ArabicKT: A Comprehensive Arabic 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 only a small fraction of specific knowledge. To address this gap, we propose an automated approach to generate comprehensive evaluation data and introduce ArabicKT, an Arabic Knowledge Taxonomy derived from Wikipedia and Wikidata. ArabicKT organizes $140,433$ categories and 1.67 million articles into a 15-layer tree structure, covering 77% of the Arabic pretraining corpus and 84% of existing Arabic benchmarks. Leveraging LLMs, we developed an automated pipeline to generate 6 million question-answer pairs for Arab-world knowledge. Our experiments reveal two key insights: (1) Models perform better on knowledge points that appear more frequently in training data, and (2) larger models exhibit superior mastery of granular cultural, religious, and historical knowledge. These findings indicates the importance of training data distribution and model scale in domain-specific knowledge acquisition, offering 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 pages (Vrandečić and Krötzsch, 2014). Even in specific domains like Arab-related knowledge, the gap is significant. Arabic-MMLU covers only 40 categories with $14,575$ questions (Koto et al., 2024), while Wikipedia has over $140,000$ Arab-related categories and 1 million pages. This huge disparity makes it hard to fully assess models in specific knowledge domains. Limited evaluation data often misses important long-tail knowledge (Üstün et al., 2024; Kim et al., 2008).

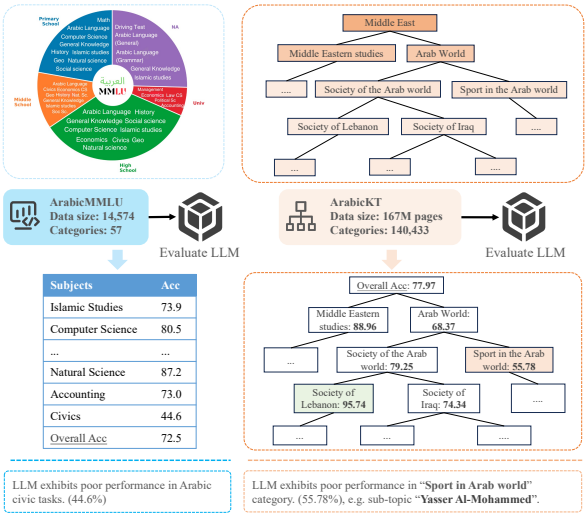


Figure 1: Overview of ArabicMMLU and ArabicKT evaluation benchmarks for assessing LLMs' Arabic knowledge. The numbers and accuracies within ArabicMMLU is from (Koto et al., 2024).

To deal with this problem, we need to move towards automatic generation of evaluation data. This approach presents two main challenges: generating high-quality evaluation data and ensuring comprehensive coverage across topics. Recent advances in LLMs make automated data generation feasible (Yang et al., 2024b). We can use LLMs to replace manual annotation, similar to the LLM-as-a-Judge (Zheng et al., 2024). For comprehensive coverage, encyclopedias serve as valuable references. This aligns with the concept of Body of Knowledge (BOK) in professional contexts (contributors, 2024). 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. While we initially planned to build an Arabic BOK, we realized it requires considerable expertise. Instead, we aim to construct an Arabic Knowledge Taxonomy (ArabicKT) as an initial prototype that could contribute to it. Wikipedia is selected as our foundation due to the commonly usages in pre-training (Touvron et al., 2023a,b) and evaluating LLMs (Geva et al., 2021; Yang et al., 2018). Additionally, Wikidata's category system, which assigns one or more categories to each article, creates a vast network representing unified concepts that can serve as ontology. However, we found several challenges in category network. First, about 83% of category definitions are missing or inaccurate, making them difficult to understand. Second, around 27% of categories have incorrect associations possibly due to editing errors, creating cycles in the network. Third, the complex graph structure makes human observation and analysis impractical (Suchanek et al., 2007). To address these issues, we developed a agentic process to correct those errors and convert the network into a more manageable tree structure. As a result, we build an Arabic Knowledge Taxonomy with 15 layers, containing $140,433$ nodes (categories) and 1.67 million articles. This taxonomy covers around 77% of Arabic pre-training corpus and 84% of Arabic benchmarks.

In addition, 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. To ensure thorough and accurate assessment, we approached question generation from multiple views and verified the generated reason for sampled questions and answers. This process yielded 6 million question-answer pairs for evaluating various language models. Our experiments revealed two key findings. First, ***models perform better on knowledge points that appear more frequently in training data***. As shown in Fig. 1, topics under "Sport in the Arab world" typically have limited coverage (around 21 tokens in Wikipedia articles) and appear infrequently (41 times) in the Arabic corpus, resulting in a lower accuracy of 55.78%. In contrast, topics under "Society of Lebanon" have extensive coverage ($5,348$ tokens) and frequent appearances ($58,014$ times), achieving a much higher accuracy of 95.74%. Second, ***model size correlates with knowledge point granularity***. Larger models can better master detailed cultural, religious, and historical knowledge, as shown in Fig. 12. These phenomena provide a foundation for understanding model capabilities and guide future improvements.

The contributions of this work are summarized as follows: First, we introduce ArabicKT, a comprehensive Arabic Knowledge Taxonomy derived from Wikipedia and Wikidata, containing $140,433$ categories and 1.67 million articles across 15 layers. 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 important patterns about how LLMs learn and retain knowledge: models perform better on frequently appearing information in training data, and larger models show superior ability in handling detailed cultural, religious, and historical knowledge. 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 ArabicKT and evaluations based on it. Our primary data sources are Wikipedia(wik, 2024) and WikiData (Vrandečić and Krötzsch, 2014), which provide extensive articles along with their hierarchical category relationships. Using WikiData's API, 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-Arabic content and articles with content lacking valid information. This initial process resulted in a directed graph of knowledge from Arab world.

We faced two main challenges in converting this graph into a practical tree architecture. The first challenge was missing or incorrect category definitions. To address the correctness of the generation, we are motivated by self-improved frameworks (Dhuliawala et al., 2023; Zhang et al., 2023; Weng et al., 2022). Following them, we developed a pair of agentic models that work together - one for generating definitions and another for critiquing them, allowing iterative improvements. The second

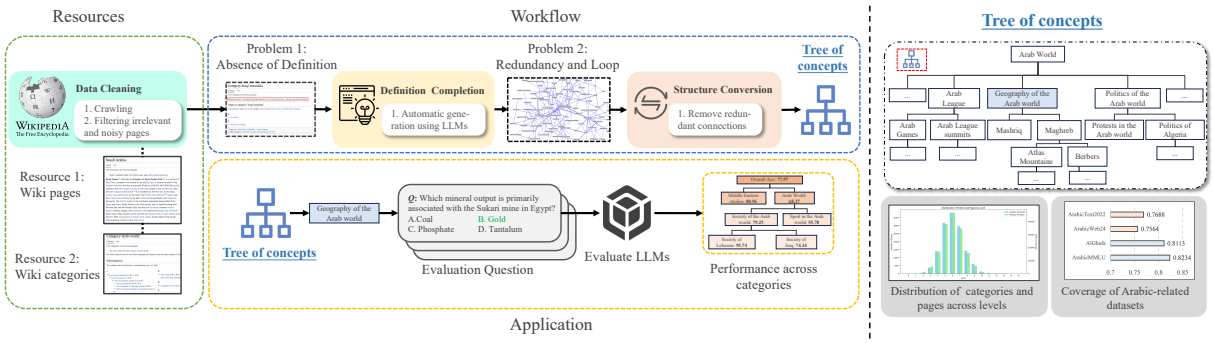


Figure 2: Constructing workflow and application of our ArabicKT (Arabic Knowledge Taxonomy).

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 ArabicKT to guide question generation for evaluating existing LLMs. This evaluation process produced what we call an Accuracy Tree, which provides detailed analysis into different language models' capabilities across various 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 approximately 0.2 million pages from our dataset.

Semantic Filtering. We also implemented a two-stage filtering combining heuristic rules and LLM. First, 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. 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. Although we are studying Arab world knowledge, language is not our main focus. For simplicity, we will use English language articles and categories in our research.

2.3 Definition Completion

Our analysis of Wikipedia categories revealed that only 17.3% contain valid definitions. They either

providing overly brief descriptions, containing irrelevant content, or lack definitions entirely. For instance, the category "Water transport in Iraq" merely states "By consensus, this category should not contain biography articles", exemplifying this widespread definitional inadequacy.

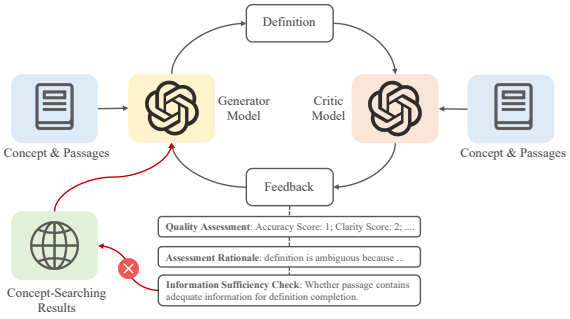


Figure 3: Workflow of definition completion.

To address this issue, we implemented a pair of agentic models for iterative definition completion, as shown in Fig. 3. 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.

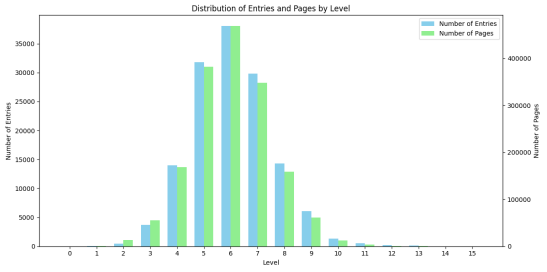
Tree Conversion. We aimed to simplify nodes that had multiple super-categories to create a more human-comprehensible structure. Our simplification process involved three steps: First, for each node with multiple super-categories, we removed redundant connections where one super-category was already a 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 Arabic Knowledge Taxonomy

Following the approach in the previous section, we constructed a ArabicKT 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



(a)

Figure 4: statistics of Arabic Knowledge Taxonomy.

The ArabicKT contains a hierarchical structure spanning 15 layers and encompassing $140,433$ dis-

tinct categories. These categories are linked to a substantial collection of 1.67 million articles. Fig. 4 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 ArabicKT's most information-rich region.

3.2 Coverage

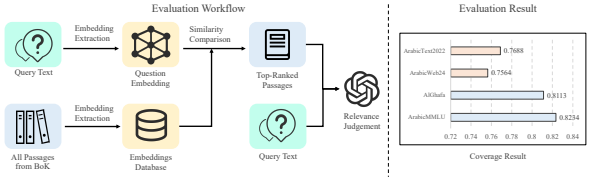


Figure 5: Evaluation workflow and result statistics of the coverage of ArabicKT.

In this section, we evaluate the coverage of ArabicKT by assessing how well ArabicKT encompasses the knowledge contained in common Arabic datasets. Here, semantic coverage refers to that the knowledge points in ArabicKT 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 dataset (AlGhafa (Almazrouei et al., 2023) and ArabicMMLU (Koto et al., 2024)) and two Arabic pre-training datasets (ArabicText2022 (BAAI et al., 2022) and ArabicWeb24 (Farhat et al., 2024)).

The evaluation workflow is shown in Fig. 5. We adopt a RAG-inspired approach (Lewis et al., 2020) for efficient retrieval and coverage assessment. More details are in App. A.7. 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. 5. The ArabicKT achieves coverage rates of 76.88% and 75.64% on training corpus, while achieves 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 ArabicKT, 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 ArabicKT. It indicates that ArabicKT contains large amount of new knowledge points to evaluate LLMs.

3.3 Precision

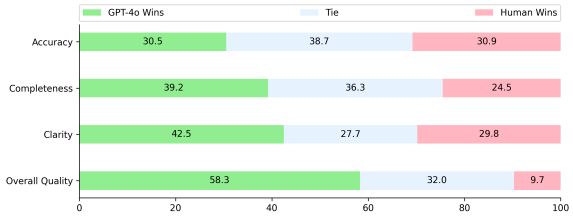


Figure 6: Comparison between the LLM-generated and human-annotated definition.

We evaluate the precision of the generated definition by comparing the performance between GPT-4o and human annotators. We recruited twelve master's students specializing in Arabic. We randomly selected 200 concepts lacking definitions. Six annotators were tasked with generating definitions for these concepts, following the style of existing Wikipedia concept definitions. Subsequently, following Alpaca-Eval (Li et al., 2023a), we performed head-to-head evaluations between LLM-generated and human-annotated definitions. Specifically, we employed a double-blind evaluation protocol where another group of six annotators were asked to compare and assess the quality of LLM-generated and human-annotated content. 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 19).

The evaluation results are shown in Fig 6, Our evaluation results demonstrate that GPT-4 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-4 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-4 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 ArabicKT, 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.

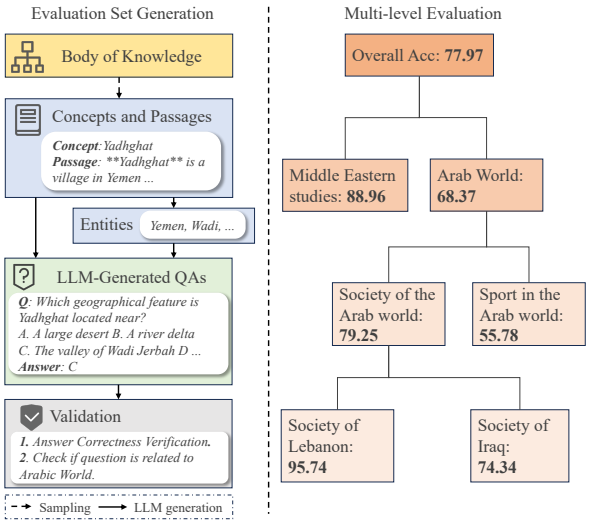


Figure 7: Evaluation workflow.

4.1 Evaluation Workflow

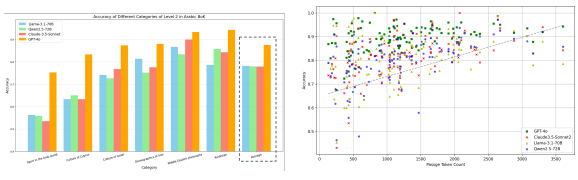
Using all articles from ArabicKT, the questions are automatically generated as shown in Fig 7. Following the construction of recent knowledge-based questions and benchmarks Yang et al. (2024b); Wang et al. (2023), we adapt their prompt and process to generate multiple-choice questions. Two types of questions 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, 118,381 evaluation questions are gathered. The prompts for generating questions, extracting entities, and revelation evaluation 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 popular open-sourced LLMs (Llama-3.1-80B-Instruct (Dubey et al., 2024) and Qwen-2.5-72B-Instruct (Yang et al., 2024a)). 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), including 3B, 7B, 14B, 32B, and 72B.

4.2 Evaluation for different LLMs

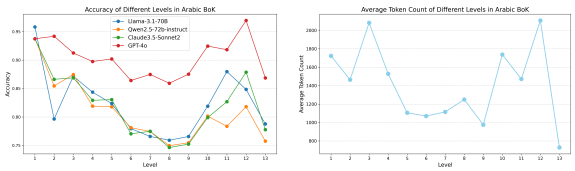


(a) (b)

Figure 8: (a) Accuracies on categories of level 2 in ArabicKT. (b) Relationship between category accuracy and average token length of the passages within the corresponding category.

Accuracy diverse on different categories but show consistency across different models. The results of four different series of LLMs are shown in Fig 8a. We randomly sample 6 categories

within Level 2 from different accuracy regions, as well as the averaged accuracy at right. Despite all the models achieving an overall accuracy exceeding 75%, they demonstrate notable performance degradation in specific knowledge domains. For instance, in the "*Culture of Saudi Arabia*" category, Llama-3.1-80B, Qwen-2.5-72B, and Claude-3.5-Sonnet2 exhibit accuracy rates below 50%, while GPT-4o's performance remains under 70%. In addition, we observed consistency in knowledge representation across different models. Categories such as "*Culture of Syria*", "*Culture of Egypt*", "*Islamic Studies*", and "*Israelites*" consistently achieve approximately 80% accuracy across all models. This consistency is further evidenced by the high correlation coefficient (0.7988) in category-wise accuracy between Qwen2.5-72B-Instruct and Claude3.5-Sonnet2.



(a) (b)

Figure 9: (a) Accuracies within different levels of ArabicKT. (b) Average token count for passages within different levels of ArabicKT.

Accuracy has a high correlation with the number of tokens appears in Wikipedia articles. For categories with low accuracy, we observe that they are mostly distributed on specific human, event, location, religion, and etc. Motivated by Allen-Zhu and Li (2023) that model requires large repetition of specific knowledge to learn it, we hypothesize these categories have: (1) insufficient contextual information within the articles and (2) limited availability of online resources. For this purpose, we first observe the relation between accuracy and token length at different levels. The token length is the summation of tokens for all passages at one level (tokenized using the Qwen2.5 tokenizer). Fig. 9b shows the comparison. It is obvious that a consistent trend between curves at left and right across all models. Moreover, we further examined the correlation between category accuracy and the average token length of passages within each category, as illustrated in Fig 8b. 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, lending support to our first hypothesis. Notably, we observed exceptions to this pattern: certain categories maintained high accuracy despite having relatively short passages. These exceptions predominantly occurred in geography and history-related categories (such as *Environment of Iran*" and *Environment of Egypt*"). While their corresponding passages were concise, the models' strong performance in these domains can be attributed to the rich abundance of related information available in online resources.

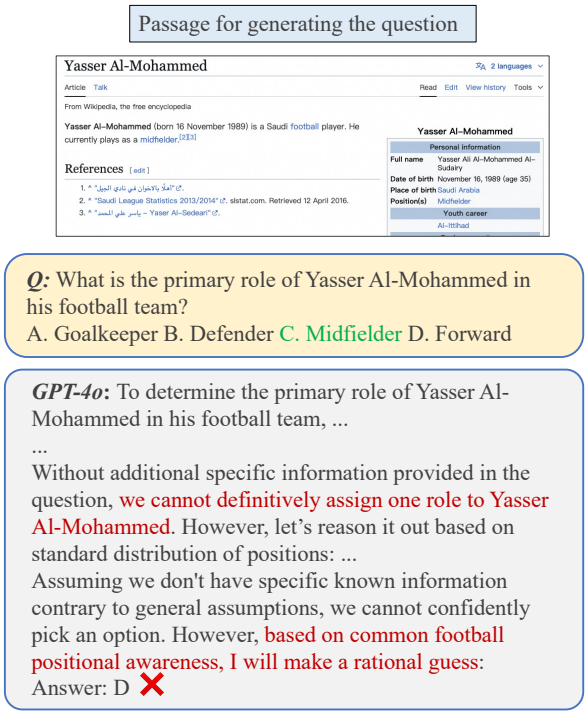
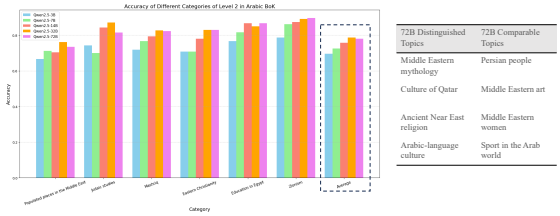


Figure 10: Example question generated from concept "*Yasser Al-Mohammed*" within category "*Culture of Saudi Arabia*" and failure answer from GPT-4o.

Further validation reveals that lower accuracy correlates to lower frequency in training corpus We further conducted analysis in model pretraining corpora. Specifically, we performed a comparative study by randomly sampling 100 concepts from two contrasting categories: a high-performing category ("*Israelites*") and a low-performing category ("*Culture of Saudi Arabia*"). By analyzing their frequency distribution in the Arabic-101 (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.

4.3 Evaluation for different size of LLMs



72B Distinguished Topics	72B 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 11: Accruracy of Qwen2.5 series models on categories of level 2 in ArabicKT.

Accuracy correlates to the size of the models Similarly, we compared the overall accuracy of Qwen2.5 series models of various sizes as well as their accuracy across different categories (Fig 11). On average, the accuracy increase as the the model size increase. This phenomenon is also consistent in most of categories that larger models will perform better.

Larger models are better at fine-grained categories. In certain categories, such as "*Zionism*" and "*Populated places in the Middle East*", models of different sizes achieved comparable accuracy levels. However, the performance disparities emerged in categories like *Middle Eastern mythology*" and *Culture of Qatar*", where larger models demonstrated markedly superior performance. Through manual analysis combining GPT-4 analysis, we identified distinguishing characteristics in the performance patterns of different-sized models. The 72B models exhibited notably superior performance in categories requiring sophisticated knowledge representation, particularly those involving nuanced historical contexts or specific cultural elements (such as Middle Eastern history, religion, and cultural practices). These domains demand enhanced capabilities in contextual understanding and knowledge integration. Conversely, in more general domains such as "*Geology of the Middle East*", the performance differential between models was minimal, suggesting that these categories involve more straightforward knowledge requirements that both large and small models can adequately process.

To further illustrate this distinction, we present

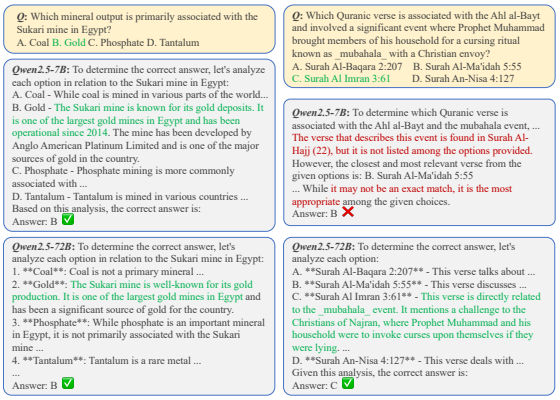


Figure 12: Responses of Qwen2.5-7B and Qwen2.5-72B for questions of different categories.

two representative cases (Fig 12). 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.

5 Conclusion and Discussions

In this work, we presented ArabicKT, 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 learn and retain knowledge about the Arab world. Our findings demonstrate that knowledge retention is strongly correlated with training data frequency, and model size impacts the ability to handle granular knowledge points.

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. Such applications could provide deeper insights into how LLMs process and utilize domain-specific knowledge, ultimately leading to more capable and interpretable models.

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 Arabic Knowledge Taxonomy (ArabicKT). 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

2024. Wikipedia: The free encyclopedia. Accessed January 2024.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.2, knowledge manipulation. *arXiv preprint arXiv:2309.14402*.

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.

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

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.

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)*.

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*.

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*.

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

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

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, ArabicMMLU (Koto et al., 2024) adapts the MMLU framework, comprising 40 tasks with 14,575 multiple-choice questions in Modern Standard Arabic, where even top-performing models achieve only 62.3% accuracy. The AlGhafa (Almazrouei et al., 2023) benchmark further enriches Arabic LLM evaluation resources, focusing on multiple-choice questions across various domains. However, these existing Arabic evaluation benchmarks share common limitations: they typically rely on manually crafted questions with limited

scale (usually thousands of samples) and may not comprehensively cover the full spectrum of Arabic knowledge and cultural domains.

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

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 13: 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 14: Example of a page in ArabicKT

In the ArabicKT knowledge system, there are two main types of nodes: pages (Fig. 14) and categories (Fig. 13). 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

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 15, 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.

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

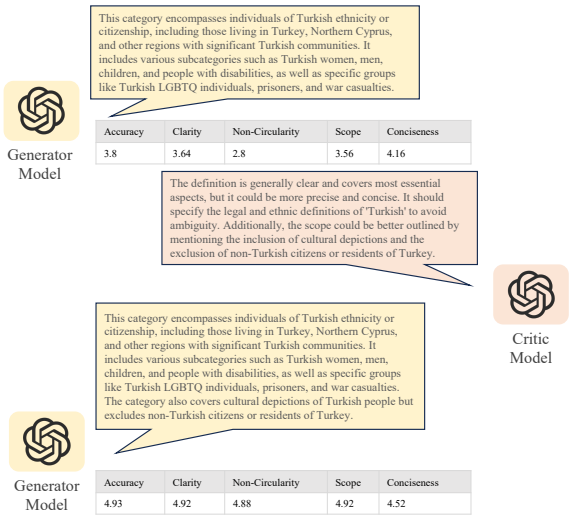


Figure 16: Example of multi-round definition completion.

A.4 Definition Completion

We implemented a dual-agent framework for iterative definition refinement, consisting of generator model for definition creation and a critique model for quality assessment. 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 17 and Figure 18, respectively. Our experimental results, as demonstrated in Figure 16, indicate that this multi-round refinement approach effectively enhances definition quality through iterative improvement.

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 24, 25, 26, and 27 respectively.

We showcased four example questions generated using different concepts from our Arabic BoK in Fig 20, 21, 22, and 23 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 28, which demands the model to first generate a chain of thoughts and then provide the answer in a specific format.

A.6 Evaluation 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 §**??**: Fig 29, 31, and 33 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 30, 32, and 34 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 Coverage Evaluation Details

We evaluate the coverage of ArabicKT by assessing how well ArabicKT 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. 5). First, we encode the knowledge point within each node in ArabicKT 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. The coverage score is defined as:

$$C(D) = \frac{|\{d \in D | \exists k \in K : I(d,k) = 1\}|}{|D|} \quad (3)$$

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. 5. The ArabicKT achieves coverage rates of 76.88% and 75.64% on training corpus, while achieves 82.34% and 81.13% on evaluation datasets. These substantial coverage rates, particularly on large-scale pre-training datasets containing millions of samples, validate the extensive breadth of our ArabicKT.

Similarly, to assess the Arabic knowledge coverage of existing evaluation datasets, we examine their knowledge coverage with respect to ArabicKT. This coverage is defined as the extent to which our ArabicKT nodes are covered by any sample in the evaluation datasets. Considering the hierarchical structure of our knowledge tree, we propose that if a node is covered, all its parent (category) nodes are considered covered as well. Formally, we define the coverage rate as:

$$C_{rev}(D) = \frac{|\{c \in C_{cat} | \exists d \in D : I(d,c) = 1\}|}{|C_{cat}|} \quad (4)$$

where C_{cat} represents all category nodes in our knowledge tree, and $I(d,c)$ indicates whether sample d covers category node c or its descendants.

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 17: 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 18: 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 19: 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 20: 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 21: 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 22: 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 23: 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 24: 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 25: 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 26: 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 27: 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 28: Question evaluation prompt following OpenAI (2024)

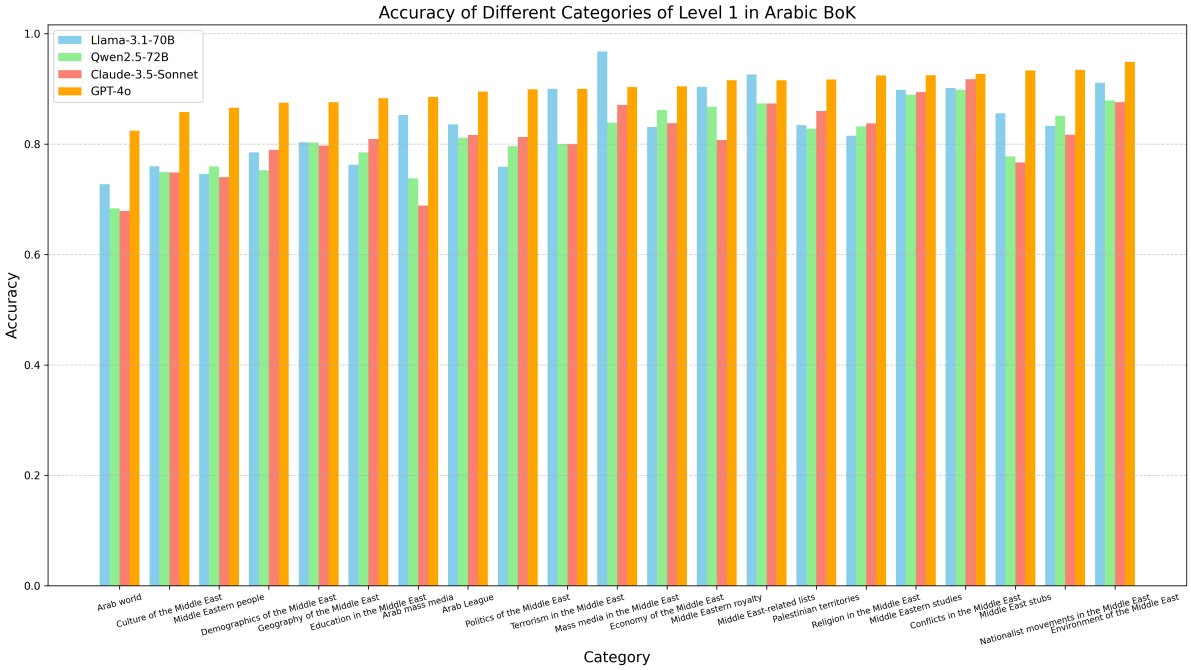


Figure 29: Accruracy on categories of level 1 in Arabic BoK for prevalent LLMs.

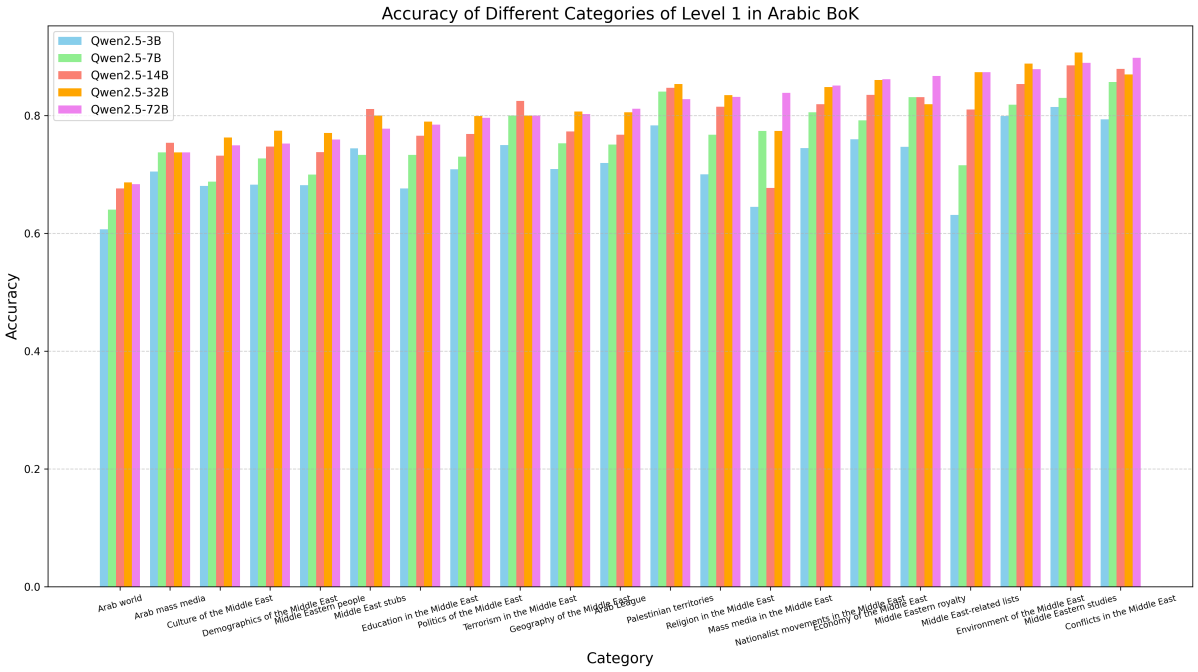


Figure 30: Accruracy on categories of level 1 in Arabic BoK for Qwen2.5 series models.

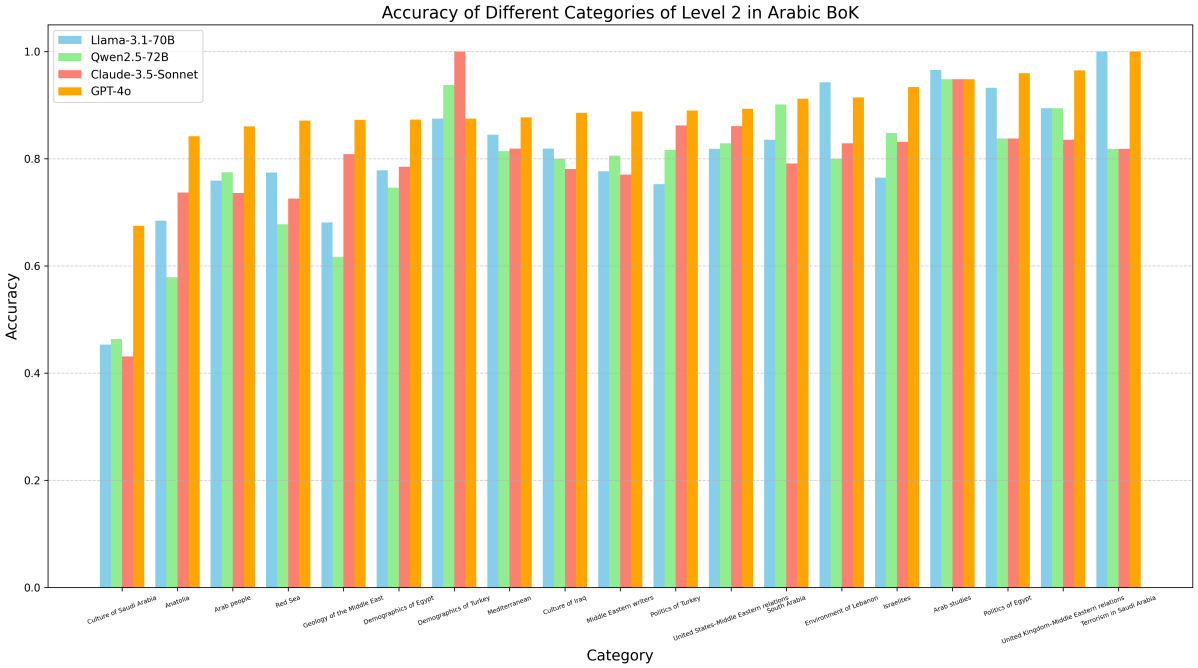


Figure 31: Accruracy on categories of level 2 in Arabic BoK for prevalent LLMs.

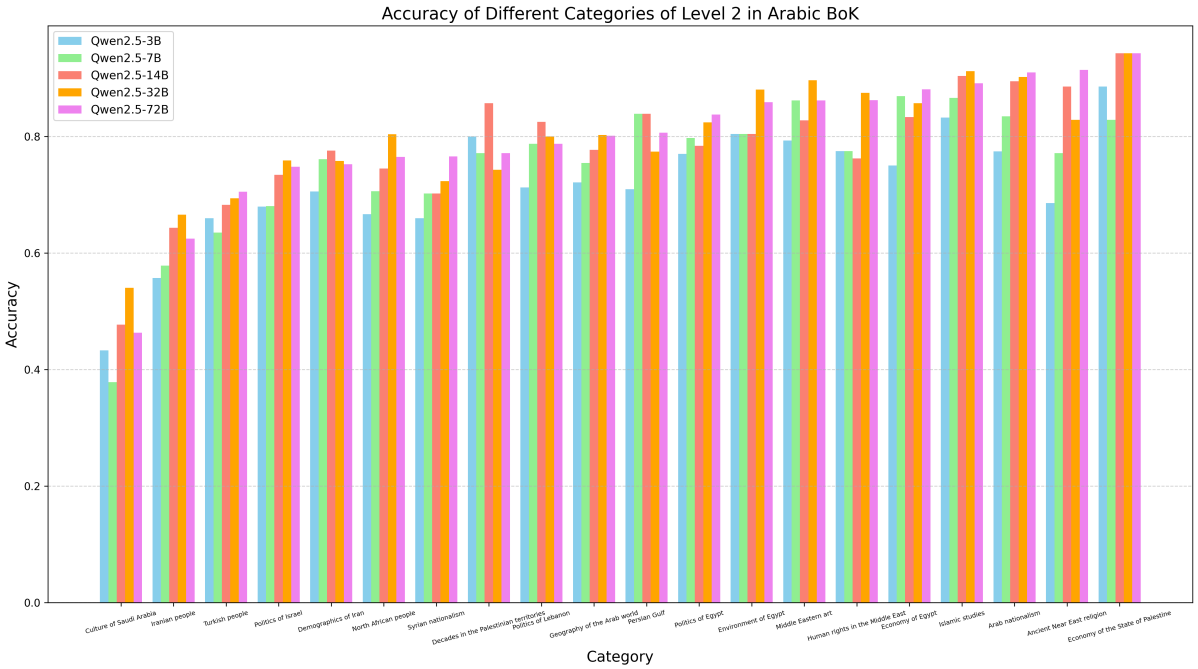


Figure 32: Accruracy on categories of level 2 in Arabic BoK for Qwen2.5 series models.

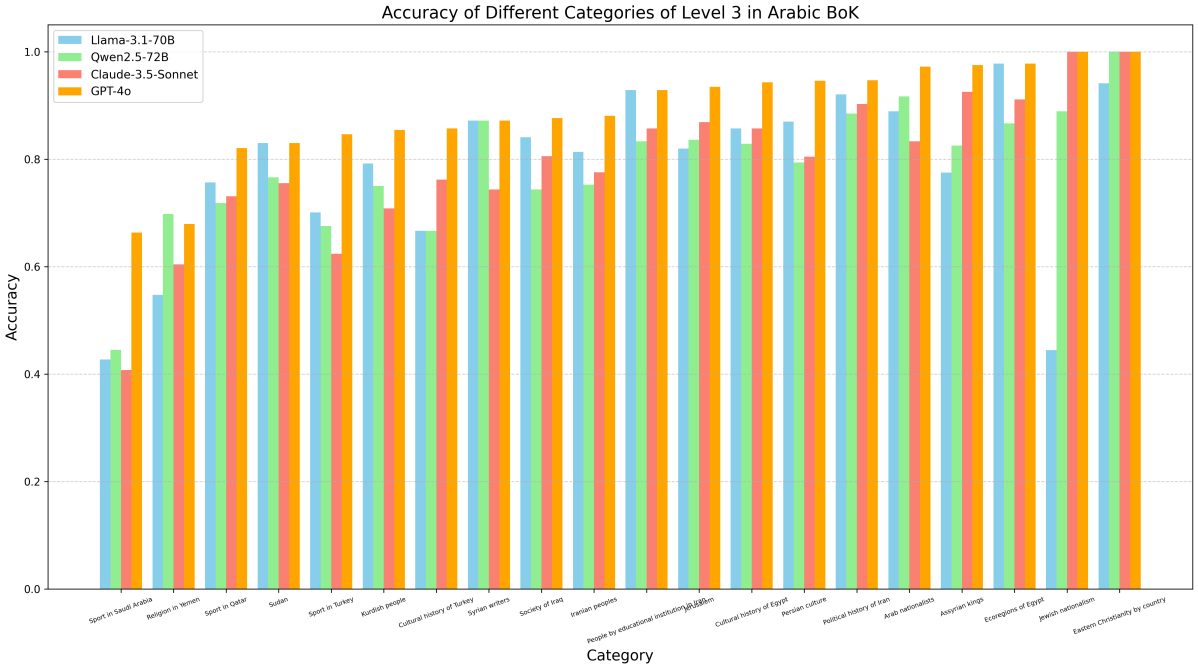


Figure 33: Accruracy on categories of level 3 in Arabic BoK for prevalent LLMs.

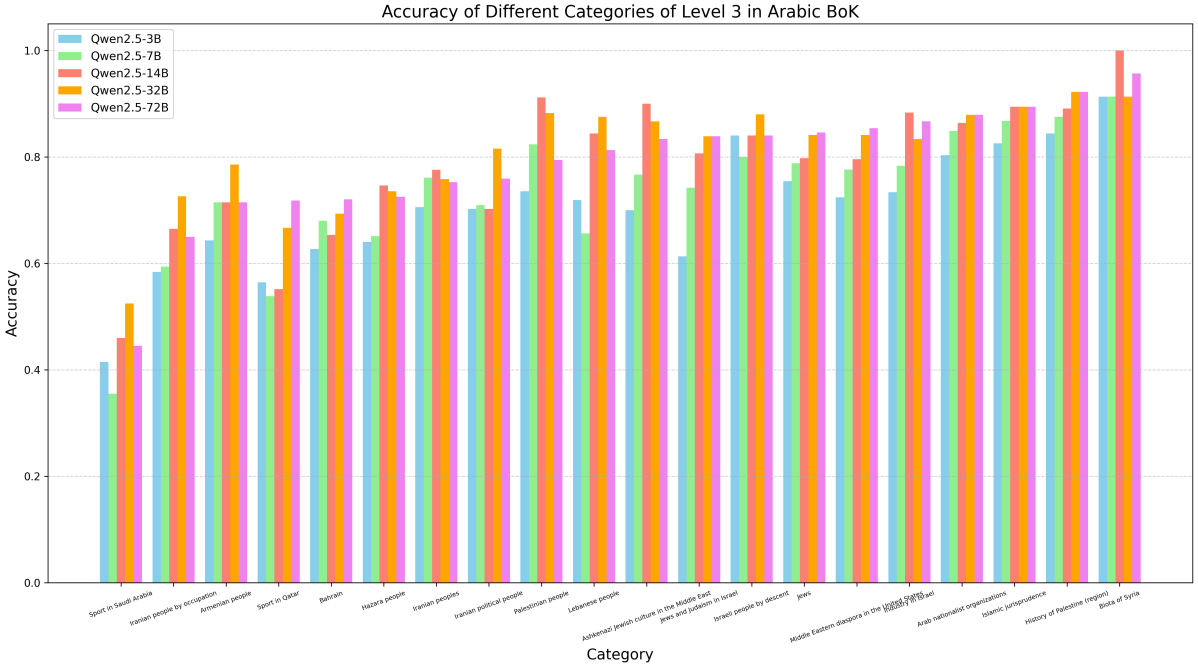


Figure 34: Accruracy on categories of level 3 in Arabic BoK for Qwen2.5 series models.