AbstentionQA: Dataset for Training and Evaluating Multilingual LLMs to Abstain When Context Is Insufficient

Anonymous ACL submission

Abstract

Faithful use of provided context is essential for retrieval-augmented generation and document analysis, where language models should generate answers using in-context grounded evidence rather than unsupported prior (parametric) knowledge. This paper examines a common failure in medium-sized large language models: answering questions without relying on the supplied document context. We evaluate models on context-grounded question-answering tasks in which the context either supports the answer, lacks necessary evidence, or is paired with unrelated questions. We find that context adherence varies significantly across models and is affected by question–context relevance and the availability of answers from parametric memory. This issue persists even in synthetic settings with fictional entities, events, and places, where parametric knowledge can not provide answers. To identify possible causes, we analyse post-training data from three open models and observe limited evidence of faithful context use or explicit abstention when evidence is insufficient. Motivated by these findings, we introduce a multilingual contextual question-answering dataset that includes both answerable and unanswerable cases, with refusal responses grounded in the provided context.

1 Introduction

A central promise of large language models (LLMs) is their ability to turn unstructured text into usable information. In many practical deployments, users provide documents—policies, reports, contracts, technical manuals, internal knowledge-base data, or search results—and ask the model to answer questions about them. These document-grounded workflows, including retrieval-augmented generation and document analysis, require the model to answer based on the supplied context rather than on general world knowledge stored in parametric memory. The model should identify information supported by the document and abstain when the document lacks sufficient evidence. These requirements are especially relevant in privacy-sensitive organisational settings, where such workflows are often deployed locally to preserve data control and security. In these environments, organisations may rely on medium-sized models that can be served on-premise, making faithful context use a practical prerequisite for real-world deployment of such models.

A substantial body of work has introduced datasets for extractive, abstractive, and open-domain question answering, including SQuAD, TriviaQA, Natural Questions, HotpotQA, and related benchmarks (Rajpurkar et al., 2016; Joshi et al., 2017; Kwiatkowski et al., 2019; Yang et al., 2018; Rajpurkar et al., 2018). However, the available resources remain heavily skewed toward English, with comparatively fewer high-quality datasets for other languages. Existing multilingual resources either cover a limited set of languages, focus on languages with large speaker populations, or are derived through translation from English rather than being developed natively in the target languages (Lewis et al., 2019; Artetxe et al., 2019; Clark et al., 2020; d'Hoffschmidt et al., 2020; Efimov et al., 2020; Gutiérrez-Fandiño et al., 2022). This creates a data gap for many European languages, particularly in settings where the goal is not only to answer correctly, but to answer faithfully with respect to a provided document.

Faithfulness is not guaranteed by question answering accuracy alone. A model may produce a correct answer for the wrong reason: because the answer is already stored in its parametric memory rather than inferred from the context. The early work on language modelling has shown that pretrained language models encode substantial factual knowledge in their parameters and can perform closed-book question answering without access to

supporting documents (Petroni et al., 2019; Roberts et al., 2020). This creates an evaluation blind spot for context-grounded question answering. If a benchmark contains widely known real-world facts, a model can appear to use the provided context while in fact relying on memorised knowledge. The problem is particularly acute when evaluating whether a model should abstain. In realistic deployments, retrieved or user-provided documents are often topically related to the question but do not contain the specific evidence required to answer it. A reliable model should recognise this gap and refuse to answer; an unreliable model may instead combine partial context with prior knowledge and produce a plausible but unsupported response.

This paper studies this failure mode in medium-sized multilingual LLMs. We evaluate six models across 13 European languages in context-grounded question answering. Our evaluation distinguishes between answerable contexts, contexts from which answer-bearing evidence has been removed (i.e., contexts related to the question but lacking the evidence to answer the question), and contexts paired with unrelated questions. We further introduce synthetic Wiki-like evaluation data involving fictional people, places, and events. These synthetic examples are designed to make it impossible for the models to answer the questions from parametric memory, allowing us to directly test whether they use the provided context. Across settings, we find that models differ substantially in their ability to remain faithful to the context.

To investigate why this behaviour arises, we analyse the post-training data of three open models. We find that current post-training mixtures provide limited signal for faithful context analysis and evidence-sensitive abstention. In particular, models are often trained on examples where a context is relevant and a question is answerable, but receive less consistent supervision for cases where the context is related yet insufficient. This imbalance encourages models to treat question answering as a generation task rather than an evidence verification task. The result is a deployment risk: when a model is asked a question about a document, it may prioritise producing a helpful-sounding answer over verifying that the answer is supported by the document.

Motivated by these findings, we release AbstentionQA, a multilingual dataset for training context-grounded question answering and abstention capabilities. AbstentionQA contains curated answerable examples in which the context supports the target answer, as well as unanswerable examples paired with refusal-style responses that explain the context and state that the information is insufficient to answer the question. The dataset is constructed from TriviaQA questions and combined with multilingual Wikipedia contexts. After multi-stage LLM validation, answer elaboration, and refusal generation, the original English-language questions and answers are translated using a minimum Bayes-risk translation workflow (Kumar and Byrne, 2004). The released version contains 6,000 answerable context-question-answer examples and 3,000 unanswerable examples per language for 12 languages. Its purpose is to provide a direct training signal for the behaviour that our experiments show is missing: answering when the context supports an answer and abstaining when it does not.

Main contributions First, we provide a systematic multilingual evaluation of six medium-sized LLMs on context-grounded question answering across 13 European languages. Second, we use Wikipedia-derived evaluation sets to show that models often fail to condition faithfully on the supplied context, especially when questions can be answered from parametric memory or when the context is topically related but insufficient. Third, we introduce synthetic fictional evaluation sets that reduce contamination from pretrained world knowledge and make context reliance easier to measure. Fourth, we analyse post-training data from open models and identify limited supervision for evidence-sensitive abstention. Finally, we release AbstentionQA, a multilingual context-question-answer dataset designed to support both grounded answer generation and abstention in practical document-based question answering settings.

2 Model Evaluation

Models We evaluate six medium-sized open-weight multilingual LLMs that are practical to serve on a single on-premise host. **ALIA 40b** (Gon) is a 40B dense decoder-only model from the Barcelona Supercomputing Center, supporting EU languages but oriented toward Iberian languages and post-trained with supervised fine-tuning (SFT) only. **EuroLLM 22b** (Ramos et al., 2026) is a 22B model targeting all official EU languages, post-trained on the EuroBlocks-22B mixture with reward-model-based answer regeneration. **Apertus 70b** (Apertus Team, 2025) is a 70B compliance-

oriented model from EPFL and ETH Zürich, post-trained with SFT and aligned with QRPO. **Gemma 3 27b** (Gemma Team, 2025) is Google's 27B model trained via knowledge distillation from a Gemini teacher and reinforcement-learning fine-tuning. **Gemma 4 31b** (Gemma Team, 2026) is the 31B successor in the Gemma family[1]. **Mistral Small 3 24b** (Mistral AI, 2025) is the 24B latency-optimised instruct model from Mistral AI[2], reportedly trained without reinforcement learning or synthetic data.

Inference setup All six models were evaluated locally with vLLM (Kwon et al., 2023) on NVIDIA H200 GPUs using greedy decoding and 0.1 repetition penalty. Each example was presented to the model with a single user-turn prompt of the form

> `{context}`
> Answer the question based on the document context or abstain if the information is insufficient.
> Question: `{question}`

For non-English evaluation languages, the instruction sentence between the context and the question was translated into the corresponding target language; the document context and question were already in the target language by construction.

LLM-as-a-judge We scored model answers with OpenAI's `gpt-4.1`. For each item, the judge received the question, the reference answer, and the model's answer, and was tasked to return a single label drawn from {`1` = correct, `0` = wrong, `2` = abstention}. We asked the judge to evaluate semantic equivalence, ignore stylistic differences and paraphrasing, mark yes/no flips as wrong, and treat hedged or non-committal responses as abstentions. The full system prompt and user message template are reproduced verbatim in Appendix A.

2.1 Evaluation Set Construction

Wikipedia-derived Evaluation Sets We constructed a multilingual, context-grounded question-answering dataset based on the rag-datasets/rag-mini-wikipedia corpus. The source dataset contains 918 question-answer pairs associated with Wikipedia articles[3]. We first created a cleaned English seed set of question-answer-context triples. We removed examples with missing or null fields and deduplicated repeated question-answer-context instances. We also manually revised underspecified questions in which the referent was expressed solely by a pronoun, replacing the pronoun with the corresponding named entity from the associated article. This process ensured that each question could be interpreted unambiguously and independently of the document title, while preserving its original answer. After filtering and revision, the final seed set comprised 794 question-answer-context triples. From this seed set, we established three evaluation conditions. First, we retained the original documents paired with their corresponding questions, resulting in an answerable condition (**Wiki question answering**) where the answer is supported by the provided document context. Second, we generated an unanswerable by ablation condition (**Wiki abstain**) by identifying and removing the paragraphs that contain the answer to the question using the Gemma 3 27b model. We subsequently conducted human validation on a random sample of 50 ablated examples. The human annotator and Gemma 3 agreed on all 50 cases. Third, we created an unrelated-question (**Wiki abstain: unrelated question**) condition by pairing document fragments with questions drawn from unrelated documents. These examples were also constructed to be unanswerable.

Synthetic Evaluation Sets During preliminary experiments, we observed that models sometimes answered questions from parametric memory rather than from the supplied context. This behaviour can inflate performance in the answerable setting, as models may produce the correct answer even when they do not rely on the provided document. Conversely, it can reduce performance on abstention in the unanswerable setting, as the models may answer based on memorised world knowledge even with insufficient evidence. This concern is consistent with prior work suggesting that language models may rely in part on internal knowledge-linked pathways when generating answers, such that correct responses do not necessarily indicate faithful use of the supplied context (Luo et al., 2026). To isolate context-grounded behaviour from parametric knowledge, we therefore constructed an additional synthetic evaluation set based on fictional Wiki-like articles.

We generated synthetic articles using Claude Opus 4.5. The model was prompted to write Wiki-like narratives about fictional personae, events, and

[1] https://huggingface.co/google/gemma-4-31B-it
[2] https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501
[3] https://huggingface.co/datasets/rag-datasets/rag-mini-wikipedia

places, set in arbitrary locations and time periods. Seed Wikipedia articles were used only to provide stylistic and structural guidance for the generated articles; the generated entities and events were fictional. This design ensured that the resulting questions referred to information unlikely to be present in a model's pretraining data, thereby reducing models' ability to answer correctly solely from parametric knowledge.

For each synthetic article, we prompted Claude Opus 4.5 to generate simple fact-seeking questions of the form *what*, *when*, *who*, and *where*, together with their corresponding answers. We then prompted the same model to classify each paragraph in the associated article as either relevant or irrelevant for answering each question. These paragraph-level relevance labels were used to construct two evaluation settings analogous to the Wikipedia-based settings. In the synthetic question answering (**Synth. question answering**) setting, questions were paired with article contexts containing the answer. In the synthetic abstention (**Synth. abstain**) setting, questions were paired with contexts from which the answer-supporting paragraphs had been removed, making the question unanswerable from the supplied context. The final synthetic dataset contained **500 question-answer pairs** for each setting.

Translation We then translated all English evaluation conditions into the target languages using the **[Anonymised] MT model**. To improve translation quality, we generated 10 candidate translations per item and applied Minimum Bayes-Risk (MBR) decoding, which selects the hypothesis with the highest expected utility, under a selected utility function (Kumar and Byrne, 2004). Questions, answers, and contexts were translated jointly to preserve cross-field consistency, but candidate selection was performed separately for each field. For questions and answers, we selected candidates using `Unbabel/wmt22-comet-da` as the utility model for MBR decoding. This model is a reference-based COMET evaluation model trained on direct assessment data (Rei et al., 2022; Unbabel). In the MBR setting, candidate translations were compared against the sampled candidate set, allowing the selected hypothesis to reflect the translation with the highest expected quality under the COMET-based utility function. For document contexts, we did not score the passages with COMET because they exceed the sentence-level regime for which the metric was trained. Instead, we used averaged paragraph-level ChrF++ as the MBR utility function (Popović, 2017).

Use of LLMs We use Gemma 3, [Anonymised] MT model, Comet, and Claude Opus 4.5 for evaluation data curation. This is a permitted use of these tools.

2.2 Results

Question Answering Figure 1 presents evaluation results on Wikipedia-derived and synthetic datasets. Figure 1 (a) indicates that all models except Alia and Apertus achieve over 90% accuracy on question answering tasks using the Wikipedia-derived evaluation sets. Figure 1 (d) demonstrates a similar pattern, indicating that most models effectively answer questions even when prior knowledge is insufficient and reliance on document context is required. However, ALIA exhibits a 16% decrease in average accuracy, indicating difficulty in generating answers consistent with the document context. In both evaluation settings, all models perform best on English, though the margin is only approximately 3%.

Abstention Figure 1 (b) highlights significant differences in the models' ability to abstain. All European models exhibit extremely low abstention rates when required to answer questions based on insufficient context. The lowest-performing models, Alia and Apertus, provide unwarranted answers in over 90% of cases, while Mistral and EuroLLM do so in approximately 75% of cases. The best-performing model, Gemma 4, achieves an average abstention rate of 73.8%. These findings indicate that most models rarely consult the document context when questions pertain to topics encountered during training and instead generate answers based on parametric knowledge.

Figure 1 (c) demonstrates that abstention is substantially easier for all models when documents and questions concern real but mutually unrelated topics. For most models, average abstention rates exceed 80%. However, Alia and Apertus continue to perform poorly, with rates of 38.9% and 54.3%, respectively.

Figure 1 (e) compares models in a scenario where questions are related to the documents, but prior knowledge is neither beneficial nor detrimental because the topics are entirely fictional. In this context, the best-performing models, Gemma 4 and Mistral, achieve an abstention rate of nearly

82%. Although this represents an improvement over scenarios involving Wikipedia documents, the abstention rate remains insufficient. Practically, these models still generate unwarranted answers for approximately one in five questions.

3 Post-Training Data Analysis

We analyse the textual post-training datasets used by ALIA, Apertus, and EuroLLM to understand where, if at all, models receive their contextual evidence-grounded question-answering training signal. The post-training datasets can be grouped according to the behaviour they are likely to induce. We find three broad categories:

- *assistant and instruction-following data* that train models to follow user instructions, answer general questions, summarise, rewrite, classify, brainstorm, and sustain multi-turn dialogue (Conover et al., 2023; Taori et al., 2023; Köpf et al., 2024; Teknium, 2023; Teknium et al., 2024; Xu et al., 2024; Lambert et al., 2025; Wang et al., 2024; Singh et al., 2024; Dang et al., 2024);
- *capability-specialized data* improving STEM, coding, function calling, table understanding, scientific-literature understanding, and domain-specific financial or web-oriented tasks (NVIDIA, 2025; Nathawani et al., 2025b,a; OLMo Team, 2025; Arcee AI, 2024);
- *multilingual data* that train translation capabilities (Team et al., 2022; Alves et al., 2024).

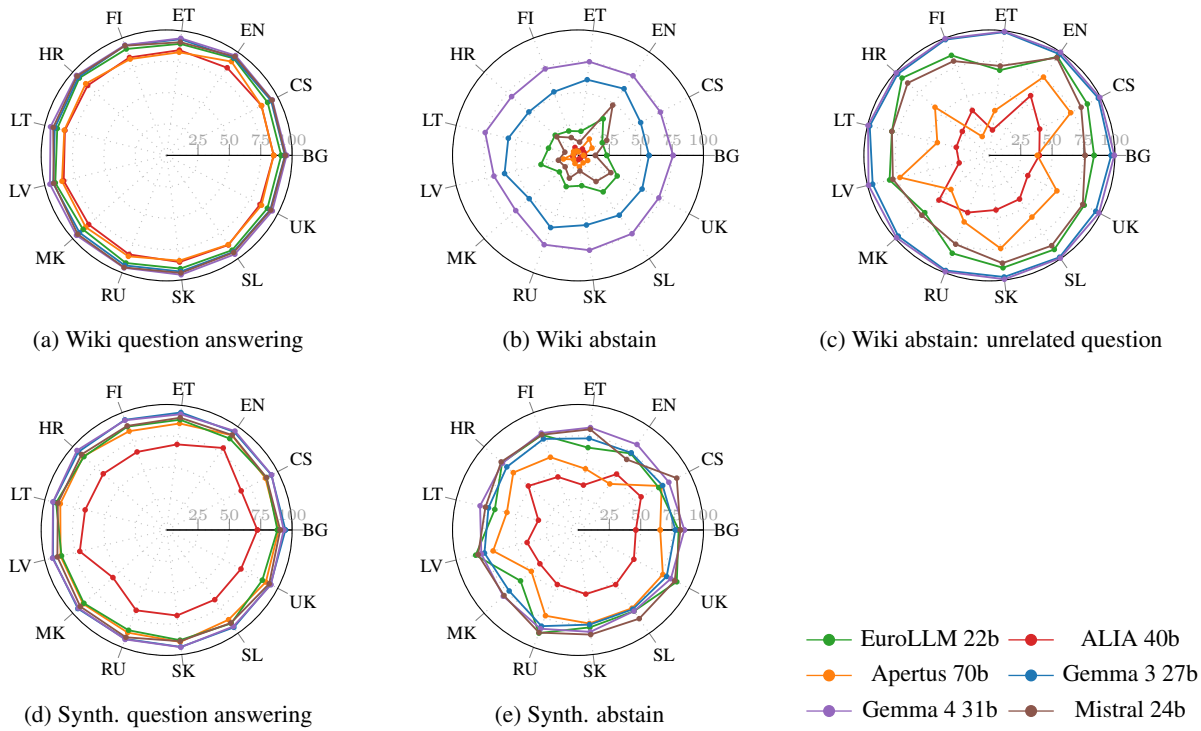
Only a subset of these sources directly trains models to answer questions from a supplied document context, and an even smaller subset provides a signal for abstention when the supplied evidence is insufficient.

ALIA 40b's contextual question-answering data comprises RAG Multilingul, CoQCat, MentorCA, MentorES, Dolly(Conover et al., 2023) and Dolly3k-Ca[4]. RAG Multilingual is the strongest RAG-oriented source: it is a synthetic multilingual QA dataset constructed from extractive QA datasets in Catalan, English, and Spanish, where each instance includes a question, a context, a generated answer, and a span from the context supporting it. This is direct training for grounded generation over retrieved passages. However, there are no negative examples that would mandate abstention. CoQCat is also strongly context-grounded: it is a Catalan conversational QA dataset based on CoQA (Reddy et al., 2019), with question-answer pairs linked to text passages and evidence spans. MentorCA, MentorES, and Dolly-derived Dolly3k-CA provide a weaker but still relevant signal, as subsets of their training examples include context fields that provide grounding for information extraction, summarisation, and question-answering tasks. We take this as evidence that ALIA receives substantial training for answering from the provided context, but little clearly documented training for refusing to answer when the context does not contain the answer.

Apertus 70b post-training data, among others, include contextual question answering datasets such as WikiQA (Yang et al., 2015), SciRIFF (Wadden et al., 2025), TableGPT (Li et al., 2024). SciRIFF trains instruction-following over scientific literature, including question answering, summarisation, claim verification, and information extraction from long scientific inputs, while TableGPT targets table-centric question answering. While this data might help context-based question answering, their inspection suggests that they are answerable by construction. The original form of the WikiQA, on the other hand, requires answer selection choosing among several sentences of which none might be the correct answer. While the Apertus technical report does not document how authors use this dataset, it could provide *some* evidence-based abstention signal. CoCoNot (Brahman et al., 2024), which trains a model to generalise to unsafe request refusal behaviour. Together, these datasets enable context-based question answering, but the abstention behaviour is not explicitly trained.

EuroLLM-22B's EuroBlocks-22B post-training corpus is built from OpenHermes 2.5, Aya Expanse, HelpSteer 2, Magpie, Hermes-3, Tülu 3 and Nemotron post-training data (Teknium, 2023; Dang et al., 2024; Wang et al., 2024; Xu et al., 2024; Teknium et al., 2024; Lambert et al., 2025; Nathawani et al., 2025b,a). This data is broad and helps induce multilingual assistant behaviour, reasoning, STEM assistance, coding, rewriting, and translation. It also likely contains some contextual QA examples, as several source mixtures include reading comprehension and contextual

[4] https://huggingface.co/collections/projecte-aina/text-datasets



(a) Wiki question answering (b) Wiki abstain (c) Wiki abstain: unrelated question

(d) Synth. question answering (e) Synth. abstain

Figure 1: Evaluation results on Wikipedia-derived and synthetic datasets indicating question answering accuracy (a), (d) and abstention rates (b), (c), (d) for six medium-sized LLMs.

question-answering. However, we find no clear evidence in the EuroLLM report for a dedicated evidence-abstention data comparable to WikiQA or SQuAD 2.0.

The conclusion is that these models are trained much more heavily for *grounded answering* than for *evidence-based abstention*. In particular, RAG Multilingual, CoQCat, MentorCA/ES, Dolly, SciRIFF, TableGPT, and WikiQA all provide signals that the model should use a supplied passage, table, or document-like context. However, most of these datasets are answerable-by-construction: the answer is present in the context; the target response demonstrates how to extract, paraphrase, or synthesise it. This trains context use, but it does not necessarily train the model to recognise the absence of evidence.

4 AbstentionQA Dataset

4.1 Dataset Curation

We created AbstentionQA as a multilingual dataset for in-context question answering, where each example pairs a question with one or more retrieved document contexts and a target answer that should be generated only when the context provides sufficient evidence. The dataset's construction pipeline starts from a preprocessed version of TriviaQA (Joshi et al., 2017) and augments it with multilingual Wikipedia contexts, LLM-based context validation, answer elaboration, refusal generation, and machine translation.

We obtained seed question-answer pairs from the deduplicated English TriviaQA train set. For each target language, we constructed a searchable Wikipedia corpus. We segmented full Wikipedia dumps into non-overlapping chunks of up to 8,192 characters. We then used the multilingual `bge-m3` embedding model (Chen et al., 2024) to embed each chunk into 1024-dimensional dense vectors. Vectors and associated metadata were indexed in language-specific Typesense instances (Typesense Inc., 2024). For every English TriviaQA question–answer pair, we concatenated the question and answer into a single retrieval query, embedded it with the same model, and retrieved the top five nearest Wikipedia chunks from each target-language index. This yielded candidate question–answer–context triplets.

We used a two-step LLM validation procedure to determine whether the retrieved context contained sufficient evidence to support all factual claims needed to answer the question. To do this, we used Gemma 3 27B as a local high-throughput filter followed by Gemini 2.5 Flash (Google DeepMind,

2025). We prompted the first model to classify each candidate context as either supporting the answer to a given question or not. The second model was used to validate the correctness of the positive labels of the initial model till we obtained a sufficiently large set of question-answer tuples paired with 5 contexts each. We split the resulting data into two subsets: the answerable subset, for which for each question-answer tuple there was at least one context that supported the answer, and the unanswerable subset, for which none of the contexts supported the answer.

For the answerable subset, we first converted the short English TriviaQA answers into full-sentence answers using Gemma 3 27b. This step was necessary because TriviaQA answers are often short spans or named entities, whereas instruction-following question answering typically requires natural, standalone responses. The model was prompted to preserve the factual content of the original answer while incorporating enough of the question semantics to form a complete sentence.

For the unanswerable subset, we generated refusal-style answers in English using Gemma 3. The generated refusal answers were designed to be informative rather than generic: each refusal first briefly describes what the provided document context is about, and then explicitly concludes that the information in the context is insufficient to answer the question. This format encourages models not only to abstain but also to ground their abstention in the supplied evidence.

All answerable and unanswerable examples were then translated into target languages using the **[Anonymised] MT model**. We used the same minimum Bayes-risk decoding procedure as for our evaluation set and translated the English questions, expanded answers, and refusals into the languages of the Wikipedia contexts.

After manual inspection of the resulting translations, we noticed examples in which answers had diverged lexically from the terminology used in the Wikipedia contexts. An example that illustrates the problem is the English *Grand Duchy of Lithuania*, which in Latvian can be translated as either the *Lietuvas lielkņaziste* or the *Lietuvas dižkunigaitija*. However, answering using one when the contexts provided used the other would be unnatural. To detect and filter such cases, we used Stanza and Classla (Qi et al., 2020; Terčon et al., 2025) to apply lemmatisation-based checks that verify whether the lemma of the translated answer appeared in the context.

Finally, we conducted an independent two-annotator evaluation on 100 Latvian and 100 Russian examples that had previously been labelled as answerable. Inter-annotator agreement was moderate, with Cohen's $\kappa = 0.5588$ for Latvian and $\kappa = 0.5240$ for Russian (Cohen, 1960). Disagreements typically involved borderline cases requiring synonymy, world knowledge, cross-lingual reasoning, or overly specific question constraints.

Unlike the broader intermediate corpus, the released dataset contains a curated subset of 6,000 answerable context–question–answer examples and 3,000 unanswerable examples for each of the following twelve languages: Bulgarian, Croatian, Czech, Estonian, Finnish, Latvian, Lithuanian, Macedonian, Russian, Slovak, Slovenian, and Ukrainian.

Use of LLMs We use Gemma 3, Gemini 2.5 Flash, [Anonymised] MT model and Comet for AbstainQA curation. This is a permitted use of the open-weight tools. As for Gemini 2.5 Flash, its licence permits its use for dataset curation if the downstream models are not trained directly on Gemini's outputs, which holds because the dataset does not contain anything generated by Gemini.

4.2 Dataset Validation

To validate that AbstentionQA provides a useful training signal for evidence-sensitive abstention, we fine-tune two variants of the same base model and evaluate them under the same protocol used for the six models above. The base model is **[Anonymised] 30B**, a 30B-parameter multilingual decoder-only model with a 64k context window covering the same European languages as our evaluation. The two variants differ only in that the baseline post-training mixture does not contain AbstentionQA, while the other model variant does.

The baseline-mix consists of public datasets: RAG-bench, which comprises 70k examples of mostly answerable questions with a very small fraction of abstentions, 45k SQuAD 2.0-derived abstention examples. Additionally, we use multilingual internal data comprising 150k contextually answerable questions sourced from EMEA, EUR-LEX, TED Talks, restaurant menus, flat and house listings, car manuals, and other sources.

We instruction-tune each variant in a supervised fashion within the GPT-NeoX framework, using 512 AMD MI250x GPUs on the LUMI supercomputer. Training is run for $\sim$3000 updates with our

final cooldown learning rate of $8 \cdot 10^{-8}$ and a batch size of approximately 1M tokens. The total budget of these experiments was about 2000 GPUh

4.3 Results

Table 1 reports ablation results for fine-tuned variants of [Anonymised] 30B trained with and without the AbstentionQA dataset. Averaged over 13 languages, adding AbstentionQA substantially improves abstention behaviour, increasing abstention rates by up to 70 percentage points while reducing question-answering accuracy by only 2.9 percentage points. Compared with Gemma 4 31b, the AbstentionQA-trained model achieves abstention rates up to 7.7 percentage points higher, while maintaining comparable, albeit slightly lower, question-answering accuracy. We take these results as an indication that the dataset does provide a useful training signal for abstention behaviour when questions are not answerable from the provided context, while preserving comparable question-answering performance.

Cross-Lingual Transfer AbstentionQA includes examples for training abstention for 12 languages, but not for English, for which we use SQUAD 2 and RAG Bench examples as available baseline datasets. Zooming in on English results for [Anonymised] 30B trained without the AbstentionQA dataset, we see meagre performance of 28.7% and 41.2% on *Wiki abstain* and *Synth. abstain* respectively. While the model trained using AbstentionQA shows higher abstention rates, (44.7% and 75.1%), they are still drastically lower than the model's averages (80.1% and 90.6%) on the other 12 languages. These results suggest that, at least as far as these experiments are concerned, the abstention behaviour learned by the model is to some extent language-specific: while adding abstention examples in 12 other languages increased performance on English, the results are much lower than they would have been if similar examples had been added for English.

Setting	Basel.	+AbsQA	Gemma 4
Wiki QA	92.3%	86.1%	95.5%
Wiki abstain	24.2%	77.4%	73.8%
Wiki abstain: unrel.	64.9%	99.2%	99.4%
Synth. QA	92.9%	90.0%	93.9%
Synth. abstain	18.4%	89.4%	81.7%

Table 1: Ablation of AbstentionQA on the [Anonymised] 30B base. Each cell is the per-task accuracy/abstention rate averaged over 13 languages. **Basel.** [Anonymised] 30B post-trained without AbstentionQA. **+AbsQA**: same recipe with AbstentionQA included. Gemma 4 31b is shown as the strongest baseline.

5 Conclusions

We conducted a systematic study of context-grounded question answering and evidence-based abstention in medium-sized multilingual LLMs. Evaluating six models across 13 European languages and five settings, we found that most models do not reliably use the provided document context. Although question-answering accuracy is generally high, abstention rates remain low when context is insufficient. Some models generate unsupported answers in over 90% of unanswerable cases. This pattern persists even with synthetic data involving fictional entities, where correct answers cannot be retrieved from parametric knowledge. These results suggest a general tendency to prioritise answer generation over evidence verification, rather than reliance on memorised world knowledge.

Analysis of post-training data from three recent European open LLMs shows that their training mixtures are biased toward examples that are answerable. We find that models receive very little signal for learning to recognise insufficient evidence. This imbalance likely explains the observed abstention failures.

Based on these findings, we introduced AbstentionQA, a multilingual dataset comprising 6,000 answerable and 3,000 unanswerable context–question–answer examples for each of the 12 European languages. Unlike in other datasets, contexts are drawn from native-language Wikipedia documents. Unanswerable examples include informative refusal responses that summarise the context and state that it is insufficient to answer the question. Fine-tuning a 30B-parameter model with AbstentionQA significantly increased abstention rates by up to 70 percentage points, while reducing question-answering accuracy by only 2.9 percentage points. This approach outperformed the best evaluated baseline on abstention tasks. We also observed limited cross-lingual transfer of abstention behaviour to English, indicating that language-specific training signals remain important, even for related tasks.

All research artefacts from this work are publicly released at `https://huggingface.co/[Anonymised]` under CC-BY-4.0.

6 Limitations

AbstentionQA Dataset This work focuses specifically on context-grounded question answering and evidence-based abstention in isolation relative to other tasks. In particular, we do not evaluate whether training on AbstentionQA affects non-QA behaviours, such as summarisation, translation, open-ended instruction-following or dialogue quality. A model could improve substantially on abstention benchmarks while becoming more conservative or less useful in other settings. Work using the AbstentionQA dataset, therefore, should evaluate these aspects independently.

Furthermore, the dataset validation experiment is performed on a single 30B base model and one supervised fine-tuning recipe. While computationally costly, additional experiments with larger and smaller models of different origins would better illustrate the dataset's benefits and limits.

The dataset construction pipeline relies on machine translation, MBR decoding, automatic validation, and lemmatisation-based filtering. These steps and techniques have reduced the cost of creating the dataset and improved its consistency, but can not eliminate all translation artefacts and terminology mismatches.

LLM evaluation experiments cover six medium-sized open-weight multilingual LLMs using a single prompt format and greedy decoding. While our setup is deliberately simple to avoid unjustified prompts, decoding parameters, retrieval formats, context-ordering strategies, or instruction templates, these aspects may affect the observed balance between answering and abstaining.

Our results characterise a simplified form of context adherence. The document contexts we use are from Wikipedia or are Wiki-like, and the questions are mostly simple fact-seeking questions of the form *what*, *when*, *who*, and *where*. Real document-grounded QA often involves longer and noisier inputs, multi-hop evidence, tables, data extraction artefacts, domain-specific terminology, contradictory passages, partially relevant retrieved documents, and questions whose answerability is genuinely ambiguous.

Cross-lingual Transfer AbstentionQA is released for 12 European languages, and the ablation experiment focuses on these languages and English. The results showed limited cross-lingual transfer from the 12 languages to English. This finding is underexplored in this work, but in practice, it is crucial when developing end-user-facing LLMs and workflows. Due to the restricted multilingual scope, additional data may be needed for other languages.

References

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Apertus Team. 2025. Apertus: Democratizing open and compliant llms for global language environments. *Preprint*, arXiv:2509.14233.

Arcee AI. 2024. The tome.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. The art of saying no: Contextual noncompliance in language models. *Preprint*, arXiv:2407.12043.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. volume 20, pages 37–46.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

John Dang, Shivalika Singh, Daniel D'Souza, and 1 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. Fquad: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208.

Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2020. Sberquad–russian reading comprehension dataset: Description and analysis. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 3–15. Springer.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team. 2026. Gemma 4 model card. Google AI for Developers. HuggingFace: `google/gemma-4-31B-it`; accessed: 2026-05-07.

Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68(0):39–60.

Mandar Joshi, Eunsol Choi, Daniel S. 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*, pages 1601–1611. Association for Computational Linguistics.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, and 1 others. 2024. Openassistant conversations: Democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP '23)*. ACM.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2025. Tülu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*, arXiv: 1910.07475.

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2024. Table-gpt: Table fine-tuned gpt for diverse table tasks. *Proceedings of the ACM on Management of Data*, 2(3):1–28.

Wen Luo, Guangyue Peng, Wei Li, Shaohang Wei, Feifan Song, Liang Wang, Nan Yang, Xingxing Zhang, Jing Jin, Furu Wei, and 1 others. 2026. Two pathways to truthfulness: On the intrinsic encoding of llm hallucinations. *arXiv preprint arXiv:2601.07422*.

Mistral AI. 2025. Mistral Small 3. Blog post. HuggingFace: `mistralai/Mistral-Small-24B-Instruct-2501`; accessed: 2026-05-07.

Divy Nathawani, Simon Ding, Vitaly Lavrukhin, and 1 others. 2025a. Nemotron-post-training-dataset-v2.

Divy Nathawani, Igor Gitman, Somshubra Majumdar, and 1 others. 2025b. Nemotron-post-training-dataset-v1.

NVIDIA. 2025. Llama-nemotron-post-training-dataset.

OLMo Team. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2463–2473. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: Words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human

languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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, and 1 others. 2026. Eurollm-22b: Technical report. *Preprint*, arXiv:2602.05879.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Ricardo Rei, José G. C. De Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 578–585. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5418–5426. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, and 1 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Teknium. 2023. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants.

Teknium, Jeffrey Quesnelle, and Chen Guang. 2024. Hermes 3 technical report. *Preprint*, arXiv:2408.11857.

Luka Terčon, Kaja Dobrovoljc, and Nikola Ljubešić. 2025. Classla-stanza: The next step for linguistic processing of south slavic languages. *Contributions to Contemporary History/Prispevki za Novejšo Zgodovino*, 65(3).

Typesense Inc. 2024. *Typesense Search Engine API Reference*. Typesense Inc. Open source search engine API documentation.

Unbabel. Unbabel/wmt22-comet-da. Hugging Face model card. Accessed: 2026-04-29.

David Wadden, Kejian Shi, Jacob Morrison, Alan Li, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, Doug Downey, Hannaneh Hajishirzi, and Arman Cohan. 2025. SciRIFF: A resource to enhance language model instruction-following over scientific literature. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6072–6109, Suzhou, China. Association for Computational Linguistics.

Zhilin Wang, Yi Dong, Olivier Delalleau, and 1 others. 2024. Helpsteer 2: Open-source dataset for training top-performing reward models. In *NeurIPS Datasets and Benchmarks Track*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, and 1 others. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *Preprint*, arXiv:2406.08464.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

A LLM-as-a-Judge Prompts

The grading prompt and user message template used with OpenAI's `gpt-4.1` judge are reproduced verbatim below.

System prompt.

```
You are a strict grader.
Given a Question, Reference Answer, and
Model Answer, return ONLY a single digit:
- 1  correct (main fact matches reference;
     paraphrases/synonyms ok)
- 0  wrong   (main fact differs)
- 2  abstention (model refuses, is unsure,
     or gives no answer)

Rules: judge semantic equivalence; ignore
style/wording; yes/no flip -> 0;
hedging -> 2.
Output: a single character: 0, 1, or 2.
```

User message template.

```
Question:
{question}

Reference Answer:
{reference_answer}

Model Answer:
{model_answer}
```