OMoS-QA: A Dataset for Cross-Lingual Extractive Question Answering in a German Migration Context

Steffen Kleinle[1,2] **Jakob Prange**[1] **Annemarie Friedrich**[1]
[1]University of Augsburg, [2]Tür an Tür Digitalfabrik GmbH
Contact: {firstname.lastname}@uni-a.de

Abstract

When immigrating to a new country, it is easy to feel overwhelmed by the need to obtain information on financial support, housing, schooling, language courses, and other issues. If relocation is rushed or even forced, the necessity for high-quality answers to such questions is all the more urgent. Official immigration counselors are usually overbooked, and online systems could guide newcomers to the requested information or a suitable counseling service.

To this end, we present OMoS-QA, a dataset of German and English questions paired with relevant trustworthy documents and manually annotated answers, specifically tailored to this scenario. Questions are automatically generated with an open-weights large language model (LLM) and answer sentences are selected by crowd workers with high agreement. With our data, we conduct a comparison of 5 pretrained LLMs on the task of extractive question answering (QA) in German and English. Across all models and both languages, we find high precision and low-to-mid recall in selecting answer sentences, which is a favorable trade-off to avoid misleading users. This performance even holds up when the question language does not match the document language. When it comes to identifying unanswerable questions given a context, there are larger differences between the two languages.

1 Introduction

Access to information is vital when moving to a new country, especially if the relocation is forced upon a person by war or persecution. Not knowing how to navigate immigration procedures and daily life in the host country can lead not only to confusion, insecurities, and delayed integration, but even to homelessness or deportation. NLP methods can and should be used to critically analyze public-policy (Beese et al., 2022; Blätte et al., 2020) and general-public discourse about immigration (Wang, 2024; Lapesa et al., 2020; Sanguinetti et al., 2018; Ross et al., 2016), to help newcomers learn new languages (Kochmar et al., 2023; Alfter et al., 2023, *inter alia*), and to provide answers to their everyday and immigration-related questions across languages and topics (this work).

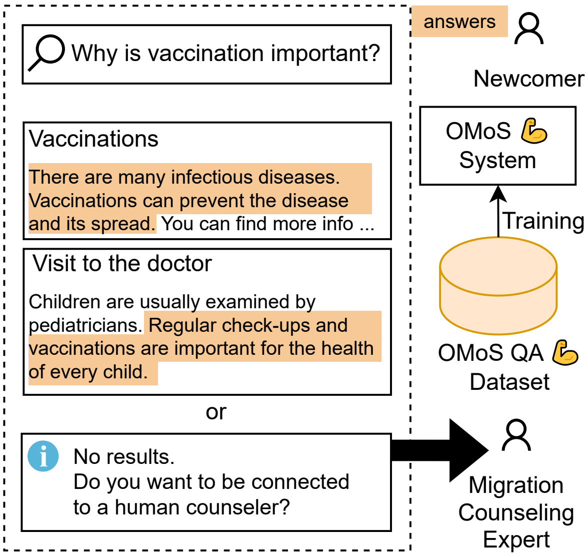

Why is vaccination important?
Vaccinations
There are many infectious diseases. Vaccinations can prevent the disease and its spread. You can find more info ...
Visit to the doctor
Children are usually examined by pediatricians. Regular check-ups and vaccinations are important for the health of every child.
or
No results.
Do you want to be connected to a human counseler?
answers
Newcomer
OMoS 💪 System
Training
OMoS QA 💪 Dataset
Migration Counseling Expert

Figure 1: Overview of our proposed task, system, and new dataset, OMoS-QA 💪: After the user asks a question, the system retrieves relevant documents and extracts answer sentences. The system is evaluated using the OMoS-QA 💪 corpus.

In this paper, we address the latter issue by presenting OMoS-QA,[1] an extractive QA dataset designed to support the development and rigorous testing of an online counseling system. We envision an application-tailored multilingual question-answering (QA) system which, given a question and a collection of informative and instructive texts, identifies sentences providing evidence for answer-

[1] *German:* **O**nline **M**igrationsberatung **o**hne **S**prachbarrieren; *English:* Online migration counseling without language barriers. Data and code available at https://github.com/digitalfabrik/integreat-qa-dataset. “omos” is also Greek for “shoulder with upper arm” 💪.

Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024), pages 231–248
September 10-13, 2024 ©2024 Association for Computational Linguistics

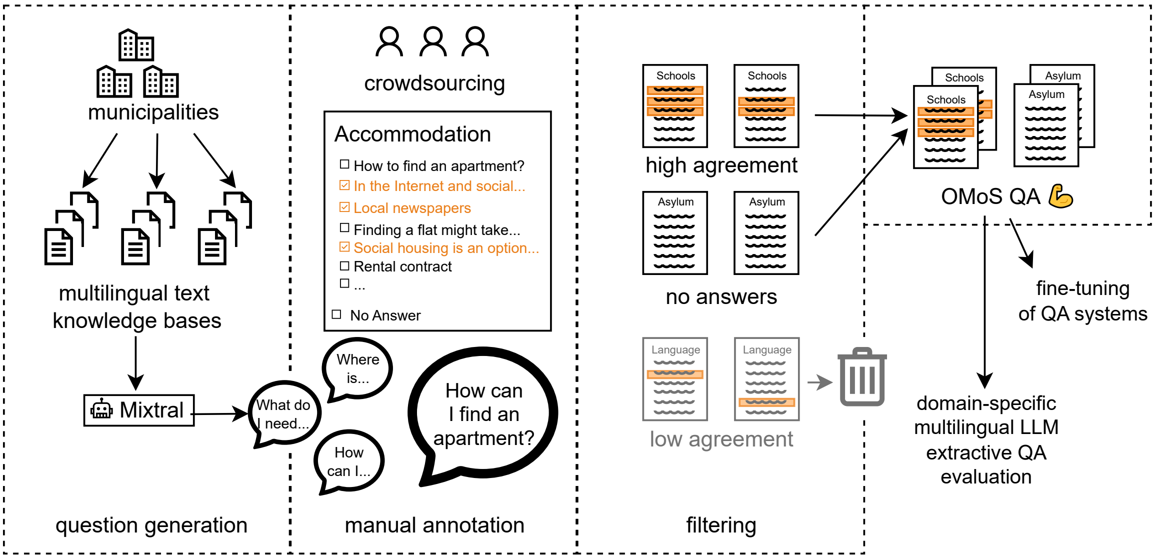


Figure 2: **OMoS-QA dataset creation.** Documents are taken from real-life multilingual knowledge bases. Questions are generated using Mixtral, but answers are annotated manually using crowdsourcing. The double-annotated dataset is then filtered on a question-level according to inter-annotator agreement.

ing the question in a relevant document (Fig. 1).

Germany has seen multiple waves of immigration since the 1950s, most recently more than one million war refugees from Syria, Iraq, and Afghanistan since 2015 and around one million war refugees from Ukraine since 2022 and ongoing. The German social system, aiming to support them, is known to be progressive but at the same time bureaucratic.[2] Providing the necessary customized information to each individual is an enormous logistical challenge. In particular during sudden crises, the counseling system has insufficient personnel capacities to sustain one-on-one counseling for less urgent inquiries. Hence, online resources are provided by cities and state governments, as well as NGOs. However, online information is scattered across many websites and portals, where it is location-specific, unstructured or structured inconsistently, and needs to be updated periodically—all on top of the language barrier.

OMoS-QA treats QA as a sentence extraction task rather than text generation, because faithfulness is of utmost importance. Well-known risks of free-text generation with large language models (LLMs), such as made-up facts and hallucinated entities (Shah and Bender, 2024; Ji et al., 2023; McKenna et al., 2023), are not acceptable in our application scenario of supporting migrants with information about social, economic, and legal processes. For the same reason, our approach aims to detect if a question is unanswerable given the provided evidence context. Extracting full sentences rather than token spans further helps with completeness and readability of the answers shown to the user. The process for constructing our new dataset is illustrated in Fig. 2. The contributions of this work are as follows.

- We present OMoS-QA, a manually annotated **corpus** of questions in German and English paired with relevant informational documents about a variety of social, economic, and legal topics and support offers. The documents were provided by three German municipalities, questions were **generated with an open-weight large language model** (LLM), and answer annotations were collected via voluntary **crowd-sourcing** (section 3).
- In order to construct a high quality dataset from the crowd-sourced annotations, we develop a filtering method based on a chance-corrected version of the Jaccard coefficient. We also present a detailed inter-annotator agreement study.
- Finally, we experiment with state-of-the-art pretrained LLMs (section 4). We compare 4 open-weight models as well as GPT-3.5, finding overall high precision in answer sentence selection and high recall in identifying unanswerable questions. A pilot cross-language QA study yields promising results.

[2] For example, there is a law that regulates who may or may not provide official immigration counseling.

2 Related Work

To ensure faithfulness of responses in our highly sensitive socio-political scenario, we focus exclusively on **extractive QA**, where the model is given a specific context to read, from which it should extract answers. Luo et al. (2022) provide a helpful comparative overview of extractive and generative approaches, and Luthier and Popescu-Belis (2020) have shown advantages of a hybrid system which dynamically chooses one of the two strategies.

Below we discuss related work on QA dataset construction, modeling extractive QA, and further NLP research in similar socio-political contexts.

QA Dataset Construction. The most popular QA datasets, such as SQuAD (Rajpurkar et al., 2016) and its derivatives (e.g. Rajpurkar et al., 2018; Möller et al., 2021), are general-purpose and thus not directly applicable to our scenario. However, curating and annotating a new QA corpus requires some finesse, especially when the target application is highly task-specific (Agarwal et al., 2022; Xu et al., 2022) or lies in a specific domain (Bechet et al., 2022; Han et al., 2022).

There is some consensus that **question generation** (QG) can be mostly automated, whereas ground-truth **answer annotations** should be provided by humans to ensure correctness. QG techniques that have proven useful include using a short summary of the context as input to the QG model (Dugan et al., 2022); question rewriting (Brabant et al., 2022); running QA as an auxiliary task and rewarding consistency between questions and answers (Yuan et al., 2023; Dugan et al., 2022); extracting QA-pairs from video transcripts (Westera et al., 2020; Pouran Ben Veyseh et al., 2022); and prompt engineering towards quality and diversity of the generated sentences (Schick and Schütze, 2021; Yuan et al., 2023). Manual answer annotation via crowd-sourcing, particularly making QA and other NLP tasks such as semantic role labeling (SRL) accessible to laypeople, has been popularized by the QA-SRL project (He et al., 2015; Roit et al., 2020; Brook Weiss et al., 2021).

In order to maintain high precision, we are particularly concerned with the option of marking a question as **unanswerable** given a context (cf. Rajpurkar et al., 2018; Liu et al., 2020; Henning et al., 2023). Moreover, Lauriola et al. (2022) have built a dataset of questions requiring clarifications, which we will consider in future work.

Finally, while **multi- and cross-linguality** remains a major challenge (Charlet et al., 2020), QA datasets in many languages (besides English) have been created in recent years, for German most notably by Möller et al. (2021).

Extractive QA Modeling. Approaches to extractive QA vary in whether they aim to predict a single span of a few tokens (Seo et al., 2017; Clark and Gardner, 2018; Hu et al., 2018), or whether the aim is to collect supporting evidence for a (possibly latent) answer (Murdock et al., 2012). To extract evidence sentences for choosing an answer in a multiple-choice QA setting, Wang et al. (2019) finetune a GPT model (Radford et al., 2018). Narayan et al. (2018) model the whole document via LSTMs over sentences before choosing sentences for answer selection and extractive summarization. Yoon et al. (2020) detect sentences for answering multihop questions with a graph neural net-based model that also takes the passage structure of the context into account.

Perhaps the closest to our problem setting in that both unanswered questions and discontiguous multi-span responses need to be accounted for (albeit in different application scenarios) are the works of Prasad et al. (2023) and Henning et al. (2023). Prasad et al. compare several pretrained BERT-style models in a multi-turn dialog setting while Henning et al. prompt a generative model to extract sentence numbers to answer questions on instructive texts.

Socio-political NLP Applications. In order to track, analyze, and predict trends in parliamentary debates about migrants and migration, Blätte et al. (2020) employ topic models while Beese et al. (2022) finetune a BERT model. A number of corpora have been compiled to study the public debate about immigration-related questions in Europe: e.g., in German and Slovene news (Lapesa et al., 2020; Zwitter Vitez et al., 2022), German and Italian social media (Ross et al., 2016; Sanguinetti et al., 2018), and UK partisan media (Wang, 2024).

3 The OMoS-QA Corpus

In this work, we present OMoS-QA, a novel dataset for QA in the context of Online Migrationsberatung ohne Sprachbarrieren (online migration counseling without language barriers). In its current version, it consists of over 900 automatically generated questions and manual answer annotations on documents contextually relevant to our problem setting in both German and English. In this section, we describe the dataset collection, annotation, and filtering, and provide corpus statistics.

3.1 Data Collection and Annotation

In an initial attempt, we tried to elicit common questions and their answers from administrative staff of migration agencies and NGO volunteers. This was unsuccessful due to their limited availability and the substantial time requirements necessary for the task. Therefore, inspired by Schick and Schütze (2021), we leverage the capabilities of LLMs to automatically generate questions. To ensure a high quality of the dataset, we collect at least two human answer annotations per question, facilitated by a new custom annotation tool. Only annotations that are largely agreed upon by two annotators are included in the final dataset.

Question Generation. We used Mixtral-8x7B-Instruct-v0.1 (henceforth abbreviated as Mixtral-8x7B; Jiang et al., 2024) to generate questions for German and English documents provided under CC BY 4.0 by three municipalities in Southern Germany.[3] The documents were retrieved using the Integreat API[4] on 2024-02-02. To facilitate the diversity of the dataset and to include both answerable and unanswerable questions, we employed two different question generation strategies for every document. In the first, the prompt contained the full document, in the second, we only provided an automatically generated three-word summary. The second strategy aimed at eliciting questions that are unanswerable given the provided document.

All questions were manually filtered, and in some cases corrected by the first author, e.g., "What are the emergency numbers provided?" was edited to "What emergency numbers are available?." In total, we collected 1,844 German questions for 548 documents and 3,062 English questions for 652 documents. Around 60% of the questions have been generated from a three-word summary such as "domestic violence support," "refugee counseling services," or "recognition of degrees."

Human Annotations. The task of finding the answers within documents resided with human annotators. As we resort to voluntary crowdsourcing, we aim to make the annotation process easy and time-efficient by creating a custom web-based annotation tool (see Appendix C) tailored to our use case. We frame the annotation task as the selection of one or multiple complete sentences that help to answer the question. Annotators are shown a question together with the text, and the option to select sentences via checkboxes. If no answer is found in the text, a separate checkbox has to be selected to consciously confirm this decision.

The annotators were recruited on a voluntary basis from German NGOs in the migration context and in the personal environment of the authors. Questions are randomly assigned to annotators on-the-fly, allowing each person to do as many (or few) annotations as they want. In total, we gathered 3,688 annotations for 1,944 questions by 238 annotators.

3.2 Question Filtering

To account for voluntary or involuntary mistakes, biases, and subjective answers by annotators, we require two annotations per question by different annotators. The annotations therefore amount to 1,744 questions with two annotations (de: 1,268, en: 476) for 863 different documents. To filter questions with low **inter-annotator agreement** (IAA), we measure question-level agreement using the *Jaccard index* over the two sets of sentences judged as relevant to answering the question by the two annotators. In a nutshell, the Jaccard index is defined as "intersection over union."

For measuring agreement, we use a chance-corrected Jaccard index. Our metric captures how much the two annotators agree on the selected set of sentences beyond chance. We assume, admittedly over-simplifying, that the prior probability of selecting a sentence is independent of the question, document, and annotator, and compute it as the total fraction of sentence selections over two times the corpus size (as each document receives two annotations). For details, see Appendix A. In our case, $P(sel)$ is 0.1856, and the expected agreement amounts to a Jaccard index of only 0.0344.

The average IAA over all questions is 0.34 (chance corrected: 0.31). This can be partly attributed to the fact that most questions are non-factoid, i.e., answers are not objective single "facts" but instead one or more relevant sentences where the boundaries around what should be the core answer and what is additional context are difficult to draw. To account for this difficulty, we modify

[3] The city of Munich and the districts (Landkreise) Augsburg and Rems-Murr-Kreis.

[4] https://digitalfabrik.github.io/integreat-cms/api-docs.html#pages

		train	dev	test	total
German	Questions	338	143	185	666
	No Answer	63 (19%)	30 (21%)	43 (23%)	136 (20%)
	Contiguous Answer	209 (62%)	86 (60%)	104 (56%)	399 (60%)
	Non-Contiguous Answer	66 (20%)	27 (19%)	38 (21%)	131 (20%)
	Documents	205	90	117	412
	Questions/Document	1.65	1.59	1.58	1.62
	Sentences/Document	27.16 ± 20.11	27.96 ± 15.87	26.91 ± 17.88	27.26 ± 18.59
	Chars/Sentence	58.62 ± 15.93	61.74 ± 16.32	61.96 ± 17.25	60.25 ± 16.44
	Chars/Question	57.85 ± 15.68	58.91 ± 17.21	59.61 ± 16.45	58.56 ± 16.23
	Agreement (Jaccard)	0.60 ± 0.33	0.59 ± 0.33	0.60 ± 0.34	0.60 ± 0.33
	with adjacent sentences	0.86 ± 0.19	0.85 ± 0.18	0.86 ± 0.19	0.86 ± 0.19
	Answer Sentences/Question	5.37 ± 6.09	5.57 ± 5.89	5.29 ± 6.84	5.39 ± 6.26
	Answers Sentences/Total Sentences	0.28 ± 0.29	0.25 ± 0.27	0.27 ± 0.28	0.27 ± 0.28
English	Questions	123	50	67	240
	No Answer	18 (15%)	8 (16%)	12 (18%)	38 (16%)
	Contiguous Answer	95 (77%)	38 (76%)	49 (73%)	182 (76%)
	Non-Contiguous Answer	10 (8%)	4 (8%)	6 (9%)	20 (8%)
	Documents	103	43	59	205
	Questions/Document	1.19	1.16	1.14	1.17
	Sentences/Document	23.51 ± 13.30	25.58 ± 16.68	25.49 ± 13.68	24.52 ± 14.14
	Chars/Sentence	65.28 ± 18.22	61.74 ± 12.72	60.48 ± 15.30	63.16 ± 16.45
	Chars/Question	59.46 ± 15.98	56.48 ± 13.22	56.51 ± 14.72	58.01 ± 15.11
	Agreement (Jaccard)	0.58 ± 0.34	0.59 ± 0.32	0.56 ± 0.34	0.58 ± 0.34
	with adjacent sentences	0.86 ± 0.20	0.84 ± 0.19	0.86 ± 0.20	0.86 ± 0.19
	Answer Sentences/Question	4.41 ± 4.98	3.90 ± 3.62	4.19 ± 4.39	4.24 ± 4.55
	Answers Sentences/Total Sentences	0.23 ± 0.23	0.20 ± 0.21	0.22 ± 0.24	0.22 ± 0.23
All	Questions	461	193	252	906
	No Answer	81 (18%)	38 (20%)	55 (22%)	174 (19%)
	Contiguous Answer	304 (66%)	124 (64%)	153 (61%)	581 (64%)
	Non-Contiguous Answer	76 (16%)	31 (16%)	44 (17%)	151 (17%)

Table 1: OMoS-QA: Overview of corpus statistics of final dataset. The Jaccard index is chance-corrected.

the annotations in a heuristic way as illustrated in Fig. 3. For each sentence marked by just one of the annotators that is adjacent to a sentence marked as relevant by both annotators, we change the annotation of the respective other annotator to "relevant" as well. We do this only if the sentence originally marked by both annotators is no more than three[5] sentences away.

After modifying the annotations to include adjacent sentences, the average Jaccard index is 0.50 (chance corrected: 0.48). To assure a high quality dataset, we filter out questions with a (non-chance-corrected) Jaccard index <0.5. This leaves us with 906 (51%) questions (de: 663, en: 243) with an average agreement of 0.86 (chance corrected: 0.86). The agreement when leaving out the adjustment of including adjacent sentences amounts to 0.61 (chance corrected: 0.59). As gold-standard answers we chose the intersection of both annotations, but including adjacent sentences as explained above.

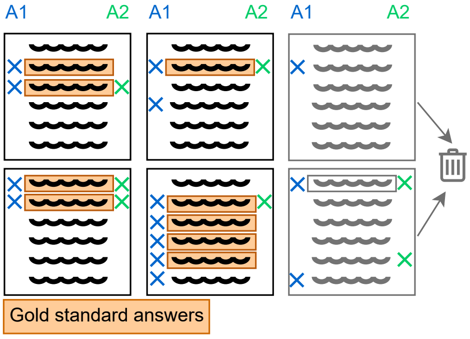


Figure 3: Gold standard construction from labels of two human annotators A1 (blue) and A2 (green). The gold standard contains sentences that A1 and A2 both mark as answers, as well as adjacent sentences marked by only one of them if at most three sentences away from the agreed-upon answer.

3.3 Final Dataset

Table 1 provides an overview of the corpus statistics of the final version of OMoS-QA. Out of the 906 QA pairs included in our **final dataset**, 151 (16%) have non-contiguous answers (i.e., the answer sentences are not adjacent), 110 (12%) have a single answer sentence and 165 (18%) questions

[5] This threshold is chosen as a middle ground between too little and too much additional context backed up by a manual inspection of samples.

have no answer in the document. The IAA did not differ substantially between German and English annotations in both the raw dataset (de: 0.34, en: 0.32) as well as the final dataset (de, en: 0.86).

Translations. To increase the size of the dataset and to take the multilingual setting into account, we translate the German questions and documents to English and vice versa using DeepL.[6] In order to preserve the gold-standard answers represented by the sentence indices, we translate each context sentence-by-sentence. Accordingly, in the German version of the dataset 240 and in the English version 666 of the 906 questions are machine-translated. We retain the information on the original languages.

Dataset Split. We split our dataset into train (51%), dev (21%) and test (28%) partitions with similar internal splits for the original language and the city the document is from. Questions without an answer, questions with contiguous and questions with non-contiguous answers are present with a similar probability over all partitions. As some questions refer to the same document, we make sure that no document occurs in multiple partitions. The proposed split is assuring a close to uniform distribution of several key properties of the dataset such as the agreement of both annotations, the document length or the annotated answer count.

4 Experiments

In this section, we describe our experiments. We evaluate several off-the-shelf LLMs as well as a finetuned sentence classifier on OMoS-QA.

4.1 Setup

We mostly follow the **prompt templates** proposed by Henning et al. (2023) for both the 0-shot and 5-shot settings, instructing the models to output a list containing the sentence IDs of the answer sentences.[7] We test the models in a 0-shot setting, only providing the prompt, but no concrete examples. In addition, we test the models in a 5-shot setting, in which we manually select and chunk examples for both German and English questions from the train partition (3 answerable, 2 unanswerable cases).[8] We use the same examples for all models and questions.

As **evaluation metrics**, we use precision (P), recall (R), and F1-score (F). To evaluate sentence-level retrieval (i.e. the binary task of selecting a sentence as an answer to the question), metrics are first computed per question at the sentence level and then macro-averaged over questions.

We also separately evaluate the binary task of identifying questions as **unanswerable** given the context. Here all metrics are at the question level. We consider two setups for extracting "unanswerable question" predictions from models: In the *inferred* setup, we run the models as before and treat generated empty lists (in the case of LLMs) or all-zero-vectors (in the case of DeBERTa) as classifying the question as unanswerable. In the *explicit* setup, we change the LLM instructions and classifier architecture to make an explicit binary prediction for each question.

During experimentation and hyperparameter selection, we evaluated only on the development split of OMoS-QA (results in Appendix D). Here we report our main results on the test split with the hyperparameters found during development.

4.2 Evaluated Models

We focus on open-weight models from MistralAI and Meta: Mixtral-8x7B (introduced in section 3.1), Mistral-7B-Instruct-v0.2 (Mistral-7B; Jiang et al., 2023) as well as Llama-3-8B-Instruct (Llama-3-8B) and Llama-3-70B-Instruct (Llama-3-70B) which are both successors of the Llama 2 model family (Touvron et al., 2023). We access these models via HuggingFace.[9] For comparison, we include results of the closed-source GPT-3.5-Turbo-0125 (GPT-3.5-Turbo) by OpenAI.[10]

4.3 Baseline

As a baseline, we run a sentence-wise classifier, consisting of a pretrained DeBERTa-v3-large encoder (He et al., 2021, accessed via HuggingFace) and a binary classification head.[11] For each sentence in a document, we pass the following input to the model: `[CLS]` `<question>` `[SEP]` `<context>` `[SEN]` `<target sentence>` `[SEN]` `<context>` `[SEP]`, where the classification is made based on the encoding of the `[CLS]` token, the target sentence is surrounded by three context sentences on

[6] https://developers.deepl.com/docs

[7] We used a model temperature of 0.75.

[8] We leave experiments with other proportions of answerable and unanswerable few-shot examples to future work.

[9] https://huggingface.co

[10] https://platform.openai.com/docs/models/gpt-3-5-turbo

[11] The head is a linear layer with 1024 input and 2 output features on top of a pooling layer. Additional hyperparameters are given in Table 6.

		Sentence-level Answers						Question-level Unanswerability					
		German			English			German			English		
Model	Setting	P	R	F	P	R	F	P	R	F	P	R	F
Mixtral-8x7B	0-shot	74.5	47.1	57.7	73.4	44.2	55.2	68.9	56.4	62.0	65.8	45.5	53.8
	5-shot	79.0	51.7	**62.5**	77.9	50.5	61.3	67.8	72.7	70.2	65.6	76.4	70.6
Mistral-7B	0-shot	69.7	47.8	56.7	74.1	47.5	57.9	**80.0**	14.5	24.6	70.0	25.5	37.3
	5-shot	**87.6**	20.3	32.9	84.3	29.5	43.7	29.2	**89.1**	43.9	30.3	72.7	42.8
Llama-3-8B	0-shot	74.9	30.0	42.9	78.2	34.8	48.1	71.1	49.1	58.1	54.7	52.7	53.7
	5-shot	81.9	42.2	55.7	82.1	44.2	57.4	54.7	85.5	66.7	53.6	**81.8**	64.7
Llama-3-70B	0-shot	85.5	46.6	60.3	84.8	46.7	60.2	69.8	67.3	68.5	74.5	63.6	68.6
	5-shot	86.7	48.2	62.0	84.9	48.4	61.6	68.3	78.2	**72.9**	64.5	72.7	68.4
GPT-3.5-Turbo	0-shot	85.3	31.6	46.1	**87.3**	31.2	45.9	50.8	60.0	55.0	54.4	67.3	60.2
	5-shot	81.8	45.1	58.1	83.8	43.9	57.6	70.9	70.9	70.9	67.2	74.5	**70.7**
DeBERTa	–	62.6	**62.4**	**62.5**	65.7	**64.2**	**64.9**	56.2	65.5	60.5	59.4	69.1	63.9
*Human Agreement**		–	–	*57.8*	–	–	*57.8*	–	–	*47.8*	–	–	*47.8*
test partition only		–	–	*76.3*	–	–	*76.3*	–	–	*100.0*	–	–	*100.0*

Table 2: Test set performance (in %) of zero-shot and 5-shot LLMs and finetuned DeBERTa on sentence-level answer extraction (left) and detection of unanswerable questions (right). The best result in each column is **bolded**. **Human Agreement* is computed from agreement before the dataset filtering step (Fig. 2) and therefore not directly comparable to model performance.

the left and right (altogether surrounded by [SEP] tokens),[12] and we add the new [SEN] special token to the vocabulary to mark the target sentence.

We finetune the full model on OMoS-QA.

4.4 Results

We present our results in the left half of Table 2. All models show very good precision (70–90%), with the highest numbers achieved by the Llama-3 models. Recall is much lower in general, with a wider span across models, reaching as low as 20.3% (Mistral-7B 5-shot in German). Most models reach between 40% and 50% recall while maintaining high precision, which seems to be a favorable trade-off. Keep in mind that selecting fewer but clearly relevant sentences, as opposed to more noisy ones, is generally in line with our goals of providing trustworthy results. The highest precision is achieved by Mistral-7B -shot for German and GPT-3.5-Turbo 0-shot for English.

Mixtral-8x7B, Llama-3-70B, and DeBERTa strike the best overall precision/recall trade-offs (F1-score). DeBERTa in particular has almost equal precision and recall.

The last row of the table presents an approximation of the "human performance" as measured via the inter-annotator agreement (F1-score) in our dataset. For each question, the data labeled by the various annotators is assigned to one of two sets randomly, and then one set is treated as the gold standard and the other as the system. As the German and English versions of the dataset consist of the same (potentially translated, see section 3.3) questions and documents, the score is the same in the two languages. We provide a version of this human score before majority voting and including adjacent sentences, which gives an idea of the difficulty of the task, even for humans—though note that these are untrained voluntary annotators and trained experts might achieve higher agreement. Due to the data mismatch, this is not directly comparable to the system evaluation setup, thus we also provide a more optimistic version after filtering ("test partition only"), which is computed on the same data as the models.

Identifying Unanswerable Questions. We report results separately for the subset of questions where the human annotators agreed that the answer is not in the text (right side of Table 2). Here, recall reflects how many of the unanswerable questions were correctly identified by the model as such. Precision indicates how many of the questions predicted as unanswerable did indeed not have an answer in the provided text.

For identifying unanswerable questions, we put higher priority on recall over precision, in line with our cautious approach to a sensitive scenario. And indeed we find that overall, recall is higher and

[12] The context size of 3 has been determined via experimentation on the dev set.

Model & Method		German P	German R	German F	English P	English R	English F
Llama-3	Exp	59.0	**83.6**	**69.2**	62.3	**78.2**	**69.4**
70B	Inf	69.8	67.3	68.5	74.5	63.6	68.6
DeBERTa	Exp	**75.0**	43.6	55.2	**75.0**	54.5	63.2
	Inf	56.2	65.5	60.5	59.4	69.1	63.9

Table 3: Test set performance (in %) of zero-shot Llama-3-70B and finetuned DeBERTa on explicit and inferred question-level unanswerability detection. The best result in each column is **bolded**. Exp=Explicit, Inf=Inferred.

precision lower than in sentence extraction. In many cases, recall is higher than precision.

In Table 3 we see that explicitly instructing or training models to recognize unanswerable questions has different effects depending on the model type. Changing the zero-shot prompt given to Llama-3-70B increases recall and decreases precision compared to inferring this decision from an empty prediction. Changing the training task of the DeBERTa-classifier has the opposite effect. This might be a result of the decrease in the amount of training data that DeBERTa receives—only one example per question in the explicit setting versus one example per document sentence per question in the inferred setting. This quantitative difference does not apply to the LLM, which instead profits from the more precisely-phrased prompt.

Zero-shot vs. Few-shot. In most conditions, few-shot learning from 5 examples is beneficial either for both recall and precision, or for recall without hurting precision too much. An exception is Mistral-7B, which overshoots on extracting fewer answers in the 5-shot scenario, with a strongly increased recall on unanswerable questions, but a worse performance on the answerable questions.

Performance by Number of Answer Sentences. In all conditions and metrics (P, R, F) we observe standard deviations over individual datapoints (questions with at least one ground-truth answer) between ± 30 and ± 40 metric points. This variance can in part be explained by the varying difficulty of questions with increasing numbers of ground-truth answer sentences. The average number of gold answer sentences (henceforth "#answers") lies between 5 and 6 in German and around 4 in English (Table 1). We show model performance as a function of #answers exemplarily for one German model in Fig. 4. As can be expected, average recall becomes roughly linearly more difficult as #answers increases, whereas average precision already starts high and approaches 1.0 for questions with more than 10 annotated answer sentences.

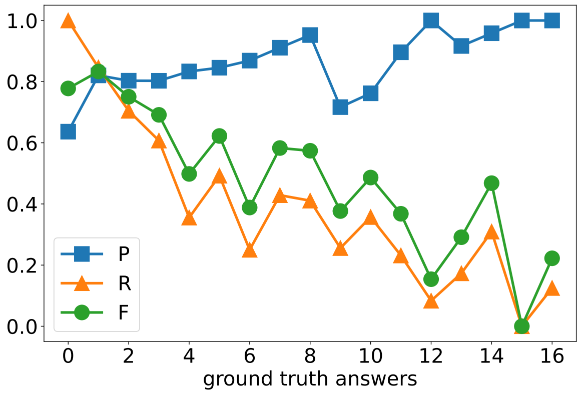


Figure 4: Test set performance as a function of the number of ground-truth answer sentences (0-shot Llama3-70B on German questions and documents).

Doc.	Q.	Answerable P	Answerable R	Unanswerable P	Unanswerable R
Ger.	Ger.	85.5	46.6	69.8	67.3
Ger.	Eng.	**85.8**	**48.2**	70.9	70.9
Ger.	Ara.	84.6	41.8	63.1	**74.5**
Eng.	Ara.	80.6	44.0	74.0	67.3
Eng.	Eng.	84.8	46.7	**74.5**	63.6
Eng.	Ger.	83.2	45.6	73.5	65.5
Ara.	Ara.	80.9	42.2	71.4	54.5
Ara.	Eng.	82.7	44.4	74.0	67.3
Ara.	Ger.	81.9	43.1	72.7	72.7

Table 4: Test set performance (%) of 0-shot Llama-3-70B on **cross-language** question-context pairs.

4.5 Cross-language QA

We also conduct a pilot cross-language QA study with German, English, and Arabic questions and documents. We compare scenarios where the question language does not match the document language against scenarios where it does. We choose Llama-3-70B over Mixtral-8x7B for this experiment, because while both perform well in section 4.4, the latter was used to generate our questions.

Our findings are shown in Table 4. Surprisingly, asking a question in a different language than the document does not hurt performance by a lot. In fact, it seems that asking questions in English works best, regardless of document language, and German documents work best, regardless of question language.

Model	Context toks (Thousands)	Params (Billions)
Mixtral-8x7B	32[13]	46.7 (12.9)[13]
Mistral-7B	32[14]	7[14]
Llama-3-8B	8[15]	8[15]
Llama-3-70B	8[15]	70[15]
GPT-3.5-Turbo	16[16]	unknown
DeBERTa-v3-large	1[17]	0.4[17]

Table 5: Model sizes. Mixtral has a total of 46.7B parameters but uses only a subset of 12.9 of them for each token.

5 Discussion

We interpret our results as largely positive, in particular with respect to our goal of building a reliable system that errs on the side of presenting fewer, higher precision results to the user. On our dataset, the newest open-weight models Mixtral-8x7B and Llama-3-70B can easily compete with closed-weight GPT-3.5.

With our various evaluation criteria and prompting setups (0-shot vs. 5-shot), we highlight different models' individual strengths: For example, the smaller LLMs Mistral-7B and Llama-3-8B are best at selectively identifying high-confidence answer sentences only, leading to extremely high sentence precision and unanswerability recall. They might thus lend themselves to an answerability filtering step, after which other models like Mixtral-8x7B and Llama-3-70B can do the heavy-lifting of higher-recall answer extraction.

It is important to keep in mind that we already use Mixtral-8x7B to generate questions, which likely contributes to its good performance (cf. Panickssery et al., 2024).

Our cross-language QA experiment suggests that translating questions asked in lower-resource languages (such as Arabic) to English and performing QA on German documents is a promising approach. Appendix E provides additional experimental results with translated and back-translated questions, which suggest that automatic translation is useful for Arabic and Ukrainian, but not so much for French, which is more similar to German and English in terms of both data availability and grammar. In future experiments, it will be interesting to introduce additional noise into questions before prompting, such as spelling errors or code-mixing, to simulate realistic user interactions and measure models' robustness.

While LLMs are indeed powerful and flexible tools that can be quickly adapted to a specialized task via in-context learning from few-shot prompts, we also see that the best-performing LLMs in our setting are the ones with the most parameters (Table 5). Much smaller, specialized models, such as task-specific classifiers built upon DeBERTa or other BERT-style encoders, are generally more controllable, interpretable, and environmentally friendly. Together with the competitive QA performance in terms of F1 and well-balanced precision and recall we observe, this emphasizes that this model class is still very much viable for practical applications in sensitive scenarios.

We will take these findings into account as we continuously work towards automating the document retrieval component and a service-ready implementation of the full QA system, and including more and more languages as potential query and document languages.

6 Conclusion

In this paper, we address the task of providing high-precision, knowledge-grounded answers to users who have freshly immigrated to Germany. We approach this challenge by compiling, manually annotating, and filtering a novel dataset, OMoS-QA, containing in total 900 document-question pairs in German and English. The dataset will be available to the research community under a CC-BY license. We also present experimental results on our new dataset from a comparison of 5 LLMs and a finetuned classifier, as well as a pilot cross-language QA study. Our results are promising and open the doors to future finetuning and large-scale multilingual experiments.

Limitations

The OMoS-QA dataset is designed to support extractive QA in an online counseling system for immigrants. In this paper, we have modeled an admittedly simplified scenario in which the document (potentially) containing the answer to a question is already provided (an assumption that is made in most currently used QA benchmarks). A full

[13] https://mistral.ai/news/mixtral-of-experts/

[14] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[15] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[16] https://platform.openai.com/docs/models/gpt-3-5-turbo

[17] https://huggingface.co/microsoft/deberta-v3-large

search scenario would of course also require identifying potentially relevant documents, i.e., include a search component.

Another limitation of our work is that annotators were not trained specifically for our task. We counterbalance this issue by double-annotations and extensive filtering.

Finally, the current version of OMoS-QA is limited to German and English documents and questions. As immigrants arrive from all over the world, an in particular in urgent crises without the possibility to study German in advance, more work is necessary to mitigate the language barrier. In future work, we plan to also conduct experiments for an extended set of languages.

Ethics Statement

During dataset construction, annotators participated on a voluntary basis and agreed to the anonymized publishing of their annotations. Before starting the annotations, they agreed to the terms shown in Appendix C. As the annotation study only included marking relevant answers to technical questions in text, i.e., annotators did not have to write text or provide personal information, no IRB review was deemed necessary.

Online migration counseling offers convenience and accessibility, but it also comes with several challenges.[13] First of all, there is a lack of a personal connection, which may be crucial in our scenario. Ensuring client confidentiality can be more challenging in an online environment. Misinterpretation of cultural cues or nuances in communication may occur, leading to misunderstandings or ineffective counseling outcomes. Finally, there are also technological barriers: not everyone has access to reliable internet connections or appropriate devices. Yet, our work is a first attempt at developing reliable language technology to support the immigration counseling process. Municipalities could, for example, provide computer terminals at the immigration authorities' offices, townhalls, or libraries. And being able to search for information in a targeted system is still much of an advantage compared to waiting for an appointment for weeks. Moreover, such a system would also lead to a more effective use of the official counselor's time, as it would relieve them from providing advice in "easy" cases.

Acknowledgments

We are extremely grateful to our voluntary annotators for their hard work and to Tür an Tür Digitalfabrik GmbH, in particular Daniel Kehne und Sven Seeberg-Elverfeldt, for supporting this work and funding the first author. We also thank Alexander Knapp and members of the Augsburg HLT group for inspiring discussions, as well as the anonymous reviewers for their helpful comments and suggestions.

References

Ankush Agarwal, Raj Gite, Shreya Laddha, Pushpak Bhattacharyya, Satyanarayan Kar, Asif Ekbal, Prabhjit Thind, Rajesh Zele, and Ravi Shankar. 2022. Knowledge graph - deep learning: A case study in question answering in aviation safety domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6260–6270, Marseille, France. European Language Resources Association.

David Alfter, Elena Volodina, Thomas François, Arne Jönsson, and Evelina Rennes, editors. 2023. *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*. LiU Electronic Press, Tórshavn, Faroe Islands.

Frederic Bechet, Elie Antoine, Jérémy Auguste, and Géraldine Damnati. 2022. Question generation and answering for exploring digital humanities collections. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4561–4568, Marseille, France. European Language Resources Association.

Dominik Beese, Ole Pütz, and Steffen Eger. 2022. FairGer: Using NLP to measure support for women and migrants in 155 years of German parliamentary debates. ArXiv preprint arXiv:2210.04359.

Andreas Blätte, Simon Gehlhar, and Christoph Leonhardt. 2020. The Europeanization of parliamentary debates on migration in Austria, France, Germany, and the Netherlands. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 66–74, Marseille, France. European Language Resources Association.

Quentin Brabant, Gwénolé Lecorvé, and Lina M. Rojas Barahona. 2022. CoQAR: Question rewriting on CoQA. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 119–126, Marseille, France. European Language Resources Association.

Daniela Brook Weiss, Paul Roit, Ayal Klein, Ori Ernst, and Ido Dagan. 2021. QA-align: Representing cross-text content overlap by aligning question-answer propositions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9879–9894, Online and Punta Cana,

[13] This list was compiled with the help of ChatGPT, yet it reflects our own opinion as well.

Dominican Republic. Association for Computational Linguistics.

Delphine Charlet, Geraldine Damnati, Frederic Bechet, Gabriel Marzinotto, and Johannes Heinecke. 2020. Cross-lingual and cross-domain evaluation of machine reading comprehension with squad and CALOR-quest corpora. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5491–5497, Marseille, France. European Language Resources Association.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia. Association for Computational Linguistics.

Liam Dugan, Eleni Miltsakaki, Shriyash Upadhyay, Etan Ginsberg, Hannah Gonzalez, DaHyeon Choi, Chuning Yuan, and Chris Callison-Burch. 2022. A feasibility study of answer-agnostic question generation for education. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1919–1926, Dublin, Ireland. Association for Computational Linguistics.

Kelvin Han, Thiago Castro Ferreira, and Claire Gardent. 2022. Generating questions from Wikidata triples. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 277–290, Marseille, France. European Language Resources Association.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*.

Sophie Henning, Talita Anthonio, Wei Zhou, Heike Adel, Mohsen Mesgar, and Annemarie Friedrich. 2023. Is the answer in the text? challenging ChatGPT with evidence retrieval from instructive text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14229–14241, Singapore. Association for Computational Linguistics.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2018. Reinforced mnemonic reader for machine reading comprehension. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4099–4106. AAAI Press.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. ArXiv Preprint 2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors. 2023. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada.

Gabriella Lapesa, Andre Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss, Jonas Kuhn, and Sebastian Padó. 2020. DEbateNet-mig15:tracing the 2015 immigration debate in Germany over time. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 919–927, Marseille, France. European Language Resources Association.

Ivano Lauriola, Kevin Small, and Alessandro Moschitti. 2022. Building a dataset for automatically learning to detect questions requiring clarification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4701–4707, Marseille, France. European Language Resources Association.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Man Luo, Kazuma Hashimoto, Semih Yavuz, Zhiwei Liu, Chitta Baral, and Yingbo Zhou. 2022. Choose your QA model wisely: A systematic study of generative and extractive readers for question answering. In *Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*, pages 7–22, Dublin, Ireland and Online. Association for Computational Linguistics.

Gabriel Luthier and Andrei Popescu-Belis. 2020. Chat or learn: a data-driven robust question-answering system. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5474–5480, Marseille, France. European Language Resources Association.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774, Singapore. Association for Computational Linguistics.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 42–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

J. W. Murdock, J. Fan, A. Lally, H. Shima, and B. K. Boguraev. 2012. Textual evidence gathering and analysis. *IBM Journal of Research and Development*, 56(3.4):8:1–8:14.

Shashi Narayan, Ronald Cardenas, Nikos Papasarantopoulos, Shay B. Cohen, Mirella Lapata, Jiangsheng Yu, and Yi Chang. 2018. Document modeling with external attention for sentence extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2020–2030, Melbourne, Australia. Association for Computational Linguistics.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. ArXiv Preprint 2404.13076.

Amir Pouran Ben Veyseh, Viet Lai, Franck Dernoncourt, and Thien Nguyen. 2022. BehanceQA: A new dataset for identifying question-answer pairs in video transcripts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7321–7327, Marseille, France. European Language Resources Association.

Archiki Prasad, Trung Bui, Seunghyun Yoon, Hanieh Deilamsalehy, Franck Dernoncourt, and Mohit Bansal. 2023. MeetingQA: Extractive question-answering on meeting transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15000–15025, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

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.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. Association for Computational Linguistics.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In *Proceedings of the 3rd Workshop on Natural Language Processing for Computer-mediated Communication (NLP4CMC)*, number 17 in Bochumer linguistische Arbeitsberichte: BLA, pages 6–9.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*.

Chirag Shah and Emily M. Bender. 2024. Envisioning information access systems: What makes for good tools and a healthy web? *ACM Trans. Web*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. ArXiv Preprint 2307.09288.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, David McAllester, and Dan Roth. 2019. Evidence sentence extraction for machine reading comprehension. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 696–707, Hong Kong, China. Association for Computational Linguistics.

Yunxiao Wang. 2024. Metaphorical framing of refugees, asylum seekers and immigrants in UKs left and right-wing media. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 18–27, St. Julians, Malta. Association for Computational Linguistics.

Matthijs Westera, Laia Mayol, and Hannah Rohde. 2020. TED-Q: TED talks and the questions they evoke. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1118–1127, Marseille, France. European Language Resources Association.

Zhuoqun Xu, Liubo Ouyang, and Yang Liu. 2022. Task-driven and experience-based question answering corpus for in-home robot application in the House3D virtual environment. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6232–6239, Marseille, France. European Language Resources Association.

Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. Propagate-selector: Detecting supporting sentences for question answering via graph neural networks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5400–5407, Marseille, France. European Language Resources Association.

Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2023. Selecting better samples from pre-trained LLMs: A case study on question generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12952–12965, Toronto, Canada. Association for Computational Linguistics.

Ana Zwitter Vitez, Mojca Brglez, Marko Robnik Šikonja, Tadej Škvorc, Andreja Vezovnik, and Senja Pollak. 2022. Extracting and analysing metaphors in migration media discourse: towards a metaphor annotation scheme. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2430–2439, Marseille, France. European Language Resources Association.

A Chance-corrected Jaccard Coefficient

For computing agreement, we use a chance-corrected version of the Jaccard coefficient. For a question q_i, it is defined as follows for two sets of selected answer sentences $A_{i_a} \subseteq S_i$ and $A_{i_b} \subseteq S_i$, where S_i is the set of all sentences of the document, and a and b index the two annotators:

$$agr_{obs} = J(A_{i_a}, A_{i_b}) = \frac{|A_{i_a} \cap A_{i_b}|}{|A_{i_a} \cup A_{i_b}|}$$

For $A_{i_a} = A_{i_b} = \emptyset$ we set $J(A_{i_a}, A_{i_b}) = 1$ as both annotators completely agree that there is no answer.

Chance Correction. In order to account for the possibility of authors just agreeing "by chance," chance correction can be applied. As the prior probability $P(sel)$ of a sentence $s_{i_k} \in S_i$ being selected we take the amount of all sentence selection in the whole corpus divided by the amount of all sentences in the corpus times 2 to account for two annotations being made:

$$P(sel) = \frac{\sum_{i=1}^{n}(|A_{i_a}| + |A_{i_b}|)}{2 * \sum_{i=1}^{n} |S_i|}$$

The probability $P(agr)$ that two random annotations agree on a sentence being an answer is then:

$$agr_{exp} = P(agr) = P(sel)^2$$

$P(agr)$ is therefore the expected agreement agr_{exp}. The observed agreement agr_{obs} is the Jaccard index $J(A_{i_a}, A_{i_b})$, such that the chance-corrected Jaccard index can be calculated as follows:

$$J_{cc}(A_{i_a}, A_{i_b}) = \frac{agr_{obs} - agr_{exp}}{1 - agr_{exp}}$$

B Prompt Template

As mentioned in section 4.1 we mostly follow the prompt template proposed by Henning et al. (2023) for both our 0-shot and 5-shot experiments. As Mixtral-8x7B and Mistral-7B do not support messages with the *system* role, we only include *user* and *assistant* messages for these models. Our complete 0-shot prompt:

system: `Your task is to select sentences from a document that answer a given question.` (Llama-3 models and GPT-3.5-Turbo only)

user (question, document): `Given the question and document below, select the sentences from the document that answer the question. It may also be the case that none of the sentences answers the question. In the document, each sentence is marked with an ID. Output the IDs of the relevant sentences as a list, e.g., "[1,2,3]", and output "[]" if no sentence is relevant. Output only these lists.`
`Question: {question}`
`Document: {document}`

We use the chunked samples shown in Fig. 5 (or their sentence-by-sentence translations) for the 5-shot experiments. For each sample we insert the following two messages to the prompt before the final user message:

user (question, document)
assistant (answers): `{answers}`

C Custom Annotation Tool

For the human annotations described in section 3.1 we developed a custom web-based annotation tool for the selection of the answer sentences. All human annotators agreed to the following conditions: *I agree to the processing and publication of my annotations and their use for machine learning. All annotations and information entered will be stored and processed anonymously.* Fig. 6 shows a screenshot of the custom annotation tool.

D Development Set Performance

We observe slightly different trends on the development set (Table 7) than on the test set (Table 2). Namely, three 0-shot model setups have a particularly low recall on sentence extraction: Mixtral-8x7B, Llama-3-8B, and GPT-3.5, which means in conjunction with high precision that they tend to generally extract fewer sentences per question. Out of these three, Mixtral-8x7B and Llama-3-8B also have particularly low precision at identifying unanswerable questions, meaning that more often than not they do not extract any answer sentence for questions which would in fact be answerable given the context. This gets largely fixed by providing few-shot examples.

Question 1: What do you need to open a bank account?
Document 1:
[9] When can I start learning to drive?
[10] In Germany, you may only drive a car with a valid driverś license.
[11] Beforehand, you have to attend a driving school and take theoretical and practical lessons, which you also have to pay for.
[12] You can get information about this at the driving school.
[13] When can I open my own bank account?
Answer 1: []

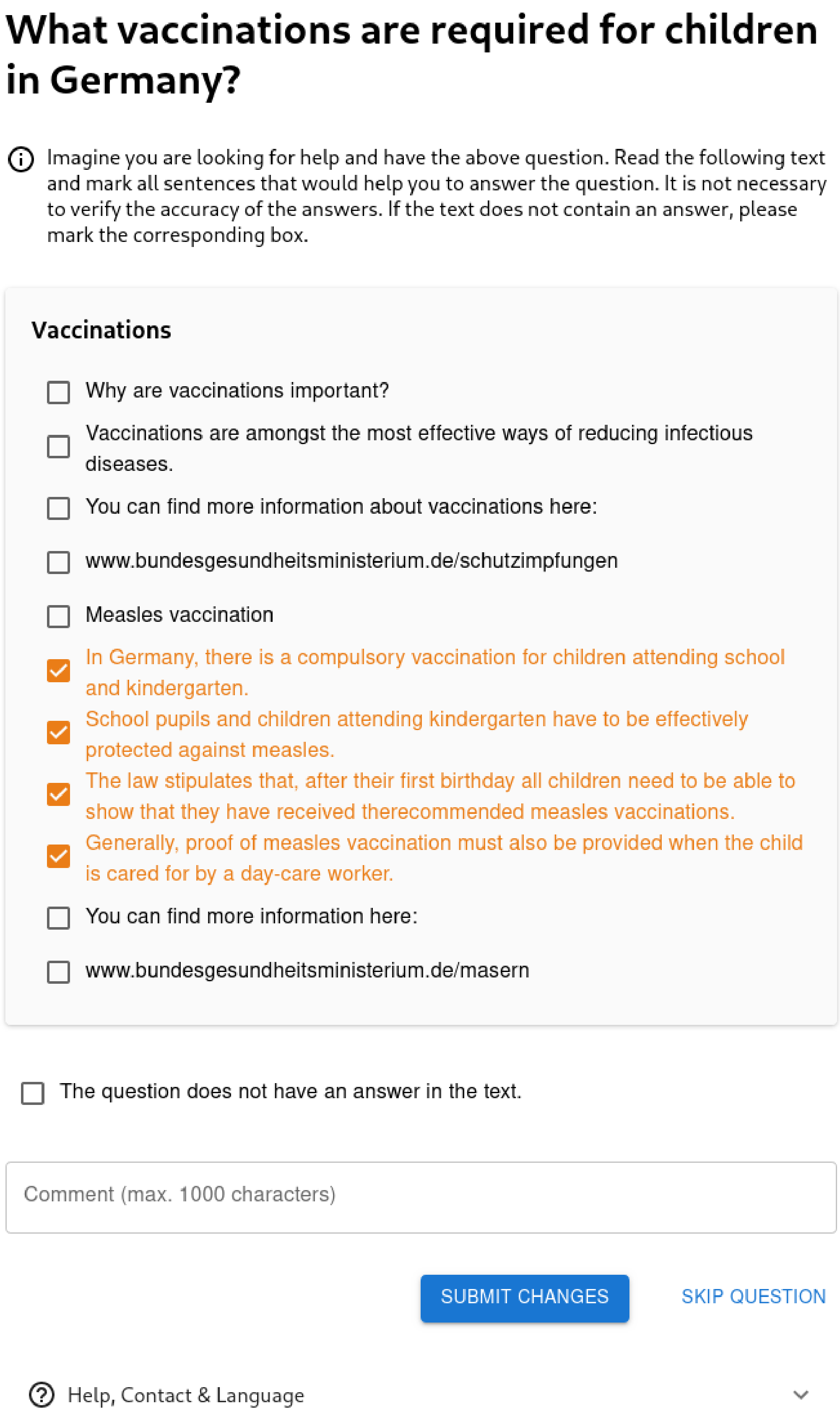
Question 2: What is a fictitious certificate?
Document 2:
[0] Residence with fictitious certificate
[1] Departure with a fictitious certificate
[2] With a fictitious certificate, you have a temporary right of residence.
[3] There are different types of fictitious certificate.
[4]Please note:
[5] Re-entry into the federal territory is only possible with a fictitious certificate in accordance with § 81 para.4 AufenthG possible.
Answer 2: [2]

Question 3: Where can I find information on admission procedures at vocational schools?
Document 3:
[11] Initial vocational training is possible at vocational schools and vocational colleges.
[12] Training can take place both in the dual system (training company and vocational school) or “purely” school-based training (vocational schools).
[13] The dates and registration requirements vary from vocational school to vocational school.
[14] Information evenings are held at vocational schools every year before enrollment.
[15] Information on the admission procedure at the vocational schools can be obtained directly from the respective school.
[5] Re-entry into the federal territory is only possible with a fictitious certificate in accordance with § 81 para.4 AufenthG possible.
Answer 3: [14, 15]

Question 4: What types of school are there in Germany?
Document 4:
[0] Support with school or personal problems
[1] Does your child need help with problems?
[2] Then these places will help you:
[3] Youth social work (JaS for short) and youth work at schools (JA for short) for school, personal or family problems:
[4] It is best to contact the school directly or the Augsburg District Office for general information:
Answer 4: [0]

Question 5: What topics are covered in the initial orientation courses?
Document 5:
[2] The German courses for initial language orientation (also known as initial orientation courses) teach both basic German language skills and information about life in Germany.
[3] They are a practical starting aid in the new living environment and make everyday life easier.
[4] A course comprises 300 teaching units of 45 minutes each and covers topics such as “Health/medical care”, “Work”, “Kindergarten/school”, “Housing”, “Local orientation/transport/mobility”.
[5] The focus is on oral communication: participants should learn as quickly as possible to find their way around in everyday life.
[6] Across all modules, initial orientation courses are also about teaching values.
Answer 5: [2, 4, 5, 6]

Figure 5: Chunked samples for 5-shot experiments.



What vaccinations are required for children in Germany?

Imagine you are looking for help and have the above question. Read the following text and mark all sentences that would help you to answer the question. It is not necessary to verify the accuracy of the answers. If the text does not contain an answer, please mark the corresponding box.

Vaccinations

- ☐ Why are vaccinations important?
- ☐ Vaccinations are amongst the most effective ways of reducing infectious diseases.
- ☐ You can find more information about vaccinations here:
- ☐ www.bundesgesundheitsministerium.de/schutzimpfungen
- ☐ Measles vaccination
- ☑ In Germany, there is a compulsory vaccination for children attending school and kindergarten.
- ☑ School pupils and children attending kindergarten have to be effectively protected against measles.
- ☑ The law stipulates that, after their first birthday all children need to be able to show that they have received therecommended measles vaccinations.
- ☑ Generally, proof of measles vaccination must also be provided when the child is cared for by a day-care worker.
- ☐ You can find more information here:
- ☐ www.bundesgesundheitsministerium.de/masern

☐ The question does not have an answer in the text.

Figure 6: Custom annotation tool

	Sentence Classification	Question Classification
Batch size	8	8
Learning rate	$2 * 10^{-6}$	$2 * 10^{-6}$
Weight decay	0.1	0.1
Warmup steps	50	50
Evaluation steps	50	10
Max. epochs	3	10
Early stopping	10	10

Table 6: The used hyperparameters for finetuning DeBERTa for answer extraction using binary sentence classification and question answerability classification.

		Answerable questions: sentence-level						Identifying unanswerable questions					
		German			English			German			English		
Model	Setting	P	R	F	P	R	F	P	R	F	P	R	F
Mixtral-8x7B	0-shot	74.1	31.7	32.4	67.7	29.0	29.3	38.8	81.6	52.5	41.1	78.9	54.1
	5-shot	74.0	**58.0**	**55.6**	72.9	**53.9**	52.4	76.2	84.2	80.0	68.9	81.6	74.7
Mistral-7B	0-shot	74.0	45.5	47.0	76.7	45.7	48.4	59.5	57.9	58.7	58.3	55.3	56.8
	5-shot	71.0	42.8	40.7	70.5	39.0	38.7	50.9	71.1	59.3	52.8	73.7	61.5
Llama-3-8B	0-shot	**89.8**	26.8	30.0	**86.2**	33.9	37.7	32.0	86.8	46.8	40.8	81.6	54.4
	5-shot	77.6	44.6	45.8	78.1	39.2	40.6	61.0	**94.7**	74.2	52.3	**89.5**	66.0
Llama-3-70B	0-shot	84.2	48.6	53.9	79.6	48.3	52.6	**81.6**	81.6	**81.6**	**77.5**	81.6	**79.5**
	5-shot	85.9	51.1	55.4	82.8	51.9	**55.0**	66.7	84.2	74.4	70.8	89.5	79.1
GPT-3.5-Turbo	0-shot	70.4	33.9	38.5	73.3	36.9	42.6	63.2	63.2	63.2	73.5	65.8	69.4
	5-shot	77.5	47.9	50.8	80.7	44.1	49.5	78.9	78.9	78.9	71.4	78.9	75.0
*Human Upper Bound**		–	–	*62.9*	–	–	*62.9*	–	–	–	–	–	–
with adjacent sentences		–	–	*88.8*	–	–	*88.8*	–	–	–	–	–	–

Table 7: Development set performance (in %) of 0-shot and 5-shot LLMs on answerable questions (left) and unanswerable questions (right). The best result in each column is **bolded**. **Human upper bound* is computed from agreement data and not directly comparable.

E Multilingual Experiments

We evaluate models on the following additional languages that are highly relevant in the migration context: Arabic (ar), French (fr), and Ukrainian (uk). These and other languages are more challenging due to their limited resources and much different language structure (German and English are closely related). Furthermore, Arabic and Ukrainian both use a non-Latin alphabet: The Arabic and Cyrillic alphabet. We use machine translation with DeepL to translate the question and, sentence-by-sentence, the document for each instance of the original OMoS-QA dataset.

In order to assess possible adverse effects of leveraging machine translation and to compare it to directly querying the model with the question in its original language, we evaluate the performance in an additional retranslation setting. To this end, we combine the original German documents with retranslated questions, i.e., questions that are first translated to the aforementioned languages and then back to German. This corresponds to the use of machine translation in the full OMoS system, as only user input (and possibly the answers) are subject to translation, while the document corpus remains unchanged. However, questions are translated twice in the retranslation setting and results should thus be considered as lower performance boundary. Since German is the original dataset language of OMoS-QA, there are no results for the retranslated setting.

E.1 Sentence-Level Results

The results are shown in Table 8. On the left side of the table, we compare sentence-level results of different languages in both a multilingual and a retranslated setting for select models. Compared to the performances on the original German dataset version, all models display lower performance in both the multilingual and the retranslated setting for Arabic, French, and Ukrainian. Llama-3-70B shows slightly higher precision for retranslated Arabic (+0.5%) and Ukrainian (+0.1%), however, this comes at a cost of a clearer decrease in recall

		Sentence-level Answers						Question-level Unanswerability					
		Multilingual			German Retrans.			Multilingual			German Retrans.		
Model	Lang.	P	R	F	P	R	F	P	R	F	P	R	F
Mixtral-8x7B	de	74.5	47.1	57.7	–	–	–	68.9	56.4	62.0	–	–	–
	ar	72.5	42.7	53.8	77.8	45.2	57.2	62.8	49.1	55.1	55.4	56.4	55.9
	fr	74.2	43.7	55.0	75.0	45.2	56.4	64.1	45.5	53.2	57.4	49.1	52.9
	uk	69.3	46.4	55.6	74.7	45.8	56.8	73.2	54.5	62.5	58.2	58.2	58.2
Llama-3-70B	de	**85.5**	46.6	60.3	–	–	–	69.8	67.3	68.5	–	–	–
	ar	80.9	42.2	55.5	**86.0**	44.1	58.3	71.4	54.5	61.9	61.0	65.5	63.2
	fr	84.1	44.9	58.5	84.3	43.5	57.4	72.9	63.6	68.0	63.8	**67.3**	65.5
	uk	82.4	41.3	55.0	85.6	43.3	57.5	**74.5**	63.6	**68.6**	**64.9**	**67.3**	**66.1**
DeBERTa	de	62.6	**62.4**	**62.5**	–	–	–	56.2	65.5	60.5	–	–	–
	ar	63.3	54.9	58.8	65.2	53.5	58.8	43.4	60.0	50.4	44.0	**67.3**	53.2
	fr	66.3	56.9	61.2	61.4	**59.9**	**60.6**	50.7	67.3	57.8	53.8	63.6	58.3
	uk	54.7	61.4	57.9	62.2	55.9	58.8	57.1	**72.7**	64.0	48.7	**67.3**	56.5

Table 8: Test set performance (in %) of zero-shot LLMs and finetuned DeBERTa on sentence-level answer extraction (left) and detection of unanswerable questions (right) for multilingual and retranslated settings. In the multilingual setting, questions and documents are machine translated to the respective language. In the retranslated setting, the question is retranslated back to German and paired with the original German document. The best result in each column is **bolded**.

(−2.5% and −3.3% respectively). For the multilingual setting, French results were the closest to German. With exception to Mixtral-8x7B, the F1-score for French is at least 2% higher. Similarly, while retranslating improves F1-score performance compared to directly querying the LLM for Arabic and Ukrainian in all settings by up to +3.4%, retranslating French comes at a performance loss for Llama-3-70B and DeBERTa. Mixtral-8x7B, on the other hand, shows a performance improvement (+1.4%) for retranslating French to German, although it is explicitly advertised as "fluent in French."[14] The biggest performance loss is displayed by Llama-3-70B in the multilingual setting in Ukrainian (−5.3%) and Arabic (−4.8%).

In general, the observed performance differences are observable but not as notable as expected. This is especially the case for Arabic and Ukrainian, as the differences in the alphabet, grammar, and language origins are significant. While machine translation seems to have a slightly better performance for these languages, a performance deterioration compared to the original German dataset is still measurable. However, the questions are translated twice in our setup, and, as a consequence, the actual implications should be smaller.

E.2 Question-Level Unanswerability

As in section 4.1, we infer question-level unanswerability from sentence-level answer extraction results. If no sentence of a document is marked as answer, we treat the question as unanswerable given the document. In contrast to question-level answer extraction, the German results are not necessarily better than those of other languages in the multilingual setting, but they always outperform the retranslated results. Surprisingly, all models perform slightly better in the Ukrainian multilingual setting than on the original German dataset (up to +3.5%, DeBERTa) and mostly considerably better than on Arabic and French (up to +13.6%). Especially Ukrainian precision is high among all models, which is in line with low precision on the sentence-level, i.e., more sentences are marked as answer. Retranslating only yields small performance improvements for French for DeBERTa and for Arabic for all models. Otherwise, directly querying models leads to better question-level results (up to +7.5%).

[14] https://mistral.ai/technology#models