When Trivia Is Not Trivial: Everyday Knowledge Failures in Multilingual LLMs

Anna Mosolova **Djamé Seddah**
INRIA Paris
firstname.lastname@inria.fr

Abstract

Quiz rooms, trivia nights, and quiz shows challenge human knowledge across a wide range of topics, from canonical facts to everyday culture. In this paper, we examine whether large language models (LLMs) can perform competitively in such settings, using quiz-style questions to test them on both common and niche topics. We introduce TRIVIAROOMQA, a multilingual benchmark designed to evaluate everyday, culturally grounded, and long-tail knowledge across 288 topics. The benchmark contains 3,300 parallel multiple-choice questions in six European languages and additional 5,340 French-only questions for a more fine-grained case study. We evaluate 30 open-weight LLMs from European, Asian, and North American providers, covering models from 7 to 70B parameters. We find that models are strong on knowledge-intensive topics such as history, geography, and mathematics, but substantially weaker on everyday popular-culture topics such as celebrities, music, movies, and news. Moreover, model performance varies across languages even for the same underlying questions, suggesting that access to factual knowledge is not always language-independent. In sum, our dataset and experiments demonstrate an important knowledge gap which is not captured by existing academic-based saturated benchmarks.

1 Introduction

Current large language models (LLMs) perform strongly on academic and reasoning-oriented benchmarks, such as MMLU (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), and AIME.[1] However, strong performance on these benchmarks does not necessarily imply reliable knowledge of everyday facts, popular culture, contemporary events, or culturally grounded information.

Recent work has started to evaluate LLM knowledge beyond academic benchmarks, focusing on cultural knowledge, long-tail facts, and general factual recall. This line of work includes a large body of papers evaluating models' knowledge of social norms and traditions in different cultures (Myung et al., 2024; Fung et al., 2024; Chiu et al., 2025; Karmim et al., 2026; Lin et al., 2026).[2] This type of knowledge, together with specialized and temporal facts, is part of low-frequency long-tail knowledge (Badhe et al., 2026), for which new benchmarks are now being constructed (Sun et al., 2024; Zhu et al., 2026). For general knowledge evaluation, several trivia-like datasets have also been proposed for English, such as TriviaQA, Natural Questions, and SimpleQA (Joshi et al., 2017; Kwiatkowski et al., 2019; Wei et al., 2024). However, these benchmarks are limited to English and do not provide fine-grained topic annotations that would allow us to analyze which groups of everyday questions are most difficult for models.

In this paper, we extend this research direction by introducing TRIVIAROOMQA, a benchmark containing typical quiz-style questions on topics covered in everyday settings. The benchmark is divided into two parts: a parallel multilingual subset and a larger French-only subset. The multilingual part of the dataset consists of 3,300 parallel questions spanning **110 topics** and covering geographical, historical, sensory, pop-culture, and everyday-life knowledge. The questions are provided in **six languages** (English, French, Italian, Spanish, German, Dutch), enabling cross-lingual evaluation of the same underlying knowledge.

To further move beyond the English-centric paradigm, we also introduce a larger French-only evaluation set of 5,340 additional questions covering **178 additional topics**. In this work, the

[1] https://artificialanalysis.ai/evaluations/aime-2025

[2] In this paper, we use the word culture to refer to *the collective beliefs, values, customs, and practices that characterize a particular group or society* (UNESCO, 2026).

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 46100–46128
October 24-29, 2026 ©2026 Association for Computational Linguistics

multilingual part allows us to compare models' knowledge across languages. The French-only part provides a deeper case study: it expands the benchmark to 8,640 French questions and covers a wider range of topics, including both general knowledge and knowledge related to French-speaking countries. We focus on French because it is widely spoken, as it is the sixth most spoken language in the world with over 334M speakers worldwide (Ethnologue, 2026), while still often appearing as a secondary language after English in LLM training data, including in multilingual or French-oriented models such as Gaperon and Lucie (Godey et al., 2025; Gouvert et al., 2025).

TRIVIAROOMQA, unlike prior benchmarks, stems from quiz-like questions originally designed for human knowledge assessment in quiz rooms, TV shows, and online trivia quizzes.[3] Each topic in our dataset includes 30 questions about the same domain with three varying levels of difficulty. In addition to topic and difficulty annotations, we group topics into broader categories and annotate questions with the time period and geographic region to which they are most closely related. This structure allows us to both report average model performance and, more importantly, analyze how models behave across topics, categories, difficulty levels, languages, time periods, and geographic regions.

We evaluate 30 open-weight LLMs from providers in Europe, North America, and Asia, across a range of model sizes from 7 to 70B parameters. The evaluation is framed as multiple-choice question answering: models are given the question and answer candidates are scored using log-likelihood-based evaluation.

Our analysis shows that models perform strongly on topics typically covered in widely available encyclopedic and educational sources, such as scientific and historical knowledge, while being less reliable on long-tail and everyday culture-specific topics. For example, only 4 models out of 30 correctly answer the question *"What can be seen on the pin Katniss received from a merchant at the Hob?"*, while all models correctly answer *"What is the diameter of the huge golden sphere located in Auroville?"*. This shows that models can recall specific encyclopedic facts while still failing on popular-culture details that are common in quiz settings but may be less systematically represented in encyclopedic sources. While the general difficulty of long-tail knowledge has been observed in the literature, in this work, we characterize these gaps more precisely through a fine-grained topic-level analysis across languages, difficulty levels, time periods, and geographic regions.

Our analysis further shows that, while humans show gradual performance degradation with increasing difficulty within a topic, LLM performance drops across all difficulty levels once a topic is outside the model's knowledge boundary. Additionally, we demonstrate that LLMs perform better on questions related to older time periods, and that some models do not answer consistently when the same question is presented in different languages.

This paper makes the following contributions:

1. We introduce TRIVIAROOMQA, a human-written quiz-style benchmark for evaluating everyday, culturally grounded, and long-tail knowledge. It contains 3,300 parallel multiple-choice questions in six European languages and 5,340 additional French-only questions, annotated with 288 topics, 24 categories, 3 difficulty levels, time-period and geographic metadata.
2. We evaluate 30 open-weight LLMs from European, North American, and Asian providers, ranging from 7B to 70B parameters and show that models perform well on scientific and encyclopedic topics, but struggle with contemporary and popular-culture questions: accuracy ranges from 0.35 to 0.64 for models with fewer than 20B parameters and from 0.41 to 0.72 for larger models. We also find that some models answer inconsistently when the same question is presented in different languages.
3. Through a pilot human comparison study, we show that LLMs and humans behave differently on difficult topics: while human performance degrades gradually with increasing question difficulty, LLMs show similarly low performance across difficulty levels on topics where their overall accuracy is low.

2 Related work

Various benchmarks have been proposed to evaluate LLM knowledge, ranging from academic and commonsense reasoning to factual and cultural knowledge.

World knowledge Most widely used benchmarks evaluate expert, academic, or science-oriented knowledge, including MMLU (Hendrycks

[3] https://www.openquizzdb.org

Benchmark	Questions	Topics	Languages	Creation
SimpleQA	4,326	10	English	Human
PopQA	14,267	16	English	Automatic
TriviaQA	95K	-	English	Human
Natural Questions (test)	7,842	-	English	Human
Ours	3,300 parallel, 5,340 French	288	6 languages	Human

Table 1: Comparison between existing trivia-style and general-knowledge benchmarks and our TRIVIAROOMQA.

et al., 2021), MMLU-Pro (Wang et al., 2024), BIG-Bench (Srivastava et al., 2023), and Humanity's Last Exam (Phan et al., 2026). Other benchmarks use school or university exam questions, such as AGIEval (Zhong et al., 2024) or MERA (Fenogenova et al., 2024), or focus on commonsense understanding, e.g., HellaSwag (Zellers et al., 2019) and PIQA (Bisk et al., 2020). While these benchmarks are useful for academic abilities, they do not target everyday knowledge.

Trivia-style factual question answering Several benchmarks evaluate factual knowledge in quiz-like settings. Joshi et al. (2017) introduce TriviaQA, a reading-comprehension dataset built from trivia questions. Natural Questions dataset (Kwiatkowski et al., 2019) evaluates models' factual knowledge using questions based on real search-engine queries, while PopQA (Mallen et al., 2023) uses questions generated from Wikidata triples. Lastly, SimpleQA (Wei et al., 2024) evaluates models' everyday knowledge on 10 topics using human-written questions. These datasets are close to TRIVIAROOMQA as they do away with academic evaluation and concentrate on factual knowledge, however, they are primarily English-centric and are not designed for fine-grained analysis of model's knowledge by topic. In Table 1, we compare TRIVIAROOMQA with these benchmarks.

Long-tail and cultural knowledge Several benchmarks evaluate models on long-tail, low-frequency, or domain-specific knowledge. For example, Head-to-Tail (Sun et al., 2024) studies model knowledge across most and least frequent entities from knowledge graphs, while LPFQA (Zhu et al., 2026) tests models on long-tail professional knowledge from forums. TRIVIAROOMQA is related to this direction, as it includes questions about entities and events that may be rare in pre-training data, such as local news or region-specific references.

Cultural knowledge can also be viewed as a domain-specific evaluation setting (Badhe et al., 2026). Culture-oriented benchmarks primarily evaluate models' understanding of social practices, customs, values, and norms associated with different cultures. Such benchmarks include CulturalBench (Chiu et al., 2025), BLEND (Myung et al., 2024), CultureAtlas (Fung et al., 2024), and LatamQA (Karmim et al., 2026). At the same time, several recent surveys on cultural evaluation (Adilazuarda et al., 2024; Zhou et al., 2025; Kabir et al., 2025; Oh et al., 2025; Alkhamissi et al., 2026) highlight that such approaches may be limited, as culture is often reduced to static facts about traditions, social preferences, or biases, commonly presented in multiple-choice format. In TRIVIAROOMQA, instead of evaluating models' understanding of cultural norms or alignment with a particular worldview, we test whether models possess everyday factual knowledge about movies, songs, events, and other topics that people may encounter through media, education, entertainment, and public life, and that together form part of everyday cultural knowledge.

3 TRIVIAROOMQA

TRIVIAROOMQA is a new dataset derived from quizzes freely available on the OpenQuizzDB platform[4], originally designed for human quiz settings. The benchmark consists of 3,300 parallel questions in French, English, Italian, Spanish, German, and Dutch, as well as 5,340 additional French-only questions (in total, 8,640 questions in French). Each question has four answer options, a short explanatory note[5], and one of the three difficulty levels: beginner, confirmed, or expert. Examples of benchmark questions across different topics and difficulty levels are provided in Appendix F.

We define two evaluation settings: TRIVIAROOMQA-MULTI, with 3,300 parallel questions in six languages (19,800 language-specific examples), and TRIVIAROOMQA-FRENCH, containing 8,640 French-only questions.

The multilingual part covers 110 topics, while the full French collection covers 288 topics. We

[4] https://www.openquizzdb.org, available under the CC BY-SA 4.0 license. TRIVIAROOMQA is released under the same license and is available at: https://github.com/anya-bel/triviaroomqa/.

[5] The explanatory note is not shown to models during evaluation, but it can be used in future experiments on model enhancement for this benchmark.

TRIVIAROOMQA-FRENCH	
# of questions	8,640
# of topics	288
# of categories	24
TRIVIAROOMQA-MULTI	
# of questions	3,300
# of topics	110
# of categories	21
Annotation	Topics, Categories, Level, Continent⋆, Time period⋆

Table 2: Statistics and annotation information for the TRIVIAROOMQA dataset. Annotations marked with ⋆ were produced automatically using Llama-70B-Instruct.

group the questions into 24 broad categories, such as history, geography, food, entertainment, music, news, sport, people, objects, tech, etc. Statistics on the number of questions and topics in each category are given in Table 5, and the full list of categories and corresponding topics is provided in Table 15 in Appendix A. General dataset statistics are provided in Table 2.

To estimate topic rarity, we used Infini-Gram (Liu et al., 2024)[6] to check the frequency of topic key words in the OLMo-3-7B-Instruct training data.[7] We found that 21% of the topics occur fewer than 100k times in its 6T-token corpus, suggesting that at least some topics in the benchmark are relatively infrequent in this particular corpus. We also checked whether OpenQuizzDB appears in FineWeb2 (Penedo et al., 2025): the full 20TB corpus contains only 214 mentions of the website, suggesting that OpenQuizzDB is not substantially represented in FineWeb2 and may be limited in CommonCrawl-derived training corpora. Another indication that the benchmark questions are not sufficiently represented in model training data is the relatively moderate performance observed across models in our experiments (Section 5), although this does not provide conclusive evidence for all questions and models.

Finally, we use Llama-70B-Instruct to assign approximate geographic and temporal metadata to each question, when applicable.[8] Specifically, we annotate questions with the continent and time period to which they are most closely related. These labels are intended for high-level analysis rather than as gold-standard annotations: they allow us to study broad geographic and temporal trends in model performance, but individual labels may contain errors. Statistics for these annotations are provided in Tables 3 and 4 in Appendix A. Human validation of a sample of the annotations is provided in Appendix C.

4 Experimental setup

We evaluate TRIVIAROOMQA using the Language Model Evaluation Harness library (lm-eval) (Gao et al., 2024)[9]. We implement the benchmark as two tasks: (i) TRIVIAROOMQA-MULTI, corresponding to the 3,300 parallel questions in six languages (19,800 language-specific examples), and (ii) TRIVIAROOMQA-FRENCH, corresponding to the full French collection of 8,640 questions. Both tasks are evaluated as multiple-choice question answering. For each example, the question is used as the prompt, and the four answer options are scored using conditional log-likelihood, following standard lm-eval protocol. The option with the highest score is selected as the model prediction, and we report accuracy. This evaluation avoids free-form answer generation and answer-normalization issues, and allows direct comparison across models and languages. Since each question has four answer options, the random baseline is 0.25.

All models are evaluated without a chat template in the main experiments, since chat templates were not available for all models. Results with chat templates are provided in Appendix L.

4.1 Models

We evaluate 30 open-weight LLMs spanning multiple scales, languages, geographic origins, and training dataset sizes. The models range from 7B to 70B parameters: 19 models have fewer than 20B parameters, and 11 models have 20B parameters or more[10]. Below, we provide the list of models used for the evaluation:

European models: EuroLLM-9B-Instruct and EuroLLM-22B-Instruct, Sabiá-7B, ALIA-40B-Instruct, Salamandra-7B-Instruct, Lucie-7B-Instruct, Ministral-8B-Instruct, Gaperon-8B-SFT and Gaperon-24B-SFT, Minerva-7B-Instruct, Velvet-14B, Apertus-8B-Instruct and Apertus-70B-Instruct, TildeOpen-30B, LLaMmlein-7B, and YandexGPT-5-Lite-8B-Instruct.

North American models: Granite-4.1-8B and Granite-4.1-30B, Gemma-3-12B-IT and Gemma-3-

[6] https://infini-gram.io

[7] We use this model as its training data is directly accessible through Infini-Gram.

[8] The exact prompts for the continent and time-period annotation tasks are provided in Appendix B.

[9] Experimental setup details are provided in Appendix O.

[10] All models were accessed using the Transformers library (Wolf et al., 2020).

27B-IT, OLMo-3-7B-Instruct and OLMo-3.1-32B-Instruct, Llama-3.1-8B and Llama-3.1-70B, as well as Aya Expanse-8B and Aya Expanse-32B.

Asian models: ALLaM-7B-Instruct, LLM-JP-4-8B-Instruct, Qwen3.5-9B, and Qwen3.5-27B.

Tables 18 and 19 in the Appendix summarize the model information, i.e., training data scale, country of origin, language coverage, citations, and links.

In Appendix E, we additionally provide a preliminary study of Llama-3.1-8B and Qwen3.5-9B with access to a search tool. This experiment estimates the possible gains from internet access when answering quiz questions.

5 Results and Discussion

We analyze model performance in the two TRIVIAROOMQA settings. First, we use TRIVIAROOMQA-MULTI to test whether models answer the same questions consistently across languages. We then use the larger TRIVIAROOMQA-FRENCH subset for fine-grained analyses by category, difficulty, time period, and geographic region. Together, these analyses allow us to separate three sources of variation: the language of the question, the type of knowledge being tested, and the metadata associated with each question.

Throughout the section, we also compare models below and above 20B parameters to assess whether scale mitigates the observed weaknesses.

The uncertainty estimates for the main analyses are provided in Appendix J.

5.1 Language: cross-lingual consistency remains uneven

Figure 1 presents the performance of 30 models on the six languages of TRIVIAROOMQA-MULTI. The main observation is that models do not always answer the same questions equally well across languages. This is true both for models that were not trained on some of the languages in TRIVIAROOMQA-MULTI, and for models whose training data includes these languages. This suggests that, for some models, factual knowledge is not accessed in a fully language-independent way, as also shown in prior work on multilingual factual knowledge (Ifergan et al., 2025). A more detailed analysis of cross-lingual answer consistency is provided in Appendix G.

Performance varies substantially across languages and models. OLMo-3-7B obtains the lowest overall performance, which is expected given

	English	French	Italian	German	Spanish	Dutch
Aya-32B	0.67★	0.69★	0.62★	0.62★	0.61★	0.61★
Apertus-70B ▲	0.63★	0.67★	0.60★	0.62★	0.62★	0.60★
Gemma-3-27B ▲	0.63★	0.66★	0.60★	0.59★	0.59★	0.59★
ALIA-40B	0.61★	0.65	0.59	0.59	0.59★	0.59
Llama-3-70B ▲	0.63★	0.66★	0.58★	0.58★	0.58★	0.58
Salamandra-7B ▲	0.59★	0.62★	0.57★	0.55★	0.55★	0.55★
Apertus-8B ▲	0.57★	0.58★	0.55★	0.55★	0.56★	0.54★
Ministral-8B	0.60	0.62	0.55	0.53	0.54	0.50
Gemma-3-12B ▲	0.59★	0.61★	0.53★	0.54★	0.52★	0.53★
Aya-8B	0.58★	0.60★	0.54★	0.54★	0.54★	0.53★
Velvet-14B	0.56★	0.57★	0.54★	0.51★	0.53★	0.43
EuroLLM-22B	0.51★	0.55★	0.49★	0.47★	0.50★	0.46★
Qwen-3.5-27B	0.50★	0.50★	0.46★	0.46★	0.45★	0.45★
Llama-3-8B ▲	0.52★	0.49★	0.46★	0.45★	0.44★	0.45
Granite-30B ▲	0.53★	0.49★	0.43★	0.44★	0.44★	0.44★
EuroLLM-9B	0.46★	0.47★	0.43★	0.42★	0.45★	0.43★
ALLaM-7B	0.51★	0.46	0.43	0.42	0.40	0.41
YandexGPT-8B ▲	0.49★	0.43	0.40	0.40	0.42	0.41
Gaperon-8B	0.44★	0.49★	0.40	0.40	0.38	0.41
Lucie-7B	0.45★	0.47★	0.40★	0.40★	0.42★	0.33
TildeOpen-30B	0.40★	0.44★	0.41★	0.41★	0.40★	0.41★
Qwen-3.5-9B	0.43★	0.43★	0.40★	0.40★	0.39★	0.39★
Gaperon-24B	0.45★	0.47★	0.39	0.38	0.38	0.38
Granite-8B ▲	0.46★	0.40★	0.38★	0.39★	0.39★	0.39★
Sabia-7B ▲	0.45★	0.41	0.38	0.38	0.40	0.38
LLaMmlein 7B	0.43	0.39	0.36	0.46★	0.36	0.37
Minerva-7B	0.45★	0.38	0.44★	0.34	0.37	0.34
Olmo-3.1-32B	0.45★	0.39	0.37	0.37	0.33	0.36
LLM-JP-8B ▲	0.42★	0.37	0.35	0.36	0.36	0.36
Olmo-3-7B	0.38★	0.34	0.32	0.32	0.32	0.32
AVG	0.51	0.51	0.47	0.46	0.46	0.45

Figure 1: Model performance on the six languages of TRIVIAROOMQA-MULTI. The reported metric is accuracy; greener cells indicate higher scores. ⋆ next to a score indicates that the model was trained on that language. ▲ marks models trained on 10T+ tokens. Model size is color-coded: ≤10B, 12-27B, 30B+.

its English-focused training setup. Other models, such as Granite-8B and Llama-3.1-8B, also drop by around 0.07 accuracy points in languages other than English, despite supporting multiple languages. Larger models improve absolute accuracy, but do not fully solve cross-lingual inconsistency: several 20B+ models still show noticeable variation across languages.

Highly multilingual models such as Salamandra-7B, Apertus-8B/70B, Gemma-3-12B/27B, and Aya-8B/32B show strong performance on all languages. Models trained specifically on one of the non-English benchmark languages, such as Gaperon-8B and Lucie-7B in French, LLaMmlein-7B in German, and Minerva-8B in Italian, often achieve comparable or better results in that language than in English. In contrast, EuroLLM and Qwen models show more uneven performance across languages in our benchmark, suggesting that cross-lingual consistency depends not only on model origin, but also on multilingual training, data scale, and model

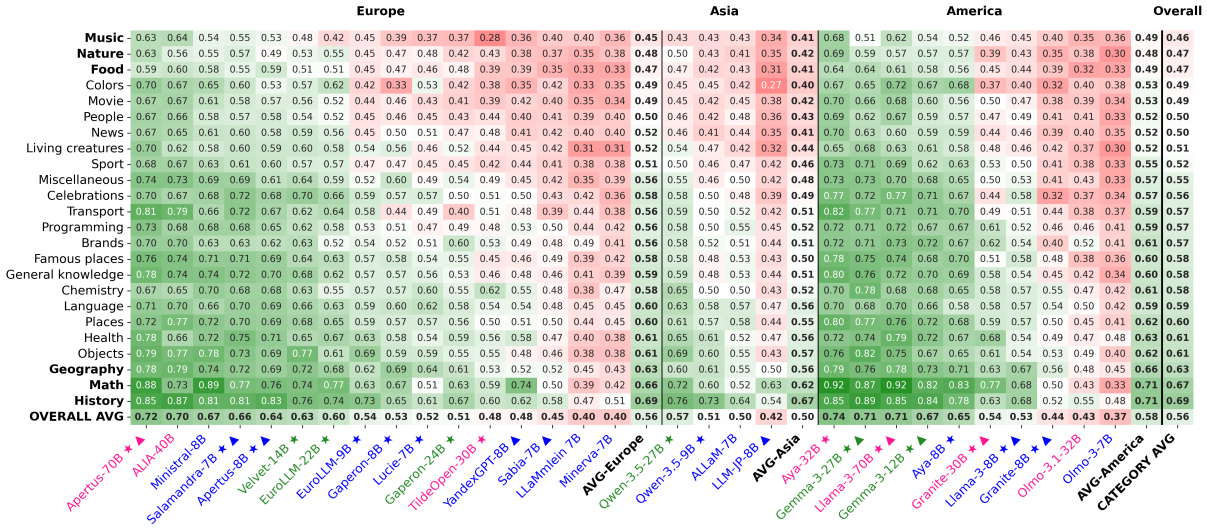

Figure 2: Model performance on **24 categories** from the TRIVIAROOMQA-FRENCH. The reported metric is accuracy; greener cells indicate higher scores. Models are sorted horizontally from left to right by performance within each region, and categories are sorted vertically from top to bottom by average performance across all models. ⋆ next to a score indicates that the model was trained on French. ▲ marks models trained on 10T+ tokens. Model size is color-coded: ≤10B, 12-27B, 30B+. The category-average columns and the three highest- and lowest-performing category labels are highlighted to facilitate comparison.

family.

We next move to the larger French subset, where we examine whether the remaining variation is explained by topic, difficulty, time period, or geographic region.

5.2 Category: encyclopedic knowledge is easier than popular culture

Figure 2 presents results for 30 models on 24 categories from TRIVIAROOMQA-FRENCH. Performance varies substantially across categories: widely documented encyclopedic categories, such as history and geography, obtain the highest average scores, while everyday and popular-culture categories, such as music, movies, people, and news, are more difficult.

This gap is visible even for strong models. For example, Ministral-8B drops by 0.27 accuracy points from History to Music, and Aya-Expanse-32B drops by 0.17. In total, music, movies, people, and news remain below 50% accuracy for 18 of the 30 models. This suggests that everyday cultural knowledge remains challenging for LLMs, possibly because such facts are less stable, less canonical, or less frequently repeated in standard factual sources than historical or geographic facts.

Models trained on more than 12T tokens, such as Salamandra, Gemma, Apertus, and Llama, achieve higher performance across many categories, and larger models such as Apertus-70B, Aya-Expanse-32B, and Granite-4.1-30B often improve on harder categories compared to their smaller counterparts. However, this pattern is not uniform across model families. In sum, scale improves absolute accuracy, but does not remove the gap between encyclopedic and everyday cultural categories.

Additional analyses in Appendices H and I further examine category-level differences, including their relationship with temporal variation and the topics and individual questions on which models most frequently fail.

5.3 Difficulty level: humans and LLMs show different patterns on unfamiliar topics

On TRIVIAROOMQA-FRENCH, model performance decreases with the annotated difficulty level: average accuracy is 0.60 on beginner questions, 0.52 on confirmed questions, and 0.48 on expert questions, with an average of 0.53. This shows that the difficulty annotations are reflected in model performance, while also confirming that the benchmark remains challenging overall. Per-model results are provided in Figure 7 in the Appendix.

To further examine whether models and humans react similarly to difficulty, we selected the six easiest and six hardest topics for models, according to their average performance, and asked five French

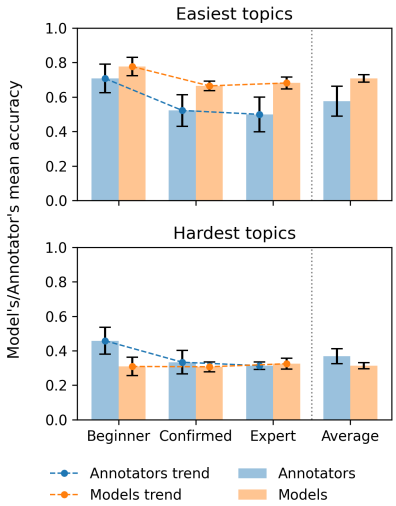

Figure 3: Average human and model performance on the six easiest and six hardest topics for models in TRIVIAROOMQA-FRENCH, grouped by question difficulty.

	timeless (1742 q.)	historical (1652 q.)	1950s-1960s (498 q.)	1970s-1980s (962 q.)	1990s-2000s (1990 q.)	2010s-2020s (1796 q.)
Aya-32B	0.70	0.79	0.76	0.70	0.71	0.70
Apertus-70B ▲	0.68	0.77	0.74	0.67	0.69	0.67
Llama-3-70B ▲	0.67	0.77	0.73	0.66	0.69	0.62
ALIA-40B	0.65	0.77	0.71	0.66	0.68	0.65
Gemma-3-27B ▲	0.70	0.78	0.70	0.63	0.66	0.65
Ministral-8B	0.65	0.73	0.66	0.60	0.62	0.60
Salamandra-7B ▲	0.63	0.73	0.66	0.59	0.60	0.59
Gemma-3-12B ▲	0.65	0.71	0.65	0.59	0.62	0.60
Apertus-8B ▲	0.61	0.73	0.65	0.59	0.59	0.56
Aya-8B	0.62	0.69	0.62	0.56	0.60	0.59
Velvet-14B	0.61	0.68	0.62	0.57	0.56	0.58
EuroLLM-22B	0.59	0.65	0.61	0.53	0.54	0.53
Granite-30B ▲	0.53	0.58	0.53	0.49	0.51	0.45
Qwen-3.5-27B	0.55	0.65	0.54	0.46	0.47	0.47
Llama-3-8B ▲	0.49	0.59	0.53	0.50	0.50	0.47
Gaperon-8B	0.55	0.59	0.55	0.46	0.48	0.46
EuroLLM-9B	0.53	0.62	0.54	0.45	0.48	0.43
ALLaM-7B	0.47	0.55	0.52	0.48	0.49	0.43
Lucie-7B	0.51	0.58	0.49	0.44	0.46	0.46
Gaperon-24B	0.51	0.57	0.52	0.44	0.44	0.45
Qwen-3.5-9B	0.49	0.56	0.50	0.45	0.43	0.41
TildeOpen-30B	0.47	0.50	0.46	0.41	0.44	0.44
YandexGPT-8B ▲	0.46	0.52	0.45	0.41	0.44	0.42
Sabia-7B ▲	0.41	0.49	0.46	0.41	0.43	0.41
Granite-8B ▲	0.44	0.49	0.43	0.42	0.41	0.38
Olmo-3.1-32B	0.40	0.45	0.44	0.41	0.41	0.38
Minerva-7B	0.36	0.42	0.42	0.40	0.38	0.38
LLM-JP-8B ▲	0.39	0.44	0.43	0.39	0.39	0.35
LLaMmlein 7B	0.36	0.43	0.39	0.39	0.39	0.38
Olmo-3-7B	0.37	0.38	0.39	0.34	0.35	0.33
AVG	0.54	0.61	0.56	0.50	0.52	0.49

Figure 4: Model performance on TRIVIAROOMQA-FRENCH grouped by the **time period** to which each question is most closely related, when applicable. The reported metric is accuracy, greener cells indicate higher scores. ▲ marks models trained on 10T+ tokens. Model size is color-coded: ≤10B, 12-27B, 30B+. The full results are provided in Figure 10 in the Appendix.

speakers to answer the corresponding questions.[11] Figure 3 compares the average performance of humans and models on these two groups of topics across difficulty levels.

On the easiest topics, LLMs outperform humans on average, but the situation is reversed on the hardest topics, where humans achieve higher scores. More importantly, humans show a gradual decrease in performance from beginner to expert questions, while models, once they are unfamiliar with a topic, remain close to the random baseline across all difficulty levels. This suggests that the difficulty scale for models is not the same as for humans, with model scale potentially related to the density of topic-related information in models' training data.

5.4 Time period: historical and timeless questions are easier for models

Figure 4 shows model performance on TRIVIAROOMQA-FRENCH grouped by the time period most closely associated with each question, including historical questions, recent questions, and timeless questions, i.e., questions that do not depend on a specific temporal frame.

Across models, historical and timeless questions obtain the highest scores, while performance decreases for more recent periods, especially for questions from the 2000-2020s. This decrease is less pronounced for several larger models, which maintain relatively strong performance on recent questions, nevertheless the temporal gap remains visible. One possible explanation is that older and timeless facts are more likely to appear repeatedly in stable sources such as encyclopedias, educational materials, and web pages. Recent questions, by contrast, often concern news or media events that may be less consistently represented in pre-training data. This trend is aligned with the category results, since recent questions include popular culture and news topics, which we showed in Section 5.2 to be among the lowest-performing categories.

5.5 Continent: performance depends on model origin and question mix

Figure 5 reports model performance on questions grouped by the continent to which they are most closely related. For models below 20B parameters, average performance by model origin broadly aligns with the question region: European models perform best on Europe-related questions, Asian models on Asia-related questions, and North American models on North America-related questions.

[11] The selected topics and participant information are provided in Appendix D.

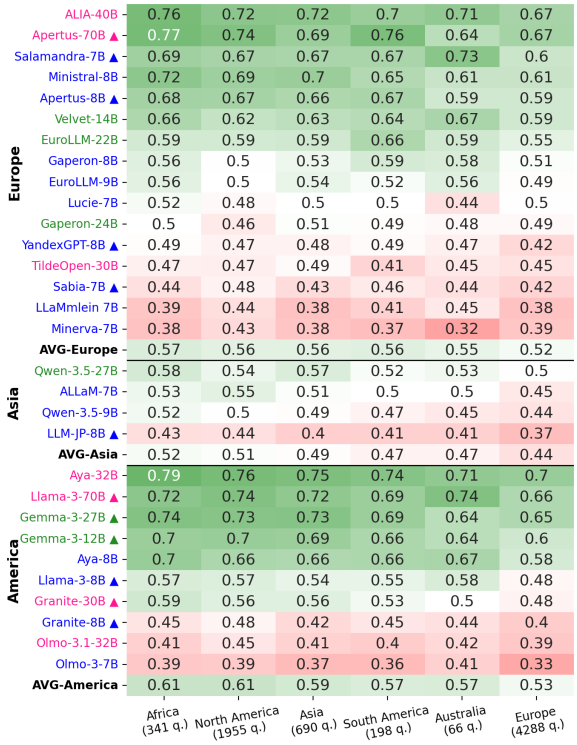

Figure 5: Model performance on TRIVIAROOMQA-FRENCH, grouped by the **continent** to which each question is most closely related, when applicable. Models are grouped by continent of origin, with the average performance for each group reported in the corresponding **AVG-{continent}** row. The reported metric is accuracy; greener cells indicate higher scores. ▲ marks models trained on 10T+ tokens. Model size is color-coded: ≤10B, 12-27B, 30B+. The full results are provided in Figure 11 in the Appendix.

This trend becomes less clear when larger models are included, since most 20B+ models in our evaluation come from North America and also have higher overall accuracy. This suggests that origin alone cannot explain the continent-level patterns.

Among individual models, EuroLLM-22B and Qwen3.5-27B show more uneven regional performance than several similarly sized multilingual models. For EuroLLM, this may be related to its smaller reported training corpus, while for Qwen the results suggest weaker performance on this particular everyday-knowledge setting.

Regional score differences appear to be driven by several factors. North America-related questions obtain high scores across most models, consistent with prior work reporting stronger model performance on North American cultural knowledge (Chiu et al., 2025). This tendency is visible even for LLaMmlein-7B, a German-only model trained from scratch, suggesting that the effect cannot be attributed only to English-language training data. Rather, it may reflect the broad coverage of U.S.-related information in web data, multilingual media, and encyclopedic sources. Similarly, the relatively strong scores on Africa-related questions can be partially explained by the topics covered in this subset, which include well-known encyclopedic topics such as Ancient Egypt and famous African cities. In contrast, Europe-related questions include more region-specific and long-tail topics, such as French cheeses, long-forgotten comedians, or Belgian beers. Thus, continent-level performance reflects both geographic knowledge coverage and the difficulty and topic distribution of the questions in each region.

5.6 Effect of model size: scale helps but does not close the gaps

The previous subsections show that larger models generally improve absolute accuracy, but do not eliminate the main performance gaps revealed by TRIVIAROOMQA. In particular, models with more than 20B parameters often rank among the strongest systems, but they still show cross-lingual inconsistency, lower performance on popular-culture and everyday categories, and sensitivity to the temporal and geographic composition of the questions.

6 Conclusion

In this paper, we introduce TRIVIAROOMQA, a multilingual benchmark for evaluating LLMs in settings inspired by quiz games.

Our results show that trivia remains challenging for open-weight LLMs. Models perform well on encyclopedic categories such as history and geography, but struggle with popular culture and everyday categories, such as music, movies, and news. They also answer the same questions inconsistently across languages, suggesting that factual knowledge is not always accessed in a language-independent way, as also shown in prior work (Ifergan et al., 2025). Larger models generally improve absolute accuracy, but do not eliminate these gaps: models above 20B parameters still show cross-lingual inconsistency and weaker performance on everyday cultural knowledge.

Our benchmark also reveals that what models find difficult does not always align with what humans find difficult. In our pilot human comparison, human performance decreases gradually as ques-

tion difficulty increases, while model performance remains close to the random baseline across difficulty levels on topics where models perform poorly overall. We also find that historical and timeless questions are easier on average than recent questions, and that North America-related questions obtain high scores across most models, although continent-level results partly reflect the topic composition of each regional subset.

Finally, a preliminary experiment with search-augmented models suggests that some errors come from information that is missing from, or difficult to retrieve from, model parameters alone, since tested models improve when given access to external search.

In summary, TRIVIAROOMQA provides a fine-grained benchmark for identifying gaps in the factual knowledge stored by LLMs and in the way this knowledge is accessed across languages, topics, time periods, and geographic regions. Future work can use the benchmark to evaluate retrieval-augmented or tool-using systems, where models have access to external information rather than relying only on parametric knowledge.

Limitations

Despite the broad coverage of TRIVIAROOMQA, this work has several limitations. First, the multilingual part of the benchmark currently covers only six European languages, and the fine-grained analysis relies mainly on the French subset. Second, we evaluate only open-weight models, so the results may not reflect the behavior of closed commercial LLMs. Third, the continent and time-period annotations are produced automatically and should be interpreted as approximate labels rather than gold annotations. Fourth, although OpenQuizzDB states that its questions are written by contributors following the platform's guidelines and reviewed by project members before publication, we do not independently verify the factual validity of every question. Finally, our human comparison study involves only five participants and should be interpreted as a pilot analysis rather than a definitive estimate of human quiz performance.

7 Ethical considerations

TRIVIAROOMQA is derived from OpenQuizzDB, which is distributed under a CC BY-SA 4.0 license. We follow the license terms and preserve attribution to the original source. The benchmark contains trivia-style questions about public factual knowledge and does not require collecting personal information. Some topics in the dataset include questions related to alcohol and sexuality. This benchmark is intended for diagnostic research on LLM factual knowledge, and its scores should not be interpreted as a general measure of model intelligence or cultural competence.

The human comparison study involved five French-speaking participants. Participants were informed about the purpose of the study, answered voluntarily, and only average accuracy statistics are reported. We do not release individual-level responses.

Metadata annotations (continent and time period information) are produced automatically with Llama-70B-Instruct, therefore, they may contain errors and should be used only for trend-level analysis rather than as gold-standard annotations.

Acknowledgments

We warmly thank our colleagues from INRIA for helping us with our pilot human comparison study: Virginie Mouilleron, Gabrielle Le Bellier, Elodie Etienne, Malik Marmonier, Adle Ben Salem, Célia Nouri, Panagiotis Tsolakis, and Nicolas Angleraud. We also thank Virginie Mouilleron for helping us with the verification of the Llama-generated annotations and Wissam Antoun for verifying the presence of our dataset in FineWeb2 and for helpful discussions of the results.

This work was performed using HPC resources from GENCI–IDRIS (Grant 2026-AD011017684) on the supercomputer Jean Zay's CSL, A100, and H100 partitions. This project was supported by the BPI Scribe project as well as by Djamé Seddah's PRAIRIE-PSAI chair, funded by the French national agency ANR, as part of the "France 2030" strategy under the reference ANR-23-IACL-0008.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Mai Alkhamissi, Yunze Xiao, Badr AlKhamissi, and Mona T. Diab. 2026. Hire your anthropologist! re-

thinking culture benchmarks through an anthropological lens. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1218–1235, Rabat, Morocco. Association for Computational Linguistics.

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, and 84 others. 2025. Apertus: Democratizing open and compliant llms for global language environments. *Preprint*, arXiv:2509.14233.

Sanket Badhe, Deep Shah, and Nehal Kathrotia. 2026. Long-tail knowledge in large language models: Taxonomy, mechanisms, interventions and implications. *Preprint*, arXiv:2602.16201.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. ALLam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Toms Bergmanis, Martins Kronis, Ingus Jānis Pretkalniņš, Dāvis Nicmanis, Jelizaveta Jelinska, Roberts Rozis, Rinalds Vīksna, and Mārcis Pinnis. 2026. Tildeopen llm: Leveraging curriculum learning to achieve equitable language representation. *Preprint*, arXiv:2603.08182.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Ethnologue. 2026. What are the top 200 most spoken languages? `http://www.ethnologue.com/guides/ethnologue200`. [Accessed 01-05-2026].

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. 2024. MERA: A comprehensive LLM evaluation in Russian. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9920–9948, Bangkok, Thailand. Association for Computational Linguistics.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *Preprint*, arXiv:2402.09369.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Nathan Godey, Wissam Antoun, Rian Touchent, Rachel Bawden, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2025. Gaperon: A peppered english-french generative language model suite. *Preprint*, arXiv:2510.25771.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, and 5 others. 2025. Salamandra technical report. *Preprint*, arXiv:2502.08489.

Olivier Gouvert, Julie Hunter, Jérôme Louradour, Christophe Cerisara, Evan Dufraisse, Yaya Sy, Laura Rivière, Jean-Pierre Lorré, and OpenLLM-France community. 2025. The lucie-7b llm and the lucie training dataset: Open resources for multilingual language generation. *Preprint*, arXiv:2503.12294.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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)*.

Maxim Ifergan, Leshem Choshen, Roee Aharoni, Idan Szpektor, and Omri Abend. 2025. Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4630–4644, Albuquerque, New Mexico. Association for Computational Linguistics.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. Break the checkbox: Challenging closed-style evaluations of cultural alignment in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24–51, Suzhou, China. Association for Computational Linguistics.

Yannis Karmim, Renato Pino, Hernan Contreras, Hernan Lira, Sebastian Cifuentes, Simon Escoffier, Luis Martí, Djamé Seddah, and Valentin Barriere. 2026. Leveraging Wikidata for geographically informed sociocultural bias dataset creation: Application to Latin America. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 177–188, Rabat, Morocco. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Peiqin Lin, Chenyang Lyu, Wenjiang Luo, Haotian Ye, Md Mehrab Hossain, Chunlan Ma, Shaoxiong Ji, Younes Samih, Bo Zeng, Fan Jiang, Yuanbin Cao, Dilda Duisenbek, Adrian Neo Sau Xun, Daria Pozdniakova, Liubou Misevich, Nevena Marinković, Ngoc Gia Linh Nguyen, Thi Khanh Linh Do, Sarakmatak Sophy, and 6 others. 2026. Culturall: Benchmarking multilingual and multicultural competence of llms on grounded tasks. *Preprint*, arXiv:2604.19262.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*.

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

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report. *Preprint*, arXiv:2506.04079.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Juhyun Oh, Inha Cha, Michael Saxon, Hyunseung Lim, Shaily Bhatt, and Alice Oh. 2025. Culture is everywhere: A call for intentionally cultural evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19156–19168, Suzhou, China. Association for Computational Linguistics.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In *Proceedings of the Tenth Italian Conference*

on Computational Linguistics (CLiC-it 2024), pages 707–719, Pisa, Italy. CEUR Workshop Proceedings.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. LLäMmlein: Transparent, compact and competitive German-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.

Long Phan, Alice Gatti, Nathaniel Li, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dan Hendrycks, Ziwen Han, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, and 281 others. 2026. A benchmark of expert-level academic questions to assess ai capabilities. *Nature*, 649(8099):1139–1146.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. *Sabiá: Portuguese Large Language Models*, page 226–240. Springer Nature Switzerland.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Miguel Moura Ramos, Duarte M. Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M. Guerreiro, Ricardo Rei, Nicolas Boizard, Amin Farajian, Mateusz Klimaszewski, José G. C. de Souza, Barry Haddow, François Yvon, Pierre Colombo, Alexandra Birch, and André F. T. Martins. 2026. Eurollm-22b: Technical report. *Preprint*, arXiv:2602.05879.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? 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 311–325, Mexico City, Mexico. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3.

UNESCO. 2026. What is Culture? Meaning, Definition. `https://www.unesco.org/en/query-list/c/culture`. [Accessed 01-05-2026].

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *Preprint*, arXiv:2411.04368.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25869–25886, Vienna, Austria. Association for Computational Linguistics.

Liya Zhu, Peizhuang Cong, Jingzhe Ding, Aowei Ji, Wenya Wu, Jiani Hou, Chunjie Wu, Xiang Gao, Jingkai Liu, Zhou Huan, Xuelei Sun, Yang Yang, Jianpeng Jiao, Liang Hu, Xinjie Chen, Jiashuo Liu, Tong Yang, Zaiyuan Wang, Ge Zhang, and Wenhao Huang. 2026. Lpfqa: A long-tail professional forum-based benchmark for llm evaluation. *Preprint*, arXiv:2511.06346.

A Dataset statistics

This section provides additional statistics for TRIVIAROOMQA-FRENCH. Table 3 reports the distribution of questions and topics by continent, Table 4 reports the distribution by time period, and Table 5 reports the distribution by category. Finally, Table 15 lists the topics associated with each category. These statistics are used to contextualize the analyses in Section 5, especially because continent and time-period results can be affected by the distribution of topics and question difficulty.

Continent	# Questions	# Topics
Europe	4288	258
North America	1955	228
Asia	690	160
unknown	1065	121
Africa	341	116
South America	198	79
Australia	66	49
Antarctica	35	6
Arctic	2	2

Table 3: Number of questions and topics per continent in TRIVIAROOMQA-FRENCH.

Time period	# Questions
Timeless	1742
Historical	1652
2010s	1601
2000s	1277
1990s	713
1980s	646
1970s	316
1960s	280
1950s	218
2020s	195

Table 4: Number of questions per time period in TRIVIAROOMQA-FRENCH.

B Data annotation prompts

Tables 16 and 17 show the prompts used to annotate questions with continent and time period, respectively. The prompts are written in English, while the example questions are provided in French to ensure consistency with the dataset language.

Category	# Questions	# Topics
Brands	90	3
Living creatures	330	11
Famous places	330	11
People	1440	48
Movie	1080	36
Places	480	16
Geography	240	8
Programming	360	12
Sport	630	21
Food	480	16
General knowledge	600	20
Health	120	4
Math	90	3
Nature	150	5
History	180	6
Objects	150	5
Miscellaneous	210	7
Colors	60	2
Transport	90	3
Music	210	7
News	960	32
Language	210	7
Celebrations	90	3
Chemistry	60	2

Table 5: Number of questions and topics per category in TRIVIAROOMQA-FRENCH.

C Human validation of the automatic annotations

As the continent and time-period annotations were produced automatically using the Llama-70B-Instruct model, we manually validate a sample of 100 random questions from TRIVIAROOMQA-FRENCH using two annotators. Table 6 reports the resulting automatic annotation accuracy and inter-annotator agreement. Continent annotations achieve 92% accuracy, while time-period annotations achieve 79.5%, with raw inter-annotator agreement of 0.90 and 0.87, respectively. At the same time, the relatively low Cohen's κ for continent annotation may partially result from the ambiguity of assigning geographic metadata to some questions, especially those related to fictional events or concerning several countries at the same time. Previous work has already shown that annotator disagreement can arise from uncertainty in item interpretation (Jiang and de Marneffe, 2022). These results show that the automatically produced metadata can be used as approximate annotations

Metric	Continent	Time period
Automatic annotation accuracy	0.920	0.795
Raw inter-annotator agreement	0.900	0.870
Cohen's κ	0.324	0.604
Positive agreement	0.946	0.918
Negative agreement	0.375	0.683

Table 6: Human validation of the automatically generated continent and time-period annotations on 100 questions from the TRIVIAROOMQA-FRENCH.

for our aggregate analyses, but we do not use them as gold-standard labels.

D Human evaluation of the six hardest and easiest TRIVIAROOMQA-FRENCH topics

In this section, we provide details on the pilot experiment comparing average human performance with the performance of the 30 LLMs used in our evaluation.

We selected the six easiest and hardest topics for models by computing average accuracy score for each topic in TRIVIAROOMQA-FRENCH.

The easiest topics are: *Cities of the world, City nicknames, Alan Turing, Famous Alberts, Microsoft, Internet Frenglish*.

The hardest topics are: *Cheeses of France, Animals of all kinds, Jean-Marie Bigard, Short quotes, Mike Horn, John McEnroe*[12].

Five French-speaking participants were recruited from among colleagues and were informed about the purpose of the study. All participants took part voluntarily. They answered questions from the twelve topics using the same multiple-choice format as the models.

Participants had an average age of 29.9 years ($\pm$5.6); 60% were women and 40% were men. Four participants were French, and one was Belgian.

This experiment is intended as a pilot comparison rather than a definitive estimate of human quiz performance and its goal is to compare general difficulty patterns: whether performance decreases gradually with annotated difficulty, and whether this behavior differs between humans and LLMs.

E Model enhancement with search tool

To test whether web access improves performance on TRIVIAROOMQA-FRENCH, we run an additional experiment with tool-using models under 20B parameters: Qwen3.5-9B and Llama-3.1-8B[13]. We provide them with a DuckDuckGo search tool and compare their performance with and without search access. Table 7 reports the results.

Model	Base	With search
Qwen3.5-9B	0.47	0.79
Llama-3.1-8B	0.51	0.63

Table 7: Comparison of model accuracy on TRIVIAROOMQA-FRENCH with and without access to a DuckDuckGo search tool.

In this preliminary experiment, search access improves performance for both tested models, with a particularly large gain for Qwen3.5-9B. This suggests that some TRIVIAROOMQA questions involve information that is difficult to retrieve from model parameters alone. However, the different gains also indicate that the improvement of scores depends on the model's ability to formulate useful search queries, which might depend on the model's recency and tool-use-specific enhancements, as in the case of Qwen.

F Benchmark examples

Table 8 presents examples of questions from different topics in our benchmark, along with their possible answers, correct answer, topic, and difficulty level.

G Cross-lingual inconsistency in model answers

To further analyze how different models perform across languages on our parallel TRIVIAROOMQA-MULTI dataset, we plot in Figure 6 the percentage of questions for which each model answers correctly in all six languages (green), in 1-5 languages (yellow), or in none of the languages (red). As shown in the figure, on average, 52% of questions show cross-lingual inconsistency, i.e., they are answered correctly in some languages but incorrectly in others.

We additionally examine cases where at least 10 models answer the English version correctly while failing in at least three of the five remaining languages. We identify 181 such questions (5.5% of the dataset). These questions span a broad range of categories, most frequently General Knowledge

[12] Jean-Marie Bigard is a French humorist, Mike Horn is a South African–born Swiss explorer, and John McEnroe is an American tennis player.

[13] Models are served using Ollama and prompted using the Pydantic AI library.

1. What did the rich of ancient Greece eat at breakfast?			
Possible answers: Fish and syrup, Bread and pure wine, Cereals and water, Fruits and honey			
Correct answer: Bread and pure wine	**Topic:** Breakfast	**Difficulty:** Confirmed	
2. Which former bodybuilder was governor of California from 2003 to 2011?			
Possible answers: Reagan, Schwarzenegger, Davis, Rolph			
Correct answer: Schwarzenegger	**Topic:** Mixed general knowledge	**Difficulty:** Beginner	
3. Who is one of the three best French football coaches of the 20th century?			
Possible answers: Albert Dieudonné, Albert Finney, Albert Batteux, Albert Lebrun			
Correct answer: Albert Batteux	**Topic:** Famous Alberts	**Difficulty:** Expert	
4. What traditional Japanese art consists of the miniaturizing cultivation of plants?			
Possible answers: Bankan, Bonsai, Niwaki, Oden			
Correct answer: Bonsai	**Topic:** Japanese folklore	**Difficulty:** Beginner	
5. Which industrial cheese is generally coated with paraffin dyed red?			
Possible answers: Babybel, Saint-paulin, Epoisses, Chavroux			
Correct answer: Babybel	**Topic:** French cheeses	**Difficulty:** Beginner	

Table 8: Examples of questions from the benchmark across different topics and difficulty levels.

(23 questions), Sport (18), Movie (15), People (15), Food (14), and Living Creatures (14). Notably, models also show lower average accuracy on most of these categories, as discussed in Section 5.2, which suggests that model knowledge in these categories may be sensitive to the language in which a question is presented.

H Temporal variation across categories

As question difficulty may also depend on the time period associated with the knowledge being tested, we examine whether the differences observed across categories in Section 5.2 could be related to their temporal distribution. For each category, we compute the mean accuracy across the eight time periods from the 1950s to the 2020s, excluding timeless and historical questions. Table 9 reports the resulting mean accuracy and standard deviation across time periods. A higher standard deviation indicates that model performance within a category varies more strongly depending on the time period of the questions.

The results show that performance varies relatively little across time periods for categories such as Chemistry, People, General Knowledge, and Movies, suggesting that their difficulty is at least not solely explained by this factor. At the same time, some other categories, such as Health and History, show higher variation across time periods while also achieving high mean accuracy. These results suggest that differences in the temporal distribution of questions do not alone explain the observed category-level differences.

I Detailed error analysis

In this section, we provide a more detailed analysis of the categories in which the models fail more and less frequently. Across all categories, 362 questions (4.2% of the dataset) are answered incorrectly by all 30 models. The highest proportions of such questions are found in Programming (7.2%), Music (6.2%), Food and Brands (5.6% each). In contrast, only 0.6% of History questions are answered incorrectly by all models, and in the Math category there are no questions that are not answered correctly by at least one model.

At the topic level, in the People category, fewer than 10 models answer more than 50% of the questions correctly for topics such as Maxime Chattam, a French writer, Maria Sharapova, a Russian tennis player, and famous French film directors. The same pattern can be observed in the Food category for questions about Belgian beers, in the Living Creatures category for questions about horses and pigeons, and in the Language category for questions about famous quotations. Examples of questions answered incorrectly by all models include: (i) *What color is the label on a bottle of Delirium Tremens?* and (ii) *Where is the pigeon's uropygial gland, a gland specific to birds, located?*

All in all, these results suggest that lower model performance on some topics may be associated with several sources of difficulty, including questions about specific named entities and fine-grained factual knowledge that is not necessarily related to a particular named entity.

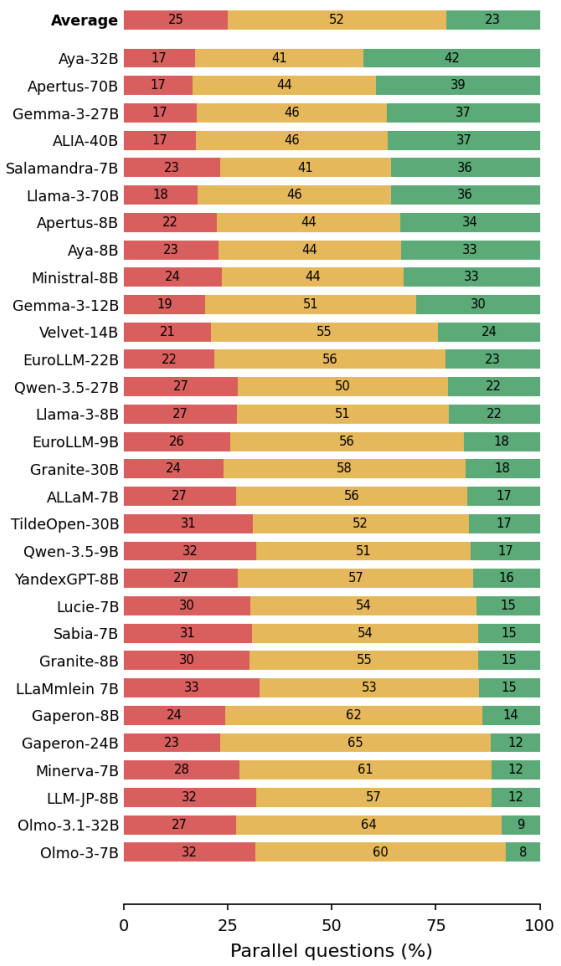

Figure 6: Percentage of questions answered correctly by each model in 0 (red), 1-5 (yellow), or 6 (green) languages on the 3,300 parallel questions from TRIVIAROOMQA-MULTI. The average across all models is shown at the top.

J Statistical Uncertainty of the Results

We report 95% confidence intervals (CIs) obtained from 1,000 non-parametric bootstrap resamples of the benchmark questions for the main analyses. For the multilingual analysis, parallel versions of the same underlying question are resampled together. For analyses aggregated across models, we additionally report the standard deviation (std) across models. The following tables provide uncertainty estimates for the language (Table 10), category (Table 11), difficulty (Table 12), temporal (Table 13) and geographic (Table 14) analyses presented in the main text.

Category	Mean accuracy [±std]
Music	40.8 [±13.0]
Colors	41.4 [±18.8]
Food	41.5 [±10.9]
Objects	45.6 [±27.9]
Living Creatures	47.1 [±15.1]
Programming	48.6 [±16.6]
People	49.6 [±5.0]
Movie	49.7 [±3.7]
Celebrations	50.1 [±17.2]
Nature	51.5 [±11.5]
Places	54.4 [±6.4]
Famous Places	54.8 [±19.0]
Sport	55.8 [±7.2]
Language	56.5 [±21.0]
General knowledge	56.5 [±5.8]
Geography	56.9 [±8.8]
Miscellaneous	58.5 [±6.0]
Transport	59.0 [±11.5]
Chemistry	61.3 [±4.7]
Brands	62.3 [±15.3]
News	63.8 [±14.9]
Health	64.1 [±30.8]
Math	68.8 [±9.7]
History	75.4 [±14.7]

Table 9: Mean accuracy and standard deviation across time periods for each category for questions from the 1950s to the 2020s in TRIVIAROOMQA-FRENCH. Timeless and historical questions are excluded.

K Additional results by difficulty level, continent, and time period

This section provides additional fine-grained results that complement the main analysis. In all figures, the reported metric is accuracy, greener cells indicate higher scores. ⋆ next to a score or a model indicates that the model was trained on that language. ▲ marks models trained on more than 10T tokens. Model size is color-coded: models with less than 10B parameters are shown in blue, with 12-27B parameters in green, with 30B+ parameters in pink.

Figure 7 reports per-model performance by question difficulty, showing that model performance generally follows the annotated difficulty levels: accuracy decreases as questions become harder. Figure 10 provides more detailed time-period results, confirming the trend observed in the aggregated results in Figure 4: model performance generally decreases on more contemporary questions.

	English	French	Italian	German	Spanish	Dutch
Aya-32B	0.67 [0.65, 0.68]	0.69 [0.67, 0.70]	0.62 [0.60, 0.63]	0.62 [0.61, 0.64]	0.61 [0.59, 0.63]	0.61 [0.60, 0.63]
Apertus-70B	0.63 [0.62, 0.65]	0.67 [0.65, 0.68]	0.60 [0.59, 0.62]	0.62 [0.60, 0.64]	0.62 [0.60, 0.64]	0.60 [0.58, 0.62]
Gemma-3-27B	0.63 [0.62, 0.65]	0.66 [0.64, 0.68]	0.60 [0.58, 0.61]	0.59 [0.57, 0.60]	0.59 [0.57, 0.60]	0.59 [0.57, 0.61]
ALIA-40B	0.61 [0.60, 0.63]	0.65 [0.64, 0.67]	0.59 [0.57, 0.60]	0.59 [0.57, 0.61]	0.59 [0.58, 0.61]	0.59 [0.57, 0.60]
Llama-3-70B	0.63 [0.61, 0.65]	0.66 [0.64, 0.68]	0.58 [0.56, 0.60]	0.58 [0.56, 0.60]	0.58 [0.56, 0.59]	0.58 [0.56, 0.59]
Salamandra-7B	0.59 [0.57, 0.61]	0.62 [0.60, 0.64]	0.57 [0.55, 0.58]	0.55 [0.53, 0.57]	0.55 [0.54, 0.57]	0.55 [0.53, 0.56]
Apertus-8B	0.57 [0.55, 0.59]	0.58 [0.57, 0.60]	0.55 [0.53, 0.57]	0.55 [0.53, 0.57]	0.56 [0.54, 0.57]	0.54 [0.53, 0.56]
Ministral-8B	0.60 [0.59, 0.62]	0.62 [0.60, 0.63]	0.55 [0.54, 0.57]	0.53 [0.51, 0.54]	0.54 [0.53, 0.56]	0.50 [0.48, 0.51]
Gemma-3-12B	0.59 [0.58, 0.61]	0.61 [0.59, 0.62]	0.53 [0.51, 0.55]	0.54 [0.53, 0.56]	0.52 [0.51, 0.54]	0.53 [0.51, 0.55]
Aya-8B	0.58 [0.56, 0.60]	0.60 [0.58, 0.61]	0.54 [0.52, 0.55]	0.54 [0.53, 0.56]	0.54 [0.52, 0.55]	0.53 [0.51, 0.55]
Velvet-14B	0.56 [0.55, 0.58]	0.57 [0.56, 0.59]	0.54 [0.53, 0.56]	0.51 [0.49, 0.52]	0.53 [0.52, 0.55]	0.43 [0.41, 0.45]
EuroLLM-22B	0.51 [0.49, 0.52]	0.55 [0.53, 0.56]	0.49 [0.47, 0.51]	0.47 [0.45, 0.49]	0.50 [0.49, 0.52]	0.46 [0.44, 0.47]
Qwen-3.5-27B	0.50 [0.48, 0.52]	0.50 [0.48, 0.52]	0.46 [0.45, 0.48]	0.46 [0.45, 0.48]	0.45 [0.43, 0.47]	0.45 [0.43, 0.46]
Llama-3-8B	0.52 [0.50, 0.53]	0.49 [0.47, 0.50]	0.46 [0.44, 0.48]	0.45 [0.43, 0.46]	0.44 [0.42, 0.46]	0.45 [0.44, 0.47]
Granite-30B	0.53 [0.51, 0.54]	0.49 [0.48, 0.51]	0.43 [0.41, 0.45]	0.44 [0.42, 0.46]	0.44 [0.42, 0.46]	0.44 [0.42, 0.46]
EuroLLM-9B	0.46 [0.44, 0.48]	0.47 [0.46, 0.49]	0.43 [0.42, 0.45]	0.42 [0.41, 0.44]	0.45 [0.43, 0.47]	0.43 [0.41, 0.44]
ALLaM-7B	0.51 [0.49, 0.52]	0.46 [0.44, 0.48]	0.43 [0.41, 0.45]	0.42 [0.40, 0.44]	0.40 [0.39, 0.42]	0.41 [0.39, 0.43]
YandexGPT-8B	0.49 [0.47, 0.50]	0.43 [0.41, 0.45]	0.40 [0.39, 0.42]	0.40 [0.39, 0.42]	0.42 [0.40, 0.43]	0.41 [0.39, 0.43]
Gaperon-8B	0.44 [0.42, 0.46]	0.49 [0.47, 0.50]	0.40 [0.39, 0.42]	0.40 [0.38, 0.42]	0.38 [0.37, 0.40]	0.41 [0.39, 0.43]
Lucie-7B	0.45 [0.43, 0.46]	0.47 [0.46, 0.49]	0.40 [0.39, 0.42]	0.40 [0.39, 0.42]	0.42 [0.41, 0.44]	0.33 [0.32, 0.35]
TildeOpen-30B	0.40 [0.39, 0.42]	0.44 [0.42, 0.45]	0.41 [0.39, 0.43]	0.41 [0.40, 0.43]	0.40 [0.39, 0.42]	0.41 [0.39, 0.43]
Qwen-3.5-9B	0.43 [0.42, 0.45]	0.43 [0.42, 0.45]	0.40 [0.38, 0.41]	0.40 [0.39, 0.42]	0.39 [0.37, 0.41]	0.39 [0.38, 0.41]
Gaperon-24B	0.45 [0.43, 0.47]	0.47 [0.45, 0.48]	0.39 [0.37, 0.40]	0.38 [0.37, 0.40]	0.38 [0.36, 0.39]	0.38 [0.37, 0.40]
Granite-8B	0.46 [0.44, 0.48]	0.40 [0.38, 0.41]	0.38 [0.37, 0.40]	0.39 [0.37, 0.40]	0.39 [0.37, 0.40]	0.39 [0.38, 0.41]
Sabia-7B	0.45 [0.43, 0.46]	0.41 [0.40, 0.43]	0.38 [0.36, 0.40]	0.38 [0.36, 0.40]	0.40 [0.38, 0.41]	0.38 [0.36, 0.40]
LLaMmlein-7B	0.43 [0.41, 0.45]	0.39 [0.37, 0.40]	0.36 [0.34, 0.37]	0.46 [0.44, 0.48]	0.36 [0.34, 0.37]	0.37 [0.36, 0.39]
Minerva-7B	0.45 [0.44, 0.47]	0.38 [0.36, 0.39]	0.44 [0.43, 0.46]	0.34 [0.32, 0.36]	0.37 [0.36, 0.39]	0.34 [0.33, 0.36]
Olmo-3.1-32B	0.45 [0.43, 0.47]	0.39 [0.38, 0.41]	0.37 [0.35, 0.39]	0.37 [0.35, 0.38]	0.33 [0.31, 0.34]	0.36 [0.34, 0.37]
LLM-JP-8B	0.42 [0.41, 0.44]	0.37 [0.35, 0.39]	0.35 [0.34, 0.37]	0.36 [0.35, 0.38]	0.36 [0.34, 0.38]	0.36 [0.35, 0.38]
Olmo-3-7B	0.38 [0.36, 0.40]	0.34 [0.32, 0.36]	0.32 [0.30, 0.34]	0.32 [0.30, 0.33]	0.32 [0.30, 0.33]	0.32 [0.30, 0.33]

Table 10: Accuracy and 95% bootstrap confidence intervals for each model across languages reported in Figure 1.

Category	Size	Accuracy [±std]	95% CI
Music	210	0.46 [± 0.10]	[0.41, 0.50]
Nature	150	0.47 [± 0.10]	[0.43, 0.52]
Food	480	0.47 [± 0.10]	[0.44, 0.50]
Colors	60	0.49 [± 0.14]	[0.41, 0.57]
Movie	1080	0.49 [± 0.11]	[0.48, 0.51]
People	1440	0.50 [± 0.10]	[0.48, 0.51]
News	960	0.50 [± 0.10]	[0.48, 0.52]
Living creatures	330	0.51 [± 0.11]	[0.47, 0.54]
Sport	630	0.52 [± 0.11]	[0.50, 0.54]
Miscellaneous	210	0.55 [± 0.12]	[0.51, 0.59]
Celebrations	90	0.56 [± 0.14]	[0.50, 0.62]
Transport	90	0.57 [± 0.14]	[0.50, 0.62]
Programming	360	0.57 [± 0.10]	[0.53, 0.60]
Brands	90	0.57 [± 0.10]	[0.50, 0.64]
Famous places	330	0.58 [± 0.13]	[0.54, 0.61]
General knowledge	600	0.58 [± 0.13]	[0.56, 0.60]
Chemistry	60	0.58 [± 0.10]	[0.49, 0.67]
Language	210	0.59 [± 0.09]	[0.55, 0.64]
Places	480	0.60 [± 0.11]	[0.58, 0.63]
Health	120	0.61 [± 0.11]	[0.56, 0.66]
Objects	150	0.61 [± 0.13]	[0.56, 0.66]
Geography	240	0.63 [± 0.11]	[0.59, 0.67]
Math	90	0.67 [± 0.17]	[0.62, 0.72]
History	180	0.69 [± 0.13]	[0.65, 0.73]

Table 11: Mean accuracy, standard deviation, and 95% bootstrap confidence intervals for each category across models, corresponding to Figure 2.

Difficulty	Size	Accuracy [±std]	95% CI
Beginner	2900	0.60 [± 0.12]	[0.59, 0.61]
Confirmed	2870	0.52 [± 0.11]	[0.51, 0.53]
Expert	2870	0.48 [± 0.10]	[0.47, 0.49]

Table 12: Mean accuracy, standard deviation across models, and 95% bootstrap confidence intervals for each difficulty level reported in Figure 7.

Figure 11 shows the full continent-level results on TRIVIAROOMQA-FRENCH, including performance on questions for which the continent parameter is not applicable.

L Results with chat template

In the main experiments, we evaluate all models without chat templates to keep the setup comparable across models. In this section, we provide the corresponding results with chat templates for models where this setting is available. In all figures, the reported metric is accuracy, greener cells indicate higher scores. ⋆ next to a score or a model indicates that the model was trained on that language. ▲ marks models trained on more than 10T tokens. Model size is color-coded: models with less than

	Timeless (1742 q.)	**Historical** (1652 q.)	**1950s** (218 q.)	**1960s** (280 q.)	**1970s** (316 q.)	**1980s** (646 q.)	**1990s** (713 q.)	**2000s** (1277 q.)	**2010s** (1601 q.)	**2020s** (195 q.)
Velvet-14B	0.61 [0.59, 0.63]	0.68 [0.66, 0.71]	0.63 [0.57, 0.70]	0.60 [0.55, 0.66]	0.60 [0.54, 0.66]	0.55 [0.51, 0.59]	0.54 [0.50, 0.57]	0.57 [0.55, 0.60]	0.57 [0.55, 0.60]	0.59 [0.53, 0.66]
Apertus-8B	0.61 [0.59, 0.63]	0.73 [0.71, 0.75]	0.67 [0.61, 0.72]	0.64 [0.59, 0.70]	0.64 [0.58, 0.69]	0.57 [0.54, 0.61]	0.60 [0.56, 0.64]	0.59 [0.57, 0.62]	0.55 [0.53, 0.58]	0.63 [0.56, 0.70]
Aya-8B	0.62 [0.60, 0.65]	0.69 [0.67, 0.71]	0.63 [0.57, 0.70]	0.61 [0.55, 0.66]	0.60 [0.54, 0.65]	0.54 [0.50, 0.58]	0.57 [0.54, 0.61]	0.62 [0.59, 0.64]	0.59 [0.57, 0.62]	0.58 [0.51, 0.65]
EuroLLM-9B	0.53 [0.51, 0.56]	0.62 [0.59, 0.64]	0.61 [0.54, 0.67]	0.49 [0.43, 0.55]	0.49 [0.44, 0.55]	0.42 [0.39, 0.46]	0.47 [0.44, 0.51]	0.48 [0.45, 0.51]	0.42 [0.40, 0.45]	0.45 [0.38, 0.52]
Gaperon-8B	0.55 [0.52, 0.57]	0.59 [0.57, 0.62]	0.58 [0.52, 0.65]	0.52 [0.46, 0.57]	0.49 [0.44, 0.54]	0.44 [0.41, 0.48]	0.48 [0.44, 0.51]	0.48 [0.46, 0.51]	0.46 [0.44, 0.48]	0.47 [0.39, 0.53]
Ministral-8B	0.65 [0.62, 0.67]	0.73 [0.71, 0.75]	0.68 [0.62, 0.74]	0.64 [0.59, 0.70]	0.63 [0.58, 0.69]	0.58 [0.54, 0.62]	0.63 [0.59, 0.66]	0.62 [0.59, 0.65]	0.60 [0.57, 0.62]	0.63 [0.55, 0.69]
Llama-3-8B	0.49 [0.47, 0.52]	0.59 [0.57, 0.61]	0.60 [0.54, 0.67]	0.48 [0.42, 0.54]	0.53 [0.48, 0.59]	0.49 [0.44, 0.52]	0.50 [0.46, 0.54]	0.50 [0.47, 0.52]	0.47 [0.45, 0.49]	0.45 [0.38, 0.52]
LLaMmlein 7B	0.36 [0.34, 0.38]	0.43 [0.41, 0.46]	0.41 [0.34, 0.47]	0.38 [0.32, 0.44]	0.37 [0.31, 0.42]	0.39 [0.36, 0.43]	0.37 [0.33, 0.40]	0.41 [0.38, 0.43]	0.38 [0.35, 0.40]	0.37 [0.30, 0.44]
Lucie-7B	0.51 [0.48, 0.53]	0.58 [0.56, 0.61]	0.52 [0.46, 0.59]	0.47 [0.41, 0.53]	0.47 [0.42, 0.53]	0.42 [0.38, 0.46]	0.46 [0.42, 0.49]	0.47 [0.44, 0.49]	0.45 [0.43, 0.48]	0.52 [0.44, 0.58]
Olmo-3-7B	0.37 [0.35, 0.39]	0.38 [0.35, 0.40]	0.39 [0.32, 0.45]	0.38 [0.33, 0.44]	0.35 [0.30, 0.41]	0.33 [0.29, 0.37]	0.34 [0.30, 0.37]	0.36 [0.33, 0.38]	0.33 [0.30, 0.35]	0.39 [0.32, 0.46]
Qwen-3.5-9B	0.49 [0.47, 0.52]	0.56 [0.54, 0.59]	0.52 [0.45, 0.59]	0.49 [0.43, 0.54]	0.51 [0.46, 0.57]	0.42 [0.38, 0.46]	0.43 [0.40, 0.47]	0.44 [0.41, 0.46]	0.40 [0.38, 0.42]	0.45 [0.38, 0.52]
Gemma-3-12B	0.65 [0.63, 0.67]	0.71 [0.69, 0.73]	0.67 [0.61, 0.73]	0.64 [0.58, 0.69]	0.63 [0.59, 0.69]	0.58 [0.54, 0.61]	0.61 [0.57, 0.65]	0.63 [0.60, 0.65]	0.59 [0.57, 0.62]	0.61 [0.54, 0.67]
ALLaM-7B	0.47 [0.45, 0.49]	0.55 [0.53, 0.58]	0.53 [0.46, 0.59]	0.51 [0.46, 0.57]	0.55 [0.50, 0.61]	0.44 [0.40, 0.48]	0.49 [0.46, 0.53]	0.49 [0.46, 0.52]	0.43 [0.41, 0.46]	0.43 [0.35, 0.49]
Granite-8B	0.44 [0.42, 0.47]	0.49 [0.46, 0.51]	0.46 [0.39, 0.53]	0.41 [0.35, 0.46]	0.47 [0.41, 0.52]	0.40 [0.37, 0.44]	0.38 [0.35, 0.42]	0.42 [0.40, 0.45]	0.38 [0.35, 0.40]	0.36 [0.29, 0.43]
LLM-JP-8B	0.39 [0.37, 0.42]	0.44 [0.42, 0.47]	0.46 [0.39, 0.52]	0.41 [0.35, 0.47]	0.39 [0.33, 0.44]	0.39 [0.35, 0.43]	0.38 [0.34, 0.42]	0.39 [0.37, 0.42]	0.35 [0.33, 0.38]	0.35 [0.29, 0.42]
Sabia-7B	0.41 [0.39, 0.44]	0.49 [0.46, 0.51]	0.49 [0.42, 0.56]	0.44 [0.38, 0.50]	0.43 [0.37, 0.48]	0.41 [0.37, 0.45]	0.43 [0.39, 0.47]	0.43 [0.40, 0.46]	0.41 [0.39, 0.43]	0.39 [0.32, 0.46]
Salamandra-7B	0.63 [0.60, 0.65]	0.73 [0.71, 0.75]	0.69 [0.63, 0.76]	0.63 [0.57, 0.69]	0.63 [0.58, 0.68]	0.58 [0.54, 0.62]	0.61 [0.57, 0.64]	0.60 [0.57, 0.63]	0.58 [0.56, 0.61]	0.67 [0.61, 0.73]
Minerva-7B	0.36 [0.34, 0.38]	0.42 [0.40, 0.45]	0.43 [0.37, 0.50]	0.41 [0.35, 0.47]	0.42 [0.37, 0.48]	0.39 [0.35, 0.43]	0.38 [0.35, 0.42]	0.38 [0.36, 0.41]	0.38 [0.36, 0.40]	0.39 [0.32, 0.46]
YandexGPT-8B	0.46 [0.44, 0.48]	0.52 [0.49, 0.54]	0.47 [0.40, 0.54]	0.44 [0.38, 0.49]	0.45 [0.40, 0.51]	0.38 [0.35, 0.42]	0.42 [0.38, 0.46]	0.44 [0.42, 0.47]	0.41 [0.39, 0.43]	0.45 [0.37, 0.52]
Apertus-70B	0.68 [0.66, 0.70]	0.77 [0.75, 0.79]	0.79 [0.73, 0.84]	0.71 [0.65, 0.76]	0.71 [0.66, 0.76]	0.65 [0.61, 0.68]	0.67 [0.63, 0.70]	0.70 [0.68, 0.73]	0.66 [0.64, 0.69]	0.68 [0.61, 0.75]
ALIA-40B	0.65 [0.63, 0.67]	0.77 [0.75, 0.79]	0.71 [0.65, 0.77]	0.71 [0.66, 0.77]	0.73 [0.68, 0.78]	0.63 [0.59, 0.66]	0.67 [0.64, 0.70]	0.69 [0.67, 0.71]	0.65 [0.62, 0.67]	0.65 [0.58, 0.72]
Aya-32B	0.70 [0.67, 0.72]	0.79 [0.77, 0.81]	0.79 [0.74, 0.85]	0.74 [0.69, 0.79]	0.73 [0.68, 0.78]	0.68 [0.65, 0.72]	0.68 [0.65, 0.71]	0.72 [0.70, 0.74]	0.71 [0.69, 0.73]	0.66 [0.59, 0.72]
EuroLLM-22B	0.59 [0.57, 0.61]	0.65 [0.63, 0.67]	0.63 [0.57, 0.69]	0.59 [0.53, 0.64]	0.58 [0.52, 0.63]	0.51 [0.47, 0.55]	0.52 [0.48, 0.55]	0.55 [0.52, 0.58]	0.52 [0.49, 0.54]	0.60 [0.53, 0.67]
Llama-3-70B	0.67 [0.65, 0.69]	0.77 [0.75, 0.79]	0.74 [0.68, 0.79]	0.71 [0.66, 0.76]	0.65 [0.59, 0.70]	0.67 [0.63, 0.71]	0.72 [0.68, 0.75]	0.68 [0.65, 0.70]	0.62 [0.60, 0.65]	0.63 [0.56, 0.69]
Qwen-3.5-27B	0.55 [0.53, 0.58]	0.65 [0.63, 0.67]	0.58 [0.52, 0.64]	0.51 [0.46, 0.57]	0.49 [0.44, 0.54]	0.44 [0.41, 0.48]	0.45 [0.41, 0.49]	0.49 [0.46, 0.51]	0.47 [0.44, 0.49]	0.48 [0.41, 0.55]
TildeOpen-30B	0.47 [0.45, 0.50]	0.50 [0.47, 0.52]	0.47 [0.41, 0.54]	0.44 [0.38, 0.50]	0.44 [0.39, 0.49]	0.40 [0.36, 0.44]	0.42 [0.38, 0.45]	0.45 [0.42, 0.48]	0.43 [0.41, 0.46]	0.51 [0.43, 0.58]
Olmo-3.1-32B	0.40 [0.37, 0.42]	0.45 [0.43, 0.47]	0.46 [0.39, 0.52]	0.43 [0.37, 0.48]	0.44 [0.39, 0.49]	0.39 [0.36, 0.43]	0.40 [0.36, 0.43]	0.42 [0.39, 0.44]	0.39 [0.36, 0.41]	0.35 [0.29, 0.43]
Gaperon-24B	0.51 [0.48, 0.53]	0.57 [0.55, 0.59]	0.54 [0.48, 0.61]	0.50 [0.45, 0.56]	0.51 [0.45, 0.56]	0.40 [0.36, 0.44]	0.41 [0.38, 0.45]	0.46 [0.43, 0.48]	0.44 [0.42, 0.47]	0.48 [0.41, 0.55]
Gemma-3-27B	0.70 [0.67, 0.72]	0.78 [0.76, 0.80]	0.74 [0.68, 0.79]	0.66 [0.61, 0.72]	0.66 [0.61, 0.72]	0.62 [0.58, 0.65]	0.66 [0.63, 0.70]	0.66 [0.63, 0.68]	0.65 [0.62, 0.67]	0.68 [0.62, 0.74]
Granite-30B	0.53 [0.51, 0.55]	0.58 [0.56, 0.60]	0.58 [0.51, 0.64]	0.49 [0.43, 0.55]	0.51 [0.45, 0.56]	0.48 [0.44, 0.52]	0.51 [0.47, 0.54]	0.52 [0.49, 0.54]	0.45 [0.43, 0.47]	0.49 [0.42, 0.56]

Table 13: Accuracy and 95% bootstrap confidence intervals for each model across time periods corresponding to Figures 4 and 10.

	Africa (341 q.)	**Asia** (690 q.)	**Europe** (4288 q.)	**N. America** (1955 q.)	**S. America** (198 q.)	**Australia** (66 q.)	**Antarctica** (35 q.)	**Arctic** (2 q.)	**unknown** (1065 q.)
Velvet-14B	0.66 [0.61, 0.71]	0.63 [0.59, 0.66]	0.59 [0.58, 0.60]	0.62 [0.59, 0.64]	0.64 [0.58, 0.71]	0.67 [0.55, 0.77]	0.57 [0.40, 0.71]	0.50 [0.00, 1.00]	0.58 [0.55, 0.60]
Apertus-8B	0.68 [0.63, 0.73]	0.66 [0.62, 0.70]	0.59 [0.58, 0.60]	0.67 [0.65, 0.69]	0.67 [0.60, 0.73]	0.59 [0.47, 0.70]	0.66 [0.51, 0.80]	0.50 [0.00, 1.00]	0.59 [0.56, 0.62]
Aya-8B	0.70 [0.65, 0.75]	0.66 [0.62, 0.69]	0.58 [0.57, 0.60]	0.66 [0.64, 0.68]	0.66 [0.59, 0.73]	0.67 [0.55, 0.77]	0.71 [0.57, 0.86]	0.50 [0.00, 1.00]	0.61 [0.58, 0.64]
EuroLLM-9B	0.56 [0.51, 0.61]	0.54 [0.50, 0.58]	0.49 [0.48, 0.51]	0.50 [0.48, 0.52]	0.52 [0.44, 0.58]	0.56 [0.45, 0.68]	0.51 [0.37, 0.66]	0.50 [0.00, 1.00]	0.51 [0.48, 0.54]
Gaperon-8B	0.56 [0.51, 0.61]	0.53 [0.50, 0.57]	0.51 [0.49, 0.52]	0.50 [0.48, 0.52]	0.59 [0.52, 0.65]	0.58 [0.45, 0.68]	0.69 [0.54, 0.83]	0.50 [0.00, 1.00]	0.50 [0.47, 0.53]
Ministral-8B	0.72 [0.67, 0.77]	0.70 [0.66, 0.73]	0.61 [0.59, 0.62]	0.69 [0.67, 0.71]	0.65 [0.59, 0.72]	0.61 [0.48, 0.73]	0.66 [0.51, 0.80]	0.50 [0.00, 1.00]	0.64 [0.61, 0.67]
Llama-3-8B	0.57 [0.52, 0.63]	0.54 [0.51, 0.58]	0.48 [0.47, 0.50]	0.57 [0.54, 0.59]	0.55 [0.47, 0.62]	0.58 [0.45, 0.70]	0.60 [0.43, 0.74]	0.50 [0.00, 1.00]	0.46 [0.43, 0.49]
LLaMmlein 7B	0.39 [0.34, 0.44]	0.38 [0.35, 0.42]	0.38 [0.37, 0.39]	0.44 [0.41, 0.46]	0.41 [0.34, 0.47]	0.45 [0.33, 0.58]	0.37 [0.23, 0.54]	0.00 [0.00, 0.00]	0.34 [0.32, 0.37]
Lucie-7B	0.52 [0.47, 0.57]	0.50 [0.46, 0.54]	0.50 [0.48, 0.51]	0.48 [0.46, 0.51]	0.50 [0.43, 0.57]	0.44 [0.32, 0.56]	0.57 [0.40, 0.71]	0.50 [0.00, 1.00]	0.48 [0.45, 0.51]
Olmo-3-7B	0.39 [0.35, 0.45]	0.37 [0.33, 0.40]	0.33 [0.32, 0.34]	0.39 [0.37, 0.42]	0.36 [0.30, 0.43]	0.41 [0.30, 0.53]	0.34 [0.20, 0.51]	0.50 [0.00, 1.00]	0.37 [0.34, 0.40]
Qwen-3.5-9B	0.52 [0.47, 0.58]	0.49 [0.46, 0.53]	0.44 [0.43, 0.46]	0.50 [0.48, 0.52]	0.47 [0.41, 0.54]	0.45 [0.33, 0.58]	0.51 [0.37, 0.69]	0.50 [0.00, 1.00]	0.50 [0.47, 0.53]
Gemma-3-12B	0.70 [0.65, 0.75]	0.69 [0.66, 0.73]	0.60 [0.58, 0.61]	0.70 [0.68, 0.72]	0.66 [0.60, 0.73]	0.64 [0.52, 0.74]	0.71 [0.57, 0.86]	0.50 [0.00, 1.00]	0.63 [0.60, 0.66]
ALLaM-7B	0.53 [0.48, 0.59]	0.51 [0.47, 0.54]	0.45 [0.44, 0.47]	0.55 [0.53, 0.58]	0.50 [0.43, 0.57]	0.50 [0.38, 0.62]	0.46 [0.29, 0.63]	0.50 [0.00, 1.00]	0.46 [0.43, 0.49]
Granite-8B	0.45 [0.39, 0.50]	0.42 [0.39, 0.46]	0.40 [0.39, 0.42]	0.48 [0.46, 0.50]	0.45 [0.38, 0.52]	0.44 [0.32, 0.56]	0.37 [0.23, 0.54]	0.00 [0.00, 0.00]	0.43 [0.40, 0.46]
LLM-JP-8B	0.43 [0.38, 0.48]	0.40 [0.37, 0.44]	0.37 [0.36, 0.39]	0.44 [0.42, 0.46]	0.41 [0.34, 0.48]	0.41 [0.29, 0.53]	0.31 [0.17, 0.49]	0.50 [0.00, 1.00]	0.39 [0.36, 0.42]
Sabia-7B	0.44 [0.39, 0.50]	0.43 [0.40, 0.47]	0.42 [0.40, 0.43]	0.48 [0.45, 0.50]	0.46 [0.39, 0.53]	0.44 [0.32, 0.56]	0.46 [0.29, 0.63]	0.50 [0.00, 1.00]	0.42 [0.38, 0.45]
Salamandra-7B	0.69 [0.64, 0.74]	0.67 [0.63, 0.70]	0.60 [0.59, 0.62]	0.67 [0.65, 0.69]	0.67 [0.60, 0.73]	0.73 [0.62, 0.83]	0.63 [0.46, 0.77]	0.50 [0.00, 1.00]	0.62 [0.59, 0.65]
Minerva-7B	0.38 [0.33, 0.43]	0.38 [0.34, 0.42]	0.39 [0.37, 0.40]	0.43 [0.41, 0.45]	0.37 [0.31, 0.44]	0.32 [0.21, 0.42]	0.43 [0.26, 0.60]	0.50 [0.00, 1.00]	0.34 [0.31, 0.37]
YandexGPT-8B	0.49 [0.44, 0.55]	0.48 [0.44, 0.51]	0.42 [0.41, 0.44]	0.47 [0.45, 0.50]	0.49 [0.42, 0.56]	0.47 [0.36, 0.58]	0.49 [0.31, 0.63]	0.50 [0.00, 1.00]	0.47 [0.44, 0.50]
Apertus-70B	0.77 [0.72, 0.81]	0.69 [0.66, 0.73]	0.67 [0.66, 0.69]	0.74 [0.72, 0.76]	0.76 [0.70, 0.82]	0.64 [0.52, 0.74]	0.66 [0.51, 0.80]	1.00 [1.00, 1.00]	0.69 [0.66, 0.72]
ALIA-40B	0.76 [0.72, 0.81]	0.72 [0.68, 0.75]	0.67 [0.65, 0.68]	0.72 [0.70, 0.74]	0.70 [0.64, 0.76]	0.71 [0.59, 0.82]	0.71 [0.54, 0.86]	0.50 [0.00, 1.00]	0.62 [0.59, 0.65]
Aya-32B	0.79 [0.74, 0.83]	0.75 [0.72, 0.79]	0.70 [0.69, 0.72]	0.76 [0.74, 0.78]	0.74 [0.68, 0.80]	0.71 [0.59, 0.82]	0.69 [0.51, 0.83]	0.50 [0.00, 1.00]	0.68 [0.65, 0.70]
EuroLLM-22B	0.59 [0.54, 0.64]	0.59 [0.55, 0.63]	0.55 [0.54, 0.57]	0.59 [0.56, 0.61]	0.66 [0.60, 0.72]	0.59 [0.48, 0.70]	0.57 [0.40, 0.71]	0.50 [0.00, 1.00]	0.58 [0.55, 0.61]
Llama-3-70B	0.72 [0.67, 0.77]	0.72 [0.68, 0.75]	0.66 [0.64, 0.67]	0.74 [0.72, 0.76]	0.69 [0.63, 0.76]	0.74 [0.64, 0.83]	0.74 [0.60, 0.89]	0.50 [0.00, 1.00]	0.67 [0.64, 0.70]
Qwen-3.5-27B	0.58 [0.53, 0.63]	0.57 [0.53, 0.61]	0.50 [0.48, 0.51]	0.54 [0.52, 0.56]	0.52 [0.44, 0.59]	0.53 [0.41, 0.64]	0.60 [0.43, 0.77]	0.50 [0.00, 1.00]	0.55 [0.52, 0.58]
TildeOpen-30B	0.47 [0.42, 0.52]	0.49 [0.45, 0.53]	0.45 [0.43, 0.46]	0.47 [0.44, 0.49]	0.41 [0.34, 0.48]	0.45 [0.35, 0.58]	0.43 [0.26, 0.60]	0.50 [0.00, 1.00]	0.46 [0.43, 0.49]
Olmo-3.1-32B	0.41 [0.35, 0.45]	0.41 [0.38, 0.45]	0.39 [0.38, 0.41]	0.45 [0.43, 0.47]	0.40 [0.33, 0.47]	0.42 [0.30, 0.53]	0.43 [0.26, 0.57]	0.50 [0.00, 1.00]	0.41 [0.38, 0.44]
Gaperon-24B	0.50 [0.45, 0.55]	0.51 [0.47, 0.54]	0.49 [0.47, 0.50]	0.46 [0.44, 0.48]	0.49 [0.43, 0.57]	0.48 [0.36, 0.61]	0.60 [0.46, 0.77]	0.50 [0.00, 1.00]	0.49 [0.46, 0.51]
Gemma-3-27B	0.74 [0.70, 0.79]	0.73 [0.70, 0.77]	0.65 [0.64, 0.67]	0.73 [0.71, 0.75]	0.69 [0.63, 0.76]	0.64 [0.53, 0.74]	0.77 [0.63, 0.91]	0.50 [0.00, 1.00]	0.68 [0.65, 0.71]
Granite-30B	0.59 [0.54, 0.64]	0.56 [0.53, 0.60]	0.48 [0.46, 0.49]	0.56 [0.54, 0.58]	0.53 [0.45, 0.60]	0.50 [0.38, 0.62]	0.66 [0.49, 0.80]	0.00 [0.00, 0.00]	0.53 [0.51, 0.56]

Table 14: Accuracy and 95% bootstrap confidence intervals for each model across continents corresponding to Figures 5 and 11.

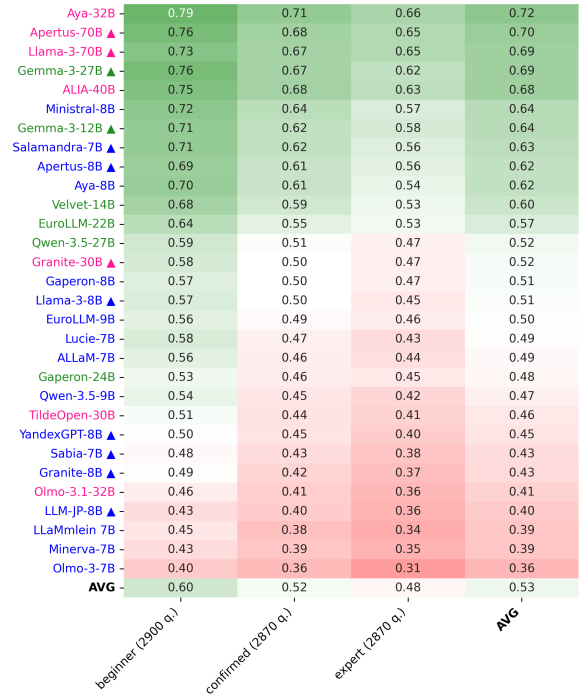

Figure 7: Model performance on TRIVIAROOMQA-FRENCH, grouped by question difficulty. The reported metric is accuracy; greener cells indicate higher scores.

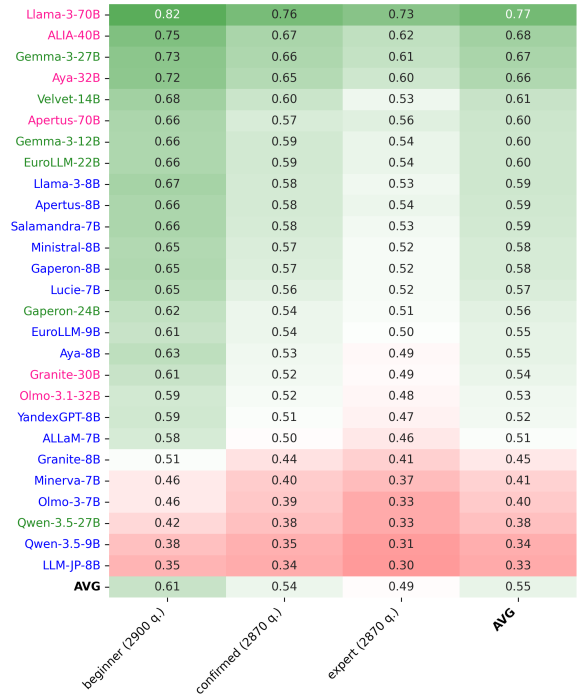

Figure 8: Model performance **with chat templates** on TRIVIAROOMQA-FRENCH, grouped by question difficulty. The reported metric is accuracy.

10B parameters are shown in blue, with 12-27B parameters in green, with 30B+ parameters in pink.

Figure 9 reports performance on the multilingual setting, while Figures 8, 12, 13, and 14 report results on the French collection by category, difficulty level, time period, and continent, respectively. Overall, the main trends remain similar to the no-chat-template setting: performance varies across languages, encyclopedic categories remain easier than popular-culture categories, and models still show variation across time periods and geographic regions.

M Model details

This section provides additional information about the models used in our experiments. Table 18 summarizes model-level metadata, including the provider region, reported training data scale, and language coverage. Table 19 provides the corresponding model references and URLs. When available, we rely on information reported in the model papers or official model cards, missing values indicate that the corresponding information was not explicitly reported by the model providers.

N GPU usage

For most evaluations described in this paper, we used a single Nvidia H100 80GB GPU. The exceptions were the evaluations of Apertus-70B-Instruct-2509 and Llama-3.1-70B-Instruct, for which we used two H100 cards. In total, evaluating the two 70B models took 30 minutes without the chat template, while evaluating the remaining 28 models took 3 hours. The same amount of time was required for the evaluations with the chat template.

The evaluation of search-augmented models, Llama-3.1-8B-Instruct and Qwen3.5-9B, also used one H100 GPU. It took 4 hours for Llama-3.1-8B-Instruct and 30 hours for Qwen3.5-9B.

O Experimental setup details

For evaluation, we used the LM Evaluation Harness (Gao et al., 2024)[14], for which we implemented two new tasks as YAML configuration files. All models were accessed through the Transformers library (Wolf et al., 2020). We evaluated each model once with a batch size of 16 and zero few-shot examples, under two conditions: with and without the chat template. Unless otherwise specified, all other parameters were left at their default values. We

[14]https://github.com/EleutherAI/lm-evaluation-harness

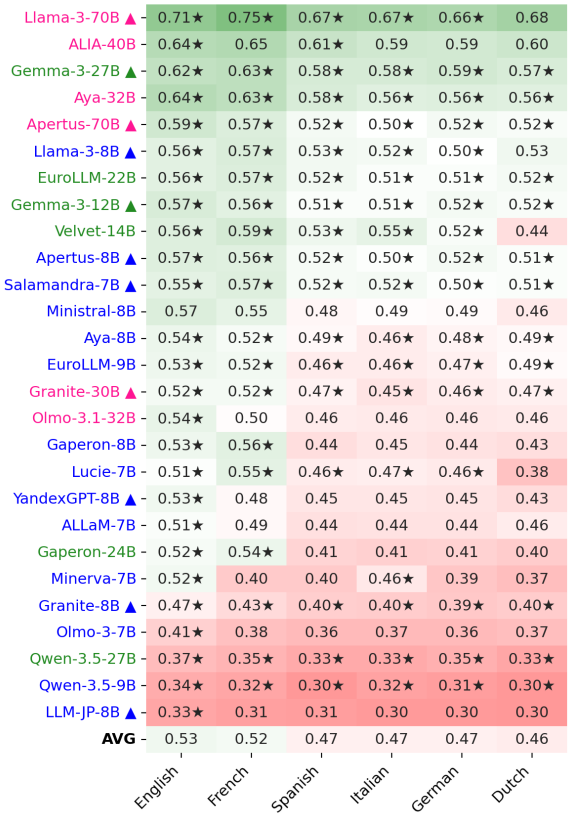

Figure 9: Model performance **with chat templates** on the six languages of TRIVIAROOMQA-MULTI. The reported metric is accuracy.

report average model's accuracy aggregated along different axes, including category, language, time period, continent, and difficulty level.

Experiments were run using LM Evaluation Harness v0.4.11 and Transformers v5.5.4.

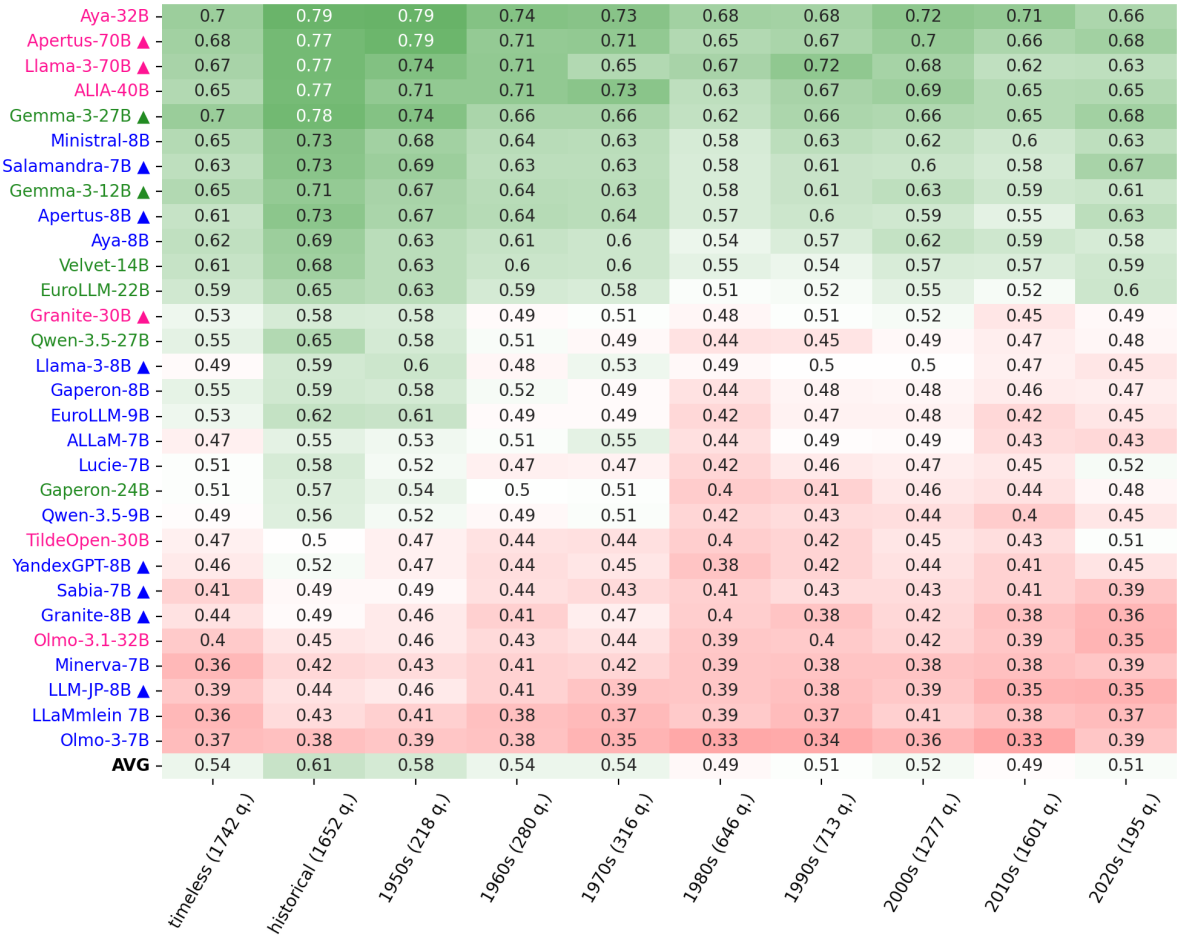

Figure 10: Model performance on TRIVIAROOMQA-FRENCH, grouped by the **time period** to which each question is most closely related, when applicable. The reported metric is accuracy.

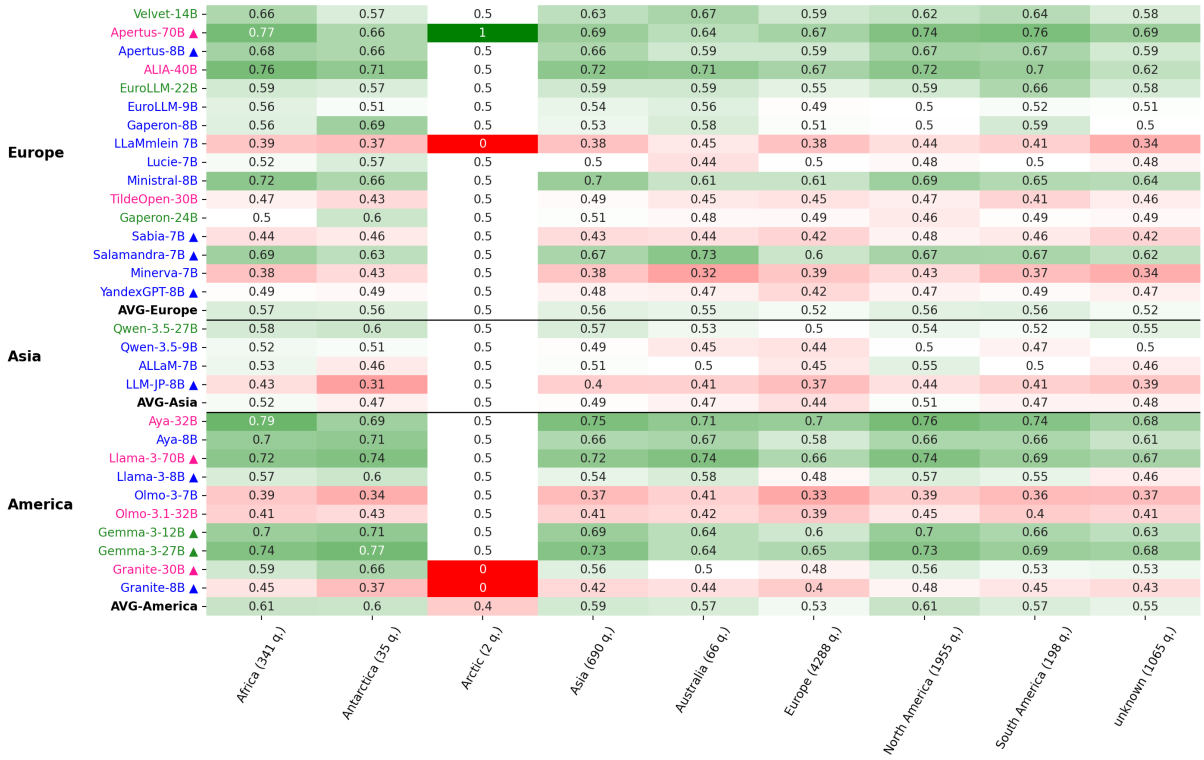

Figure 11: Model performance on TRIVIAROOMQA-FRENCH, grouped by the **continent** to which each question is most closely related, when applicable. Models are grouped by continent of origin, with the average performance for each group reported in the corresponding **AVG-{continent}** row. The reported metric is accuracy.

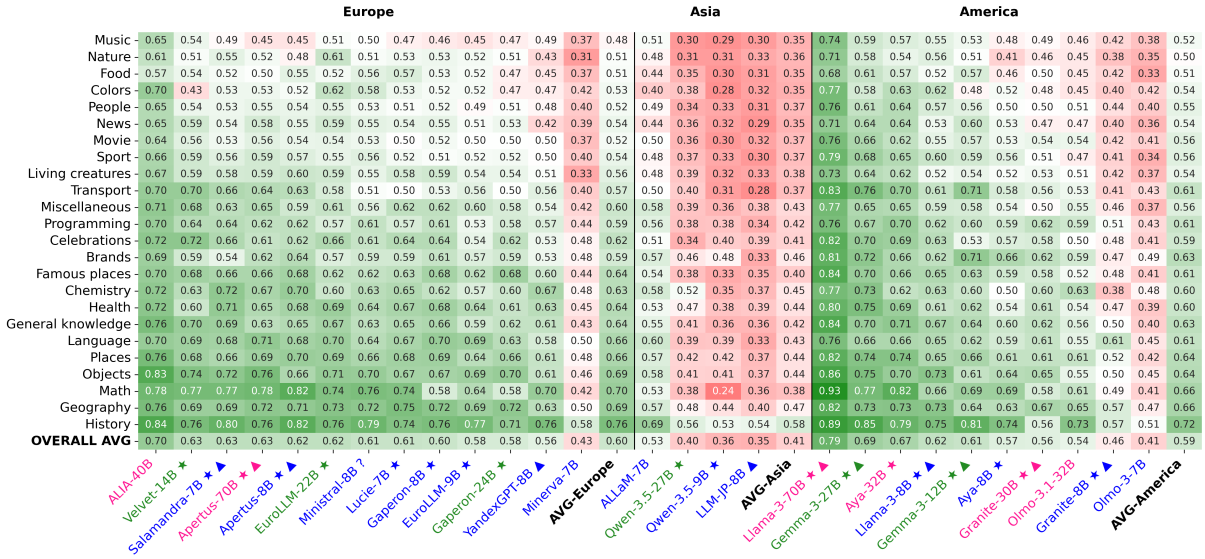

Figure 12: Model performance **with chat templates** on **24 categories** from TRIVIAROOMQA-FRENCH. The reported metric is accuracy. Models are sorted horizontally from left to right by performance within each region, and categories are sorted vertically from top to bottom by average performance across all models.

	timeless (1742 q.)	historical (1652 q.)	1950s (218 q.)	1960s (280 q.)	1970s (316 q.)	1980s (646 q.)	1990s (713 q.)	2000s (1277 q.)	2010s (1601 q.)	2020s (195 q.)
Llama-3-70B ▲	0.75	0.83	0.82	0.8	0.8	0.75	0.78	0.77	0.73	0.73
ALIA-40B	0.67	0.76	0.73	0.69	0.75	0.62	0.64	0.68	0.64	0.7
Gemma-3-27B ▲	0.65	0.73	0.72	0.68	0.68	0.61	0.65	0.64	0.65	0.7
Aya-32B	0.64	0.71	0.69	0.69	0.62	0.61	0.64	0.67	0.64	0.61
Velvet-14B	0.62	0.69	0.65	0.57	0.6	0.55	0.54	0.59	0.58	0.61
Gemma-3-12B ▲	0.57	0.66	0.67	0.56	0.59	0.55	0.6	0.58	0.58	0.63
Llama-3-8B ▲	0.59	0.68	0.67	0.6	0.6	0.54	0.59	0.58	0.54	0.57
Apertus-8B ▲	0.6	0.7	0.67	0.59	0.6	0.53	0.55	0.57	0.52	0.61
EuroLLM-22B	0.61	0.68	0.7	0.57	0.57	0.5	0.58	0.58	0.56	0.59
Apertus-70B ▲	0.6	0.68	0.67	0.57	0.57	0.54	0.58	0.58	0.57	0.53
Gaperon-8B	0.59	0.67	0.66	0.58	0.64	0.51	0.52	0.57	0.51	0.56
Salamandra-7B ▲	0.62	0.7	0.62	0.56	0.58	0.54	0.53	0.56	0.52	0.57
Ministral-8B	0.6	0.67	0.64	0.57	0.58	0.52	0.54	0.56	0.52	0.56
Lucie-7B	0.61	0.67	0.64	0.54	0.55	0.48	0.52	0.54	0.53	0.53
Gaperon-24B	0.57	0.66	0.6	0.56	0.58	0.47	0.51	0.54	0.5	0.57
EuroLLM-9B	0.56	0.66	0.62	0.53	0.53	0.5	0.49	0.54	0.48	0.54
Granite-30B ▲	0.55	0.61	0.6	0.56	0.58	0.5	0.53	0.53	0.48	0.49
Aya-8B	0.54	0.61	0.6	0.54	0.55	0.5	0.52	0.55	0.53	0.47
Olmo-3.1-32B	0.52	0.6	0.59	0.54	0.54	0.49	0.52	0.52	0.48	0.47
YandexGPT-8B ▲	0.53	0.63	0.57	0.52	0.53	0.5	0.51	0.51	0.44	0.44
ALLaM-7B	0.5	0.58	0.55	0.51	0.55	0.49	0.51	0.52	0.45	0.45
Granite-8B ▲	0.45	0.52	0.5	0.47	0.45	0.42	0.44	0.45	0.41	0.38
Minerva-7B	0.4	0.46	0.44	0.43	0.44	0.37	0.4	0.41	0.38	0.38
Olmo-3-7B	0.38	0.44	0.4	0.41	0.42	0.37	0.41	0.41	0.36	0.36
Qwen-3.5-27B	0.39	0.43	0.43	0.39	0.37	0.34	0.33	0.36	0.36	0.35
Qwen-3.5-9B	0.37	0.38	0.36	0.31	0.33	0.32	0.32	0.34	0.32	0.29
LLM-JP-8B ▲	0.35	0.38	0.38	0.33	0.34	0.34	0.32	0.29	0.29	0.29
AVG	0.55	0.62	0.6	0.54	0.55	0.5	0.52	0.53	0.5	0.52

Figure 13: Model performance **with chat templates** on TRIVIAROOMQA-FRENCH, grouped by the **time period** to which each question is most closely related, when applicable. The reported metric is accuracy.

		Africa (341 q.)	Antarctica (35 q.)	Arctic (2 q.)	Asia (690 q.)	Australia (66 q.)	Europe (4288 q.)	North America (1955 q.)	South America (198 q.)	unknown (1065 q.)
Europe	ALIA-40B	0.74	0.74	0.5	0.7	0.62	0.66	0.72	0.66	0.66
	Apertus-70B ▲	0.66	0.63	0.5	0.61	0.56	0.57	0.64	0.67	0.59
	Apertus-8B ▲	0.64	0.6	0.5	0.65	0.61	0.57	0.62	0.65	0.59
	EuroLLM-22B	0.64	0.66	0.5	0.64	0.58	0.59	0.6	0.62	0.58
	EuroLLM-9B	0.61	0.66	0.5	0.58	0.56	0.54	0.56	0.6	0.53
	Gaperon-24B	0.65	0.66	0.5	0.59	0.59	0.55	0.55	0.58	0.54
	Gaperon-8B	0.65	0.66	0.5	0.62	0.56	0.58	0.56	0.57	0.56
	Lucie-7B	0.64	0.6	0.5	0.59	0.53	0.57	0.56	0.59	0.57
	Minerva-7B	0.43	0.43	0.5	0.4	0.42	0.41	0.44	0.37	0.37
	Ministral-8B	0.61	0.69	0.5	0.63	0.58	0.55	0.61	0.62	0.57
	Salamandra-7B ▲	0.69	0.6	0.5	0.63	0.61	0.56	0.62	0.6	0.6
	Velvet-14B	0.7	0.66	0.5	0.66	0.59	0.58	0.63	0.67	0.59
	YandexGPT-8B ▲	0.6	0.66	0.5	0.58	0.53	0.48	0.58	0.54	0.53
	AVG-Europe	0.63	0.63	0.5	0.6	0.56	0.56	0.59	0.59	0.56
Asia	ALLaM-7B	0.55	0.57	0.5	0.56	0.48	0.47	0.59	0.49	0.49
	LLM-JP-8B ▲	0.35	0.43	0.5	0.34	0.41	0.32	0.34	0.34	0.34
	Qwen-3.5-27B	0.38	0.51	0.5	0.38	0.29	0.37	0.39	0.4	0.38
	Qwen-3.5-9B	0.33	0.4	0.5	0.35	0.38	0.34	0.35	0.34	0.34
	AVG-Asia	0.4	0.48	0.5	0.41	0.39	0.38	0.42	0.39	0.39
America	Aya-32B	0.72	0.69	0.5	0.7	0.74	0.63	0.7	0.65	0.61
	Aya-8B	0.62	0.57	0.5	0.58	0.53	0.52	0.59	0.6	0.53
	Gemma-3-12B ▲	0.65	0.69	0.5	0.64	0.61	0.57	0.64	0.61	0.56
	Gemma-3-27B ▲	0.73	0.83	0.5	0.7	0.59	0.64	0.72	0.66	0.65
	Granite-30B ▲	0.57	0.69	0.5	0.61	0.53	0.5	0.6	0.55	0.54
	Granite-8B ▲	0.5	0.54	0.5	0.5	0.47	0.42	0.5	0.52	0.45
	Llama-3-70B ▲	0.82	0.86	0.5	0.81	0.76	0.75	0.82	0.82	0.74
	Llama-3-8B ▲	0.65	0.66	0.5	0.66	0.67	0.56	0.66	0.62	0.55
	Olmo-3-7B	0.43	0.4	0.5	0.43	0.44	0.37	0.44	0.39	0.37
	Olmo-3.1-32B	0.6	0.6	0.5	0.57	0.64	0.48	0.6	0.58	0.51
	AVG-America	0.63	0.65	0.5	0.62	0.6	0.54	0.63	0.6	0.55

Figure 14: Model performance **with chat templates** on TRIVIAROOMQA-FRENCH, grouped by the **continent** to which each question is most closely related, when applicable. Models are grouped by continent of origin, with the average performance for each group reported in the corresponding **AVG-{continent}** row. The reported metric is accuracy.

Category	Topics
Brands	Brands, logos and slogans; Coca-Cola Company; Microsoft
Celebrations	Merry Christmas; Halloween; Saint Nicholas Day
Chemistry	Magnesium; Chemistry
Places	Canada; Italy; Nice; Belgium; United Kingdom; Rome; Brittany; Cities of the world; Vannes; Auroville; Périgord; Mouscron; City nicknames; Brussels today; Istanbul; Japan
Colors	Colors; Full of color
General knowledge	International knowledge; Knowledge and personalities; Folklore; Romanticism; Saints; General knowledge 2; General knowledge 3; General knowledge 4; General knowledge 5; Mixed knowledge; Mixed knowledge 2; Unbeatable; Youth knowledge; Mixed knowledge 3; Mixed knowledge 4; Mixed knowledge 5; Mixed knowledge 6; Japanese folklore; Sculpture; General knowledge
Famous places	The call of the open sea; Machu Picchu; Teotihuacan; World monuments; Castles and fortified castles; Central Park; Mont Saint-Michel; Tourist sites; Very long bridges; Louvre Museum; Chambord
Food	Desserts and pastries; Wines from elsewhere; Sugar; Coffee; Chocolate; Breakfast; Gin; Potato; Herbs and spices; Mineral waters; Belgian beers; Apples; Foreign gastronomy; French cheeses; Divine wines; Non-alcoholic drinks
Geography	Mediterranean; Geography for everyone; The Durance; Water stories; Antarctica; Active volcanoes; Forests of France; Long-distance hikes
Health	The human body; Health and well-being; The Omicron threat; Hair
History	Ancient Egypt; Political history; History of France; Gladiators; Wars and battles; Major dates of the 20th century
Language	Literary quotations; Famous expressions; Short quotations; Internet Frenglish; Spelling quiz; Crosswords; Crosswords 2
Living creatures	Sharks; Famous animals; Birds; Horses; Animals in numbers; Animals of all kinds; Beehive bees; Pigeon racing; Ants; Animals and habitats; Cats
Movie	Pokémon; SpongeBob SquarePants; Tintin; Movie robots; Television serials; Playboy; French comedies; Youth and cartoons; Star Trek; Secret Story; Disney princesses; Superheroines; Reality TV; César Awards ceremony; Alien: the saga; Fiction for everyone; Comedy films; Money Heist; Harry Potter; The X-Files; Breaking Bad; Marvel heroes; Game of Thrones; The Fifth Element; American series; The Big Blue; The Lion King; The Visitors; Jurassic Park; The Hunger Games; Toy Story turns 20; Star Wars heroes; Star Wars; Movie secrets; Imaginary characters
Music	Eighties bands; French variety music; The Cure; Disco hits; New wave; Reggae; Acid jazz
Nature	Fruit trees; Snow; Cactus; Japanese garden; Field fauna and flora
News	Celebrity news: February 2017; Celebrity news: January 2017; Celebrity news: March 2017; Celebrity news: April 2017; Celebrity news: May 2017; Celebrity news: June 2017; Celebrity news: August 2017; Celebrity news: September 2017; Social issues; Celebrity news: October 2017; Celebrity news: January 2018; Celebrity news: February 2018; Celebrity news: March 2018; Celebrity news: April 2018; Celebrity news: May 2018; Celebrity news: June 2018; Celebrity news: July 2018; Celebrity news: August 2018; Celebrity news: September 2018; Celebrity news: October 2018; It happened in 2019; COVID-19; Celebrity news: February 2021; Celebrity news: March 2021; Celebrity news: April 2021; Celebrity news: May 2021; Celebrity news: June 2021; 2021 retrospective; Celebrity news: August 2015; Celebrity news: February 2015; Celebrity news: January 2015; Celebrity news: September 2015
Math	Find the number; Questions about numbers; Units of measurement
Objects	Inventions; Objects and instruments; Makeup; Ceramics and pottery; Musical instruments
Miscellaneous	Global warming; Fashion victim; Mysteries of the world; Dragons yesterday and today; Board games; Languages of love; Peoples of the world
People	Mothers; Depeche Mode; Jenna Jameson; Clara Morgane; Donald Trump; Classical authors; Maria Sharapova; Mythical movie couples; Nikola Tesla; Steven Spielberg; Handsome men; Johnny Hallyday; Mike Horn; France Gall; Alan Turing; Global stars; Meghan Markle; Famous first names; WAGs; Traci Lords; Britney Spears; Brad Pitt in film; Pamela Anderson; French TV hosts; Florence Foresti; Charlize Theron; John McEnroe; Famous Alberts; Electro artists; Celebrity gossip 2013; Jean-Marie Bigard; Virginie on screen; Patrick Sébastien; Clint Eastwood; Celebrity gossip 2014; Cinema beauties; Bruce Willis; Sophie Marceau; French female directors; Jean Michel Jarre; Marilyn Monroe; Celebrities; Tutankhamun; Maxime Chattam; International singers
Programming	Geocaching; Internet; PlayStation 2; Cryptocurrencies; OpenBSD; Instagram; World of Warcraft; Software and web applications; Nintendo games and consoles; Linux; iPhone; Social networks
Sport	Rio 2016; Weightlifting; Boxing; Football of the past; Sports for everyone; Ironman; Ligue 1 stadiums; Climbing; PyeongChang 2018; Golf; Tennis; Russia 2018; Olympic Games; Football 2010–2020; Football 2000–2010; Football 1990–2000; Tennis court; Team sports; NBA players and franchises; European basketball; FC Barcelona
Transport	Means of transport; Car manufacturers; Automobile

Table 15: Categories and associated topics in TRIVIAROOMQA-FRENCH.

Continent classification prompt
You are a continent classifier. Task: Identify the continent most strongly associated with the quiz question. You may use: - the question - the candidate answers - the anecdote - the quiz topic/category Allowed outputs ONLY: - Africa - Antarctica - Arctic - Asia - Europe - North America - Australia - South America - unknown Rules: - Output ONLY one valid continent name from the allowed list. - If the question is not related to a specific continent, output: unknown - Never answer the quiz question itself. - Never explain. - Never ask for clarification. - Never output countries, cities, companies, objects, animals, or sentences. - Output exactly one value from the allowed list. Examples: Question: Quelle marque de jeans mondialement connue fabrique le célèbre 501 ? Choices: Levi's, Wrangler, Diesel, Lee Anecdote: Cette compagnie de vêtements est en effet mondialement connue, principalement pour son Blue-jeans. Topic: Marques et entreprises Continent: North America Question: De quelle filiale du groupe Nestlé George Clooney est-il devenu le séduisant ambassadeur ? Choices: Nespresso, Perrier, Maggi, KitKat Anecdote: Le fameux slogan 'What else ?', associé au beau George Clooney, a aujourd'hui été souvent parodié et a fait le tour du monde. Topic: Publicité Continent: Europe Question: Grâce à quoi les requins peuvent-ils respirer, de la même manière que les raies ? Choices: Leurs poumons, Leurs nageoires, Leurs branchies, Leur peau Anecdote: Les requins possèdent cinq à sept fentes branchiales sur les côtés de la tête. Topic: Animaux Continent: unknown Question: {context} Choices: {", ".join(answers)} Anecdote: {anecdote} Topic: {topic} Continent:

Table 16: Prompt used for continent classification.

Time-period classification prompt
You are a quiz question time-period classifier. Task: Identify the decade or era most strongly associated with the subject of the quiz question. You may use: - the question - the candidate answers - the anecdote - the quiz topic/category Allowed outputs ONLY: - 2020s - 2010s - 2000s - 1990s - 1980s - 1970s - 1960s - 1950s - historical - timeless Rules: - Output ONLY one valid value from the allowed list. - Select the period most directly associated with the main subject of the question. - Use the period during which the referenced person, event, product, company, work, trend, or phenomenon is most culturally relevant or famous. - For questions about discoveries, inventions, brands, celebrities, songs, movies, politics, sports, or technology, choose the decade most associated with them. - For questions about ancient history, the Middle Ages, Antiquity, monarchies, wars before 1950, mythology, or old civilizations, output: historical - For questions about biology, mathematics, language, geography, nature, physics, animals, or other knowledge not tied to a specific era, output: timeless - If multiple periods are possible, choose the one that is most commonly associated with the topic. - Never answer the quiz question itself. - Never explain. - Never ask for clarification. - Never output anything other than one allowed value. - Output exactly one value. Examples: Question: Quelle marque de jeans mondialement connue fabrique le célèbre 501 ? Choices: Levi's, Wrangler, Diesel, Lee Anecdote: Cette compagnie de vêtements est en effet mondialement connue, principalement pour son Blue-jeans. Topic: Marques et entreprises Time period: historical Question: De quelle filiale du groupe Nestlé George Clooney est-il devenu le séduisant ambassadeur ? Choices: Nespresso, Perrier, Maggi, KitKat Anecdote: Le fameux slogan 'What else ?', associé au beau George Clooney, a aujourd'hui été souvent parodié et a fait le tour du monde. Topic: Publicité Time period: 2000s Question: Grâce à quoi les requins peuvent-ils respirer, de la même manière que les raies ? Choices: Leurs poumons, Leurs nageoires, Leurs branchies, Leur peau Anecdote: Les requins possèdent cinq à sept fentes branchiales sur les côtés de la tête. Topic: Animaux Time period: timeless Question: Quelle ville française était la « cité des papes » au Moyen-Âge ? Choices: Metz, Nantes, Tours, Avignon Anecdote: C'est l'une des rares villes françaises à avoir conservé ses remparts et son centre historique, composé du palais des papes. Topic: Histoire de France (Paris vaut bien une messe) Time period: historical Question: {context} Choices: {", ".join(answers)} Anecdote: {anecdote} Topic: {topic} Time period:

Table 17: Prompt used for time-period classification.

Model name	# Training Tokens	Country	Languages
		Europe	
EuroLLM-9B-Instruct	4T	European Union	35 languages (Dutch, English, French, German, Italian, Spanish)
EuroLLM-22B-Instruct-2512	4T		
sabia-7b	11.4T	Portugal	English, Portuguese
ALIA-40b-instruct-2601	9.37T	Spain	Dutch, English, French, German, Italian, Spanish
salamandra-7b-instruct	12.875T	Spain	35 languages (Dutch, English, French, German, Italian, Spanish)
Lucie-7B-Instruct-v1.1	3T	France	English, French, German, Italian, Spanish
Ministral-8B-Instruct-2410	-	France	-
Gaperon-1125-8B-SFT	4T	France	English, French
Gaperon-1125-24B-SFT	2T		
sapienzanlp–Minerva-7B-instruct-v1.0	2.5T	Italy	English, Italian
Velvet-14B	4T	Italy	English, French, German, Italian, Portuguese, Spanish
Apertus-8B-Instruct-2509	15T	Switzerland	1800 languages (English, French, German, Italian, Portuguese, Spanish)
Apertus-70B-Instruct-2509	15T		
LLaMmlein-7B	3T*	Germany	German
TildeAI–TildeOpen-30b	2T	Lithuania	34 languages (Dutch, English, French, German, Italian, Spanish)
YandexGPT-5-Lite-8B-instruct	15.32T	Russia	English, Russian
		America	
Llama-3.1-8B-Instruct	15T	USA	8 languages (English, French, German, Italian, Spanish)
Llama-3.1-70B-Instruct	15T		
granite-4.1-8b	15T	USA	Multilingual (includes Dutch, English, French, German, Italian, Spanish)
granite-4.1-30b	15T		
gemma-3-12b-it	12T	USA	140 languages
gemma-3-27b-it	14T		
Olmo-3-7B-Instruct	6.05T	USA	English
Olmo-3.1-32B-Instruct	6.2T		
aya-expanse-8b	-	Canada	23 languages (Dutch, English, French, German, Italian, Spanish)
aya-expanse-32b	-		
		Asia	
ALLaM-7B-Instruct-preview	5.2T	Saudi Arabia	Arabic, English
llm-jp–llm-jp-4-8b-instruct	11.7T	Japan	English, Japanese
Qwen3.5-9B	-	China	200+ languages
Qwen3.5-27B	-		

Table 18: Training data details for the models used in the evaluation, including provider country and language coverage. *The authors of LLaMmlein (Pfister et al., 2025) do not explicitly report the training dataset size for LLaMmlein-7B, but state that the same training dataset was used for all model sizes (LLaMmlein 120M, 1B, 7B). We therefore report the number of training tokens given for the LLaMmlein-1B model.

Model name	URL
EuroLLM-9B-Instruct (Martins et al., 2025)	https://huggingface.co/utter-project/EuroLLM-9B-Instruct
EuroLLM-22B-Instruct-2512 (Ramos et al., 2026)	https://huggingface.co/utter-project/EuroLLM-22B-Instruct-2512
sabia-7b (Pires et al., 2023)	https://huggingface.co/maritaca-ai/sabia-7b
ALIA-40b-instruct-2601 (Gonzalez-Agirre et al., 2025)	https://huggingface.co/BSC-LT/ALIA-40b-instruct-2601
salamandra-7b-instruct (Gonzalez-Agirre et al., 2025)	https://huggingface.co/BSC-LT/salamandra-7b-instruct
Lucie-7B-Instruct-v1.1 (Gouvert et al., 2025)	https://huggingface.co/OpenLLM-France/Lucie-7B-Instruct-v1.1
Ministral-8B-Instruct-2410	https://huggingface.co/mistralai/Ministral-8B-Instruct-2410
Gaperon-1125-8B-SFT (Godey et al., 2025)	https://huggingface.co/almanach/Gaperon-1125-8B-SFT
Gaperon-1125-24B-SFT	https://huggingface.co/almanach/Gaperon-1125-24B-SFT
sapienzanlp–Minerva-7B-instruct-v1.0 (Orlando et al., 2024)	https://huggingface.co/sapienzanlp/Minerva-7B-instruct-v1.0
Velvet-14B	https://huggingface.co/Almawave/Velvet-14B
Apertus-8B-Instruct-2509 (Apertus et al., 2025)	https://huggingface.co/swiss-ai/Apertus-8B-Instruct-2509
Apertus-70B-Instruct-2509	https://huggingface.co/swiss-ai/Apertus-70B-Instruct-2509
LLaMmlein-7B (Pfister et al., 2025)	https://huggingface.co/LSX-UniWue/LLaMmlein_7B
TildeAI–TildeOpen-30b (Bergmanis et al., 2026)	https://huggingface.co/TildeAI/TildeOpen-30b
YandexGPT-5-Lite-8B-instruct	https://huggingface.co/yandex/YandexGPT-5-Lite-8B-instruct
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
Llama-3.1-70B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct
granite-4.1-8b	https://huggingface.co/ibm-granite/granite-4.1-8b
granite-4.1-30b	https://huggingface.co/ibm-granite/granite-4.1-30b
gemma-3-12b-it (Team, 2025)	https://huggingface.co/google/gemma-3-12b-it
gemma-3-27b-it	https://huggingface.co/google/gemma-3-27b-it
Olmo-3-7B-Instruct (Olmo et al., 2026)	https://huggingface.co/allenai/Olmo-3-7B-Instruct
Olmo-3.1-32B-Instruct	https://huggingface.co/allenai/Olmo-3.1-32B-Instruct
aya-expanse-8b (Dang et al., 2024)	https://huggingface.co/CohereLabs/aya-expanse-8b
aya-expanse-32b	https://huggingface.co/CohereLabs/aya-expanse-32b
ALLaM-7B-Instruct-preview (Bari et al., 2025)	https://huggingface.co/humain-ai/ALLaM-7B-Instruct-preview
llm-jp–llm-jp-4-8b-instruct	https://huggingface.co/llm-jp/llm-jp-4-8b-instruct
Qwen3.5-9B (Qwen Team, 2026)	https://huggingface.co/Qwen/Qwen3.5-9B
Qwen3.5-27B	https://huggingface.co/Qwen/Qwen3.5-27B

Table 19: Model references and URLs. When multiple model sizes come from the same paper, the citation is included only once.