Sea",
"Persian Plateau", "Zagros Mountains", "Taurus Mountains", "Arabian Desert", "Syrian Desert", "Nile Delta",
"Tigris-Euphrates Valley", "Dead Sea Rift", "Mount Lebanon", "Mount Hermon", "Sinai Peninsula",
"Shatt al-Arab", "Strait of Hormuz", "Strait of Tiran", "Strait of Bab el Mandeb", "Gulf of Aqaba", "Gulf of Oman",
"Gulf of Suez", "Sumer", "Median Empire",

# Arab Countries
"Arab", "Arabian", "Algeria", "Algerian", "Bahrain", "Bahraini", "Comoros", "Comorian",
"Djibouti", "Djiboutian", "Egypt", "Egyptian", "Iraq", "Iraqi", "Jordan", "Jordanian",
"Kuwait", "Kuwaiti", "Lebanon", "Lebanese", "Libya", "Libyan", "Morocco", "Moroccan",
"Oman", "Omani", "Palestine", "Palestinian", "Qatar", "Qatari", "Saudi Arabia", "Saudi", "Saudi Arabian",
"Somalia", "Somali", "Sudan", "Sudanese", "Syria", "Syrian", "Tunisia", "Tunisian",
"United Arab Emirates", "Emirati", "UAE", "Yemen", "Yemeni",

# Middle East Countries
"Iran", "Iranian", "Persian", "Israel", "Israeli", "Turkey", "Turkish", "Cyprus", "Cypriot",
"Tehran", "Isfahan", "Shiraz", "Mashhad", "Tabriz", "Qom", "Ahvaz", "Karaj", "Kermanshah", "Urmia", "Rasht", "Kerman",
"Jerusalem", "Tel Aviv", "Haifa", "Beersheba", "Rishon LeZion", "Ashdod", "Netanya", "Petah Tikva", "Beit Shemesh", "Bnei Brak",
"Istanbul", "Ankara", "Izmir", "Bursa", "Adana", "Gaziantep", "Konya", "Antalya", "Kayseri", "Mersin", "Eskisehir", "Diyarbakır",
"Samsun", "Denizli", "Şanlıurfa", "Malatya", "Kahramanmaraş",
"Nicosia", "Limassol", "Larnaca", "Famagusta", "Paphos", "Kyrenia", "Morphou", "Polis", "Magheramason", "Deryneia",

# Arabic Cities
"Algiers", "Oran", "Constantine", "Manama", "Riffa", "Muharraq", "Moroni", "Mutsamudu", "Fomboni", "Djibouti", "Ali Sabieh", "Tadjourah",
"Cairo", "Alexandria", "Giza", "Luxor", "Aswan", "Port Said", "Suez", "Mansoura", "Tanta", "Ismailia", "Hurghada", "Sharm El Sheikh",
"Baghdad", "Basra", "Mosul", "Erbil", "Najaf", "Kirkuk", "Karbala", "Sulaymaniyah",  "Samarra", "Dohuk",
"Amman", "Aqaba", "Irbid", "Zarqa", "Russeifa", "Kuwait City", "Al Ahmadi", "Hawalli",
"Beirut", "Tripoli", "Sidon", "Tyre", "Baalbek", "Tripoli", "Benghazi", "Misrata", "Sabha", "Tobruk",
"Rabat", "Casablanca", "Marrakech", "Fes", "Tangier", "Agadir", "Meknes", "Oujda", "Kenitra",
"Tetouan", "Tetfouth", "Nador", "Muscat", "Salalah", "Sohar", "Nizwa", "Buraimi",
"Ramallah", "Gaza City", "Hebron", "Nablus", "Bethlehem", "Jericho", "Jenin",
"Doha", "Al Rayyan", "Al Wakrah", "Al Khor",
"Riyadh", "Jeddah", "Mecca", "Medina", "Dammam", "Khobar", "Ta'if", "Tabuk", "Buraidah", "Najran", "Al Khafji",
"Mogadishu", "Hargeisa", "Bosaso", "Kismayo", "Baidoa", "Galkayo",
"Khartoum", "Omdurman", "Nyala", "Port Sudan", "Kassala", "El Obeid",
"Damascus", "Aleppo", "Homs", "Latakia", "Hama", "Deir ez-Zor", "Raqqa", "Idlib", "Tartus",
"Tunis", "Sfax", "Sousse", "Bizerte", "Kairouan", "Gabès", "Gafsa", "Kasserine", "Monastir", "Mahdia",
"Dubai", "Abu Dhabi", "Sharjah", "Al Ain", "Ajman", "Fujairah", "Ras Al Khaimah", "Umm Al Quwain",
"Sana'a", "Aden", "Taiz", "Al Hudaydah", "Ibb", "Dhamar", "Mukalla", "Hadramawt", "Dhi Qar",

# Historical Cities
"Babylon", "Nineveh", "Persepolis", "Uruk", "Byblos", "Tyre", "Sidon", "Petra", "Carthage", "Antioch", "Ephesus", "Palmyra",
"Hatra", "Samarra", "Ur", "Susa", "Mari", "Harran", "Seleucia", "Byzantium", "Ctesiphon", "Tyropolis",
"Edessa", "Antiochia Parva", "Apamea", "Arsamosata", "Halab", "Corinthopolis", "Seleucia on the Tigris",
"Damascus", "Aleppo", "Homs",

# Nation and Culture related Vocabulary
"Bedouin", "Kurdish", "Amazigh", "Berber", "Druze", "Persian", "Turkic", "Phoenician", "Circassian",
"Assyrian", "Aramean", "Coptic", "Maronite", "Nabatean", "Mamluk", "Seljuk", "Abbasid",
"Umayyad", "Fatimid", "Safavid", "Sassanid", "Achaemenid", "Neo-Assyrian", "Elamite",
"Lydian", "Urartu", "Zoroastrian", "Mithraism",

# Language-related Vocabulary
"Arabic", "Hebrew", "Persian", "Turkish", "Kurdish", "Aramaic", "Syriac", "Berber",
"Amharic", "Akkadian", "Coptic", "Ottoman Turkish", "Elamite", "Sumerian",

# Religious Concept
"Islam", "Sunni", "Shia", "Sufism", "Christianity", "Eastern Orthodox", "Coptic Christianity", "Baha'i", "Antiochian",
"Maronite Christianity", "Judaism", "Zoroastrianism", "Bahá'i Faith", "Safaidi",
"Yazidism", "Druze Faith", "Kabbalah", "Ahmadiyya", "Shiite Islam", "Wahhabism",
"Salafism", "Hasidism", "Mysticism", "Quran", "Hadith", "Sharia", "Fiqh",
"Fatwa", "Madrasa", "Caliphate", "Imam", "Mosque", "Minaret", "Hajj",
"Ramadan", "Eid al-Fitr", "Eid al-Adha", "Mecca", "Medina", "Kaaba",
"Hijab", "Jihad", "Dhikr", "Sufi Whirling", "Rumi", "Alchemy in Islamic World",

# Conture Concept
"Islamic Golden Age", "Majlis", "Bazaar", "Souk", "Hammam", "Qanat", "Calligraphy", "Arabic Art", "Persian Carpets", "Islamic",
"Mosques", "Minarets", "Sufi Poetry", "Islamic Jurisprudence", "Ottoman", "Abbasid", "Umayyad",
"Seljuk", "Fatimid", "Safavid", "Shia Islam", "Sunni Islam", "Andalusian", "Bedouin",
"Nomadic Culture", "Maqam", "Tarab", "Dabke", "Henna", "Islamic Calligraphy", "Persian",
"Persian", "Arabic",

# Myth and Philosophy
"Sumerian", "Babylonian", "Assyrian", "Mesopotamian", "Zoroastrian", "Mithraism", "Neoplatonism", "Islamic", "Avicennian",
"Al-Farabi", "Ibn Sina", "Ibn Rushd", "Rumi's", "Falasifa"
```

Figure 13: Keywords for filtering Arab-related wiki pages

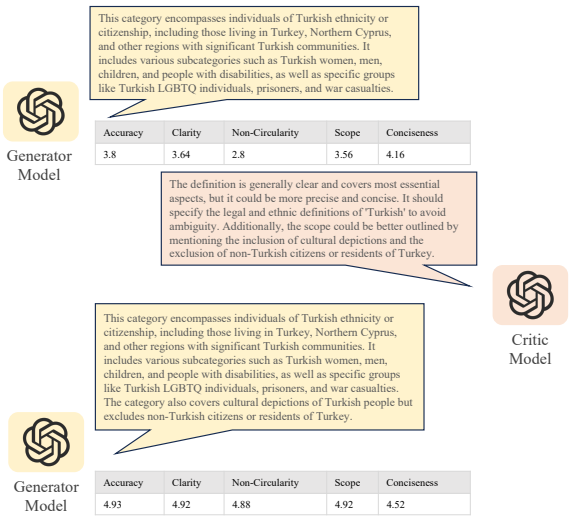


Figure 14: Example of multi-round definition completion.

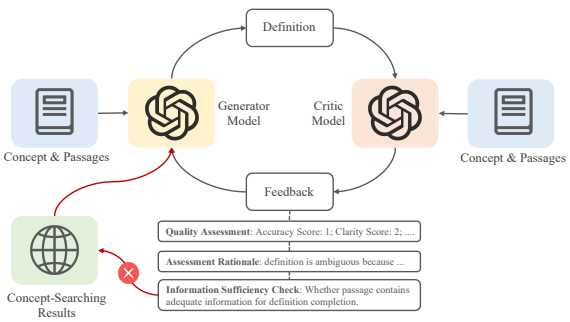


Figure 15: Workflow of definition completion.

Prompt for Definition Completion

```
**You are a Wikipedia Category Definition Expert.**

**Your task is to create a clear, concise, and accurate
definition for a given Wikipedia Category based on the
provided information. Follow these guidelines to ensure
the definition meets Wikipedia's standards:**

1. **Be Clear and Concise:** Use straightforward language
without unnecessary complexity. Aim for brevity while
ensuring all essential aspects of the category are covered.
2. **Define the Scope:** Clearly outline what is included
in the category and, if necessary, what is excluded. Specify
any relevant geographical, temporal, or organizational boundaries.
3. **Avoid Redundancy and Circular Definitions:** Do not
use the category title or its synonyms within the definition
to prevent circular reasoning.
4. **Include Necessary Context:** Provide any additional
context that helps in understanding the category, such
as related organizations, time periods, or specific
attributes relevant to the category.
5. **Maintain Objectivity:** Present the definition in
an unbiased manner without subjective opinions or evaluations.
6. **Use Consistent Formatting:** Adhere to Wikipedia's
style guidelines for category definitions, ensuring
uniformity across all definitions.

---

**Input Information:**
- **Title**: `{title}`
- **Subcategories**:
  ```
  {subcategories_and_definition_str}
  ```
- **Pages**:
  ```
  {pages_and_definition_str}
  ```

---

**Output:**
Generate a single, well-structured sentence or a short
paragraph that serves as the definition for the given
Wikipedia Category, adhering to the guidelines outlined above.

---

**Example:**
*If provided with the following input:*
- **Title**: Category:American poets
- **Subcategories**:
  - **20th-century American poets**: Poets from America
  who were active in the 20th century.
  - **African-American poets**: Poets of African-American
  heritage.
- **Pages**:
  - **Maya Angelou**: American poet, memoirist, and civil
  rights activist.
  - **Robert Frost**: Renowned American poet known for
  his depictions of rural New England life.

*The generated definition should be:*
This category encompasses poets from the United States
across various time periods and diverse backgrounds,
recognized for their contributions to literature.

---
**Your Task:**
Using the provided input information, generate an
appropriate Wikipedia Category definition following the
structure and guidelines above.
```

Figure 16: Prompt for definition completion

Prompt for Definition Critique

```
Evaluate this following definition for:
1. Accuracy: Assess if the definition accurately and
completely describes the category without omitting
critical attributes.
2. Clarity: Determine if the definition is clear, precise,
and free from ambiguity.
3. Non-Circularity: Ensure the definition does not repeat
the category name or use synonyms that lead to circular
reasoning.
4. Scope: Verify that the definition correctly outlines
what is included in the category without being overly broad
or too narrow.
5. Conciseness: Check if the definition is succinct while
maintaining completeness.

You should be evaluating only and not attemping to solve
the task.
Only output "PASS" if all criteria are met and you have no
further suggestions for improvements.
Output your evaluation concisely in the following **JSON** format.

{
  "evaluation": "<PASS, NEEDS_IMPROVEMENT, or FAIL>",
  "feedback": "<What needs improvement and why>"
}
```

Figure 17: Prompt for definition critique

Instructions for Head-to-head Evaluation

```
Evaluation Instructions:
1. You will receive multiple pairs of concept definitions
(A and B), with each pair containing two different
definitions for the same concept
2. Please do not focus on the source of the definitions
(human/machine), evaluate solely based on the definitions
themselves
3. Compare each pair of definitions and mark your assessment
on the corresponding scales

-----------------
Evaluation Section
-----------------

Concept: [Specific concept name]

Definition A: [First definition]
Definition B: [Second definition]

1. Accuracy Comparison
Which definition is more accurate in terms of facts and content?
_ Definition A is more accurate
_ Both are equally accurate
_ Definition B is more accurate

2. Completeness Comparison
Which definition provides more comprehensive coverage of the
concept's core elements?
_ Definition A is more complete
_ Both are equally complete
_ Definition B is more complete

3. Clarity Comparison
Which definition is clearer and easier to understand?
_ Definition A is clearer
_ Both are equally clear
_ Definition B is clearer

4. Overall Quality Comparison
Considering all factors, which definition do you think is
better overall?
_ Definition A is better
_ Both are equally good
_ Definition B is better
```

Figure 18: Instructions for head-to-head evaluation of LLM-generated definition and human-annotated definition

What is the historical significance of the village of Nagrig?
A. It was the capital of Ancient Egypt.
B. It bordered the Ancient city of Sais in the fifth nome of Lower Egypt.
C. It was founded during the reign of Saladin.
D. It was a major city in the Middle Kingdom of Egypt.

Figure 19: An example QA (q_B) for concept "Avraham Kalfon".

What is the Yerushalmi, or Jerusalem Mixed Grill, primarily associated with?
A. A religious text
B. A culinary dish
C. A historical event
D. A style of dance

Figure 20: An example QA (q_B) for concept "Zabid".

How does the American Jewish volunteer organization AMIT incorporate Jewish values into its educational mission for students in Israel?
A. By focusing solely on technological education and ignoring religious studies
B. By creating an exclusive environment for students from affluent backgrounds
C. By providing a balanced education that includes religious Jewish studies alongside academic and technological subjects
D. By prioritizing military training over academic subjects

Figure 21: An example QA (q_R) for concept "Collège Élite (Beirut)".

What position did Habibullah Khan Karzai hold at the United Nations?
A. Afghan Ambassador to the United States
B. Permanent Representative from Afghanistan
C. Special Envoy to the European Union
D. Afghan Delegate to the World Bank

Figure 22: An example QA (q_R) for concept "Ahmad al-Khatib".

Prompt Template for q_B Generation

```
**Instructions:**
You are an educator designing assessment questions to test
understanding of a specific knowledge point. Based on the
provided article, generate a set of new close-book questions
that vary in type and difficulty. The questions should
comprehensively cover the key aspects of the knowledge point.

**Knowledge Point:**
{concept}

**Article:**
<article>
{passage}
</article>

Instructions:
- **Language:** English
- **Number of Questions:** 3
- **Types of Questions:** Multiple-choice
- **Difficulty Levels:** Vary the difficulty from basic recall to
higher-order thinking skills
- **Content Requirements:**
  - Ensure questions are directly related to the information in the
article
    - Do not mention the article in the questions
  - Do not require referring back to the original context; questions
should be self-contained
    - Avoid ambiguity; questions should be clear and precise, all
entities should be defined and avoid using pronouns and ambiguous
terms like "the book", "the article", etc.
  - Ensure that each correct answer is distinct, clear, definite,
and unambiguous
    - Provide correct answers for each question.
    - Please use A,B,C,D to format your options.
    - The questions should focus on the topic of {concept}
    - Provide a reason for the correct answer.

**Output Format:**
1. **Question:** [Question Text]
    - A) [Option A]
    - B) [Option B]
    - C) [Option C]
    - D) [Option D]
    - **Correct Answer:** [A/B/C/D]
    - **Reason:** [Reason for the correct answer]

---
**Your Questions:**
```

Figure 23: Prompt template for q_B generation

Prompt Template for Entity Extraction

```
As a knowledge analyzer, your task is to dissect and understand
a lecture passage (with title) provided by the user. You are
required to perform the following task:
**Extract Entities**: Identify and list all significant "nouns"
or entities mentioned within the script. These entities should
include, but are not limited to:
    * People: Any lecturers, historical figures, or individuals
mentioned.
    * Places: Specific locations or institutions referenced.
    * Objects: Any concrete objects or tools discussed within
the context of the lecture.
    * Concepts: Key academic concepts, theories, or themes
that are central to the lecture's discussion.

Ensure that your summary is brief yet comprehensive, and the
list of entities is detailed and accurate. Structure your
response in a JSON format to organize the information effectively.
Do not include the title of the passage as an entity in your response.

Here is the format you should use for your response (in JSON):

"entities": ["entity1", "entity2", ...]

**Input**:
<Title>
{title}
</Title>
<Passage>
{passage}
</Passage>
```

Figure 24: Prompt template for entity extraction

Prompt Template for q_R Generation

```
**Instructions:**
You are an educator designing assessment questions to test
understanding of a specific knowledge point. Based on the
provided article, generate a question discussing the interaction
between the knowledge point and the provided entity within the
context of the article.

**Knowledge Point:**
{concept}

**Entity:**
{entity}

**Article:**
<article>
{passage}
</article>

Instructions:
- **Language:** English
- **Number of Questions:** 1
- **Types of Questions:** Multiple-choice
- **Content Requirements:**
  - Ensure questions are directly related to the information in
the article
    - Do not mention the article in the questions
    - Do not require referring back to the original context;
questions should be self-contained
    - Avoid ambiguity; questions should be clear and precise, all
entities should be defined and avoid using pronouns and ambiguous
terms like "the book", "the article", etc.
  - Ensure that each correct answer is distinct, clear, definite,
and unambiguous.
    - Provide correct answers for each question.
    - Please use A,B,C,D to format your options.
    - Provide a reason for the correct answer.

**Output Format:**
1. **Question:** [Question Text]
    - A) [Option A]
    - B) [Option B]
    - C) [Option C]
    - D) [Option D]
    - **Correct Answer:** [A/B/C/D]
    - **Reason:** [Reason for the correct answer]

---
**Your Questions:**
```

Figure 25: Prompt template for q_R generation

Prompt Template for Determining if Question is Related to Arab-knowledge

```
**Task Description**

You are an AI language model analyst. Your task is to
determine whether a given question is of good quality
based on the following criteria:
**Related to Arabic Culture**: the question is directly
related to **OR** belongs to `Generalized Middle East`
**OR** `Broader Arab world` with highly related historical
or geographical connections. It could be:
1. Countries including `Afghanistan`, `Algeria`, `Comoros`,
`Cyprus`, `Djibouti`, `Morocco`, `Iraq`, `Iran`, `Pakistan`,
`Turkey`, `Tunisia`, `Syria`, `Somalia`, `Yemen`, `Sudan`,
`Libya`, `Egypt`, `Saudi Arabia`, `United Arab Emirates`,
`Qatar`, `Bahrain`, `Kuwait`, `Oman`, `Jordan`, `Lebanon`,
`Palestine`, `Israel`, `Jerusalem`, `Hebron`, `Gaza`,
`Jericho`, etc, modern or ancient.
2. Cities or Locations in the above areas.
3. Peoples, Organizations, or Persons in the above areas.
4. Cultures, Societies, Works, Art, Science, Religions,
Educations, Histories, Geographies, Politics, Economies,
etc in the above areas.
5. Other related concepts.

If the question satisfies the criteria, then the
question is of good quality. Otherwise it is of bad quality.

**Output Instruction**

You should return GOOD or BAD first, then give an explanation
in English. Please output the results in the following format:
<quality>GOOD/BAD</quality>
<explanation>provide a explanation here</explanation>

**Input**
{question}
```

Figure 26: Prompt Template for determining if question is related to Arab-knowledge

Question Evaluation Prompt

```
Answer the following multiple choice question. The last
line of your response should be of the following format:
'Answer: $LETTER' (without quotes) where LETTER is one
of ABCD. Think step by step before answering.

Who were the founding members of the band Speed Caravan?

A) Mehdi Haddab, Hamdi Benani, and Keziah Jones
B) Mehdi Haddab, Pascal « Pasco » Teillet, and Hermione
Frank
C) MC Spex, Rachid Taha, and Rodolphe Burger
D) The Chemical Brothers, The Cure, and Jimmy Page
```

Figure 27: Question evaluation prompt following OpenAI (2024)

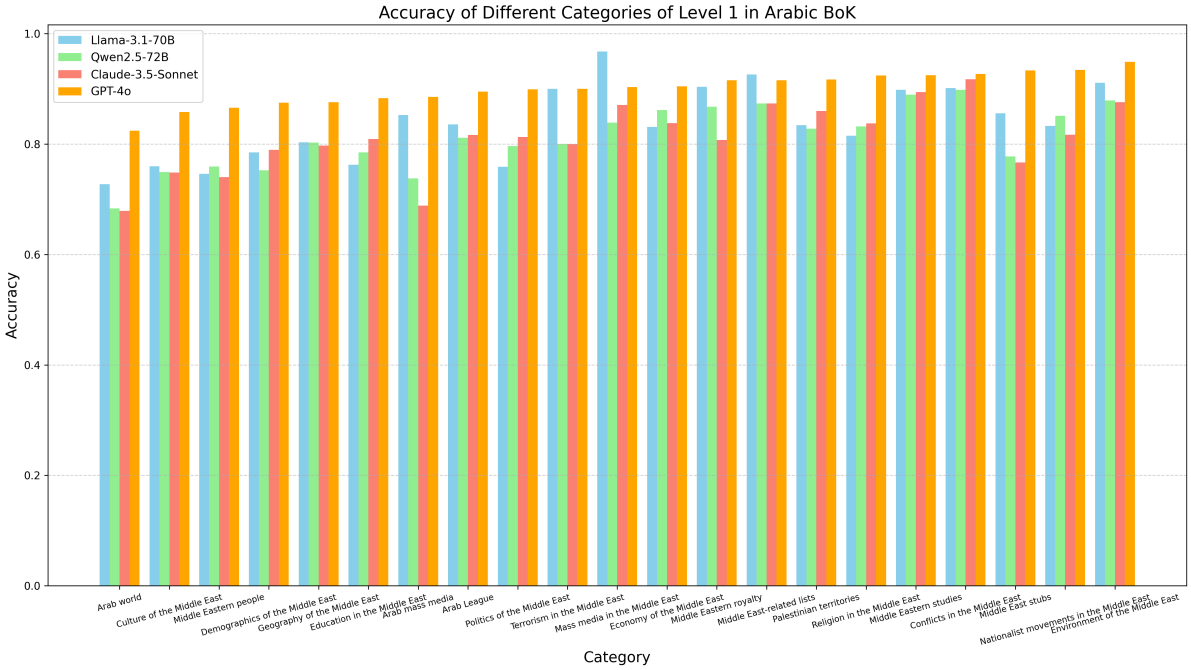


Figure 28: Accruracy on categories of level 1 in Arabic BoK for prevalent LLMs.

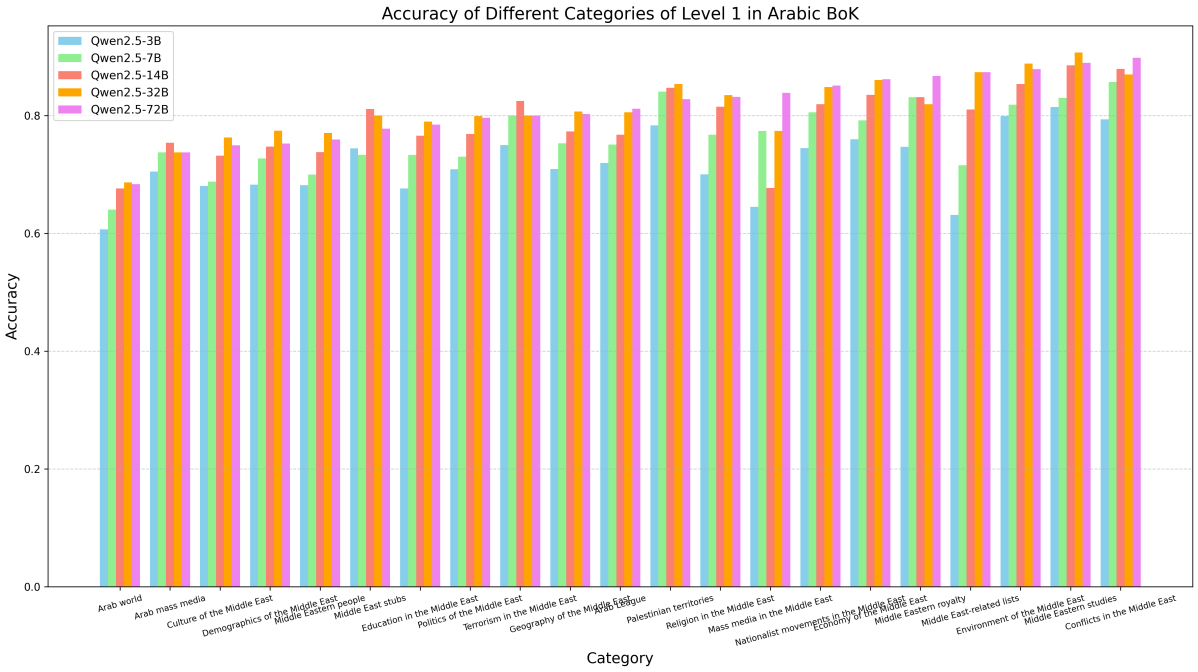


Figure 29: Accruracy on categories of level 1 in Arabic BoK for Qwen2.5 series models.

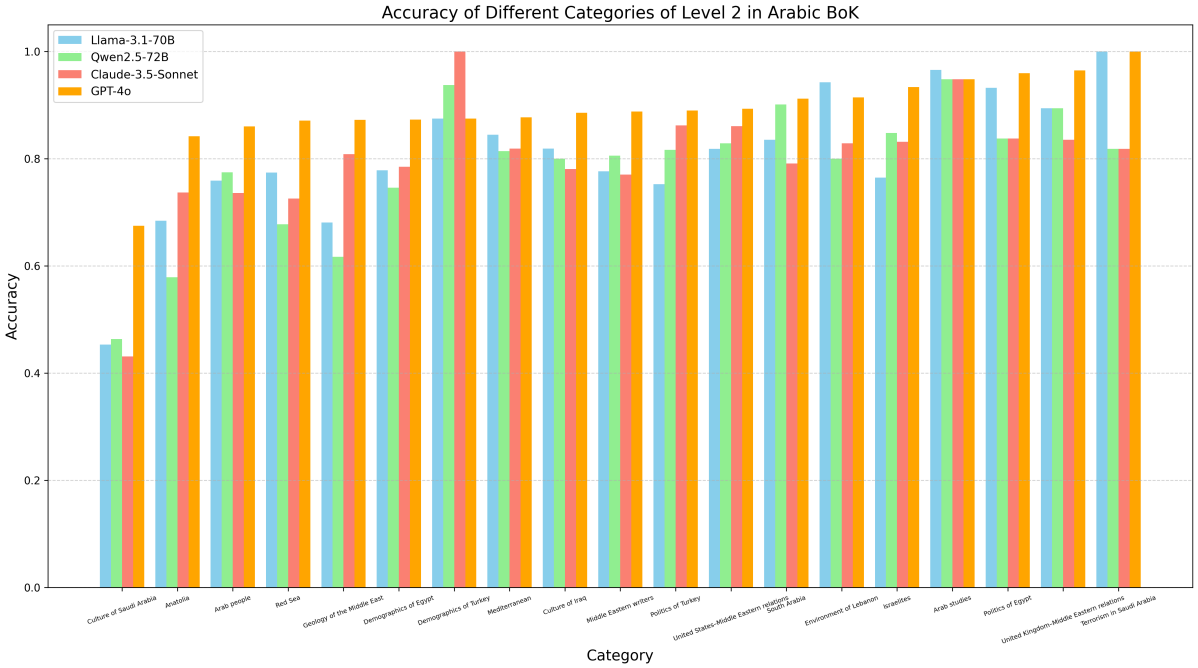


Figure 30: Accruracy on categories of level 2 in Arabic BoK for prevalent LLMs.

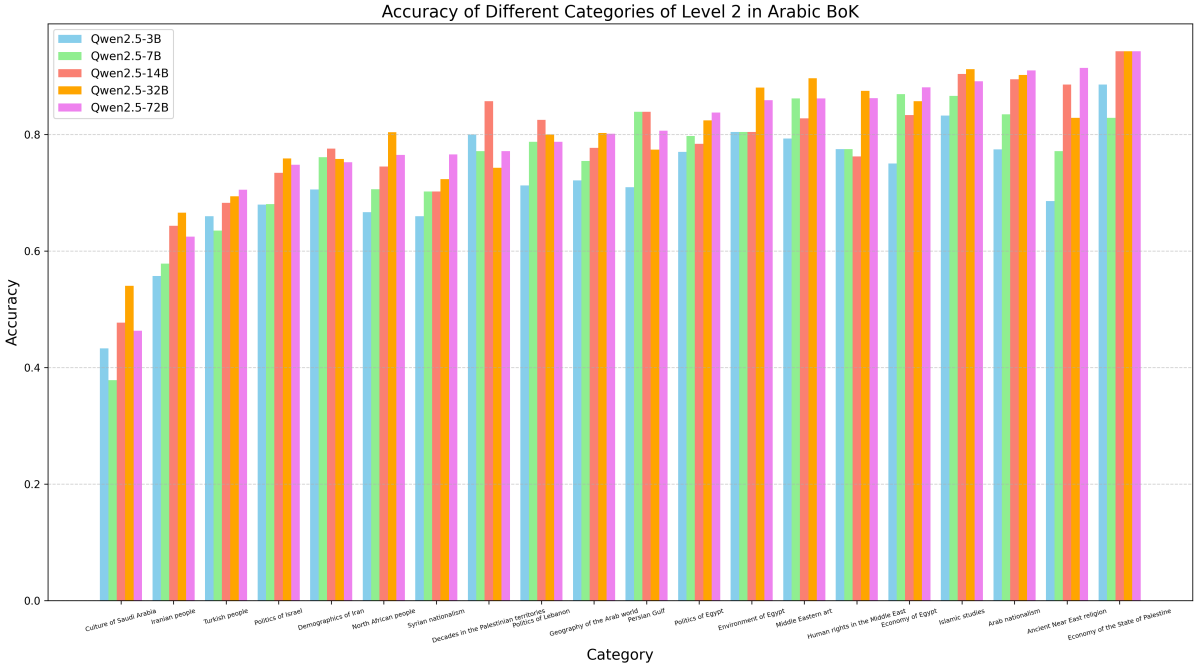


Figure 31: Accruracy on categories of level 2 in Arabic BoK for Qwen2.5 series models.

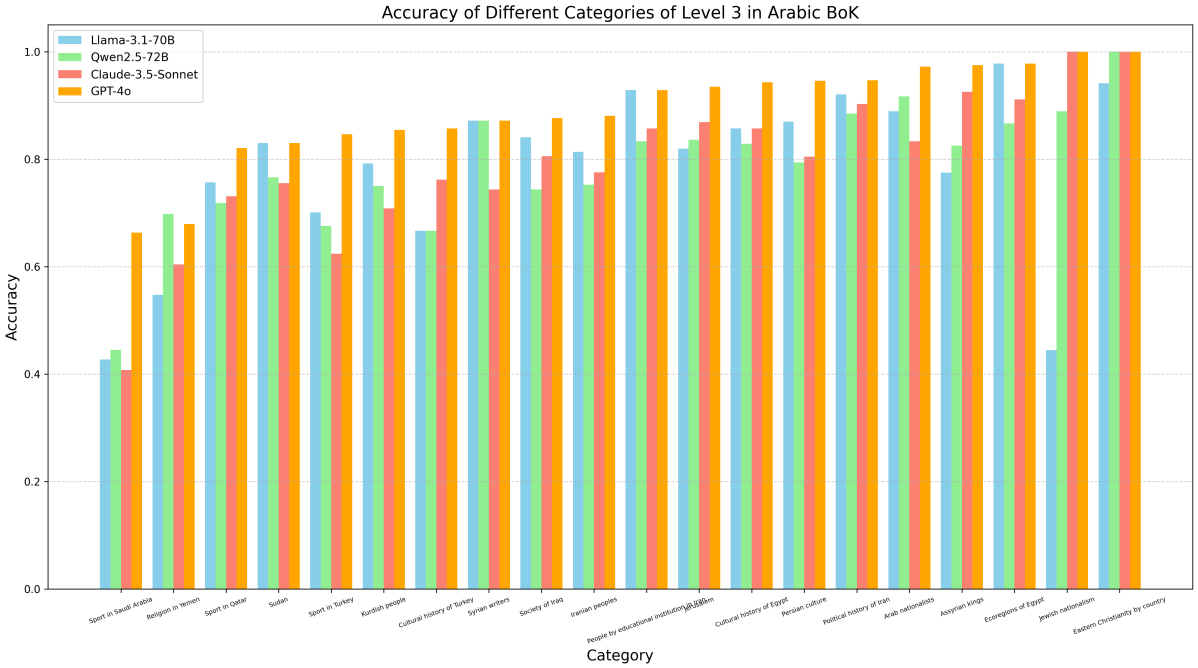


Figure 32: Accruracy on categories of level 3 in Arabic BoK for prevalent LLMs.

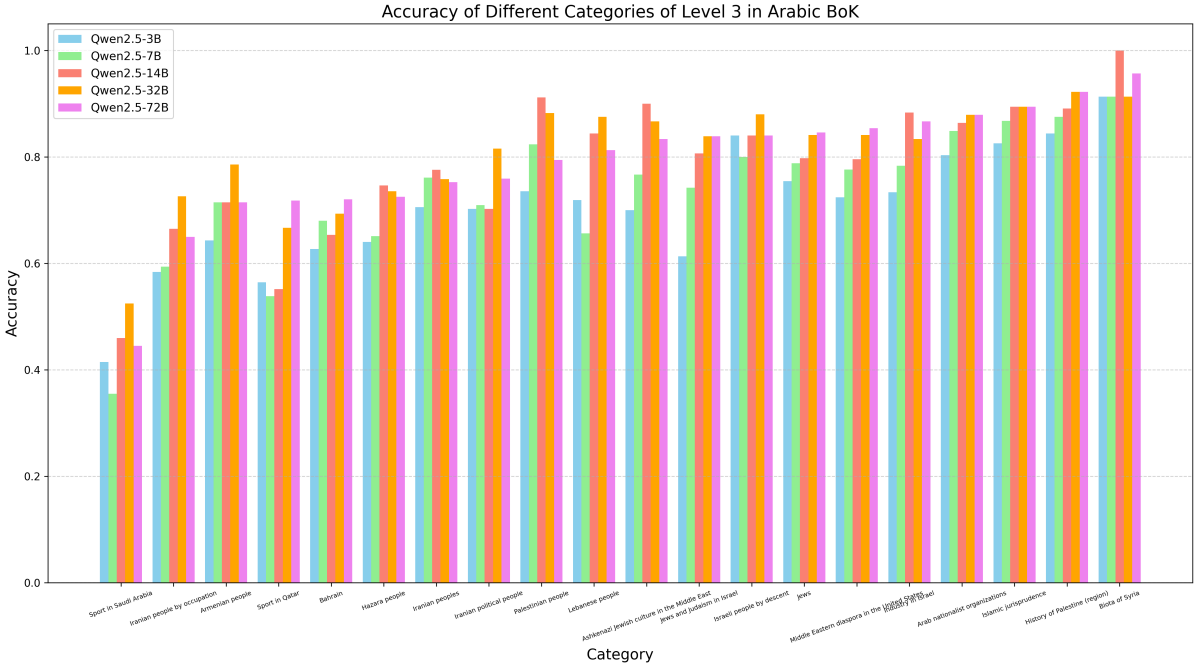


Figure 33: Accruracy on categories of level 3 in Arabic BoK for Qwen2.5 series models.