Indirect Question Answering in English, German and Bavarian: A Challenging Task for High- and Low-Resource Languages Alike

Miriam Winkler[▲], **Verena Blaschke**[▲,■], **Barbara Plank**[▲,■]
[▲]MaiNLP lab, Center for Information and Language Processing, LMU Munich, Germany
[■]Munich Center for Machine Learning, Germany
mi.winkler.wald@gmail.com, verena.blaschke@cis.lmu.de, b.plank@lmu.de

Abstract

Indirectness is a common feature of daily communication, yet is underexplored in NLP research for both low-resource as well as high-resource languages. Indirect Question Answering (IQA) aims at classifying the polarity of indirect answers. In this paper, we present two multilingual corpora for IQA of varying quality that both cover English, Standard German and Bavarian, a German dialect without standard orthography: InQA+, a small high-quality evaluation dataset with hand-annotated labels, and GenIQA, a larger training dataset, that contains artificial data generated by GPT-4o-mini. We find that IQA is a pragmatically hard task that comes with various challenges, based on several experiment variations with multilingual transformer models (mBERT, XLM-R and mDeBERTa). We suggest and employ recommendations to tackle these challenges. Our results reveal low performance, even for English, and severe overfitting. We analyse various factors that influence these results, including label ambiguity, label set and dataset size. We find that the IQA performance is poor in high- (English, German) and low-resource languages (Bavarian) and that it is beneficial to have a large amount of training data. Further, GPT-4o-mini does not possess enough pragmatic understanding to generate high-quality IQA data in any of our tested languages.

Keywords: Indirect Question Answering, Low-resource Dialect, LLM Data Generation, Data Augmentation, Corpora and Annotation

1. Introduction

Indirect Question Answering (IQA) is about recognizing the polarity of indirect answers, as illustrated in Table 1. Our motivation to study the task in both standard and non-standard languages lies in the practical relevance of IQA for communication with digital assistants, where users often interact in everyday, non-standard varieties. As Blaschke et al. (2024c) show, German dialect speakers have interest in using conversational technologies in their native language – for which indirectness is as relevant as it is common. For example, Sanagavarapu et al. (2022) find 77 % of 5-minute telephone conversations contain at least one yes/no question with an indirect answer. However, a large number of NLP (Natural Language Processing) tasks that are relevant for digital assistants are primarily researched in English. In contrast, Bavarian is a low-resource, non-standardized German dialect. It has recently been the subject in several NLP studies, including for tasks like slot and intent detection (van der Goot et al., 2021; Winkler et al., 2024; Krückl et al., 2025; Blaschke et al., 2026), which can be pragmatically demanding. However, no IQA datasets for Bavarian exist yet.

In general, IQA is understudied, especially in the multilingual or low-resource domain (§2). Performance of limited previous work systems is usually rather poor and not up-to-par with the performance of tasks like POS tagging in monolingual (Awwalu et al., 2020; Li et al., 2022; Mohammad et al., 2025), multilingual (Snyder et al., 2008; Naseem et al., 2009; Plank et al., 2016) and low-resource (Mitri et al., 2025; Haq et al., 2025; Deore et al., 2025) settings, or natural language inference (Bowman et al., 2015; Storks et al., 2020; Ning et al., 2025).

	I'm early?	A little.
	Komm ich zu früh?	Ein bisschen.
	Bin i z'fria?	A bissal.

Table 1: Parallel example sentences from InQA+ in all three available languages. The indirect answers correspond to *Yes*.

Although the performance of IQA is not outstanding even in standard languages like English, the qualitative challenges underlying the task are rarely discussed beyond quantitative scores. Yet especially for tasks requiring deep pragmatic understanding, non-trivial language phenomena can introduce unexpected difficulties for modelling.

Contributions

- We release two datasets, InQA+ (§3) and GenIQA (§4),[1] in English, German and Bavarian with pairs of questions and indirect answers featuring natural translations on the one hand

[1]https://github.com/mainlp/Multilingual-IQA

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4268–4289
11–16 May 2026.

and artificial data generated by GPT-4o-mini (OpenAI, 2024) on the other.

- We experiment with and test the effect of data quality and quantity (§6) and demonstrate how to alleviate challenges arising during the experimentation.
- We analyse and discuss the implications of indirectness (§7) in an overview of the IQA task and why it is so difficult for humans and state-of-the-art large language models (LLMs) alike (§§3.2 and 4.2).

2. Related Work

Current research seldomly reflects on the specific challenges of IQA, especially the role of indirectness. Since understanding implicit messages relies on good pragmatic capabilities, dataset design can strongly influence model performance. The following related works not only include available resources but also highlight the diverse challenges of indirectness in NLP tasks.

IQA resources are good, but sparse. Even in English, IQA is one of the lesser researched tasks, with three notable data resources available: Circa (Louis et al., 2020), Friends-QIA (Damgaard et al., 2021) and SwDA-IA (Sanagavarapu et al., 2022). Friends-QIA and SwDA-IA draw their data from existing resources like TV show scripts or phone call transcriptions, whereas Circa was created by large-scale crowdsourcing. The datasets consist of polar questions with one or multiple indirect answers in English which are annotated with labels denoting the polarity of the indirect answer.

Multilingual IQA work is even more rare: Wang et al. (2023) created a high-quality but unavailable multilingual corpus for 8 languages. IndirectQA EN (Müller and Plank, 2024) provides the first publicly available parallel multilingual dataset in English, French and Spanish. The data stems from parallel subtitles from the opensubtitles v2018 corpus[2] (Lison and Tiedemann, 2016) and is also hand-annotated. It serves as the base of InQA+ (§3). In our experiments, we replicate the experiments of Müller and Plank (2024)[3] and get close results.

Analysing the IQA performance in previous work grants insights into possible challenges and performance to be expected. All datasets use 5 to 9 labels for classification. For IQA, we observe from prior research that, for example, the size of the context around the indirect answer matters for the performance (Louis et al., 2020; Damgaard et al., 2021). More context can increase the performance of both machines and human annotation. Paci et al. (2025) show this for implicit message understanding in an Italian politics domain, which could inform IQA, too.

Indirectness is diverse. Indirectness is involved in a variety of NLP tasks, most prominently in sarcasm detection (Zhang et al., 2021; Jayaraman et al., 2022; Khan et al., 2025), implicit message understanding (Paci et al., 2025) and IQA (cf. previous paragraph). Because indirectness is a common occurrence in everyday human communication, it is prevalent in the real-world.

Indirectness can challenge model performance, as Paci et al. (2025) underscore with their work on implicit political message understanding: indirectness occurs frequently and many models struggle with the required pragmatic understanding.

3. InQA+: Hand-Curated Test Dataset

InQA+ is our multilingual extension of IndirectQA (Müller and Plank, 2024; §2). It serves as a small, yet high-quality test dataset. It consists of 438 hand-annotated question-answer pairs sourced from the opensubtitles v2018 (Lison and Tiedemann, 2016) corpus (examples in Table 2). The size is derived from the French and Spanish data splits of IndirectQA. The question answer pairs are parallel in English, Standard German and Bavarian. InQA+ is completely new and separate from IndirectQA. The English sentences are distinct from the sentences in IndirectQA because only a few sentences from IndirectQA were parallel with the German opensubtitles v2018 (Lison and Tiedemann, 2016) corpus, so no direct extension to IndirectQA could be produced. The data statement can be found in Appendix A.1.

As motivated in §1, we want to contribute more Bavarian data resources to the NLP space. English as the highest resource language builds the baseline for our experiments, whereas Standard German rounds off our selection as another well-researched standard language that is closely related to Bavarian.

To find suitable candidate pairs, we lightly prefilter the raw opensubtitles v2018 corpus, that is parallel between English and German, for questions that are followed by a line that does not contain the direct answer particles *Yes* or *No*. We only select indirect answers that sound plausible as an answer to the preceding question and that do not contain any direct answer particles including colloquial ones like *Yeah* or *Nope* (also see Appendix A.1). We

[2] http://www.opensubtitles.org/

[3] Müller and Plank (2024) compare the IQA performance across test genres (comedy vs. crime). We do not see any distinctive performance differences between the genres and provide our per-genre results in Appendix B.1.

1 - Yes This is Long Tieng, right? *Right.*	**2 - No** You hungry? *Just black coffee for me.*	**3 - Conditional Yes** Will you be home tonight? *I will be if you want me to be.*
4 - Neither yes nor no Is anyone still in there? *I don't know.*	**5 - Other** Another trick question, right? *Open the window.*	**6 - Lacking context** Hey, you guys want sangria? *It's Guys' Night!*

Table 2: Example questions and *answers* from InQA+ illustrating each **label**'s meaning.

manually label the data with the label set described in §3.1 and show the label distribution in Table 3.

3.1. Annotations and label definitions

We use the same label set as IndirectQA (Müller and Plank, 2024) to maintain data compatibility. The labels are defined as follows:

1. *Yes*: A clear yes or all gradients of yes (including weaker forms, e.g., maybe yes).
2. *No*: A clear no or all gradients of no.
3. *Conditional Yes*: A yes that only holds if certain conditions are true.
4. *Neither Yes nor No*: A neutral answer that lies in the middle of yes and no.
5. *Other*: The sentence does not match the questions as an answer.
6. *Lacking Context*: Without further context, the answer cannot be clearly categorized.

We illustrate the meaning of each label further with the examples in Table 2. The label set is a modification from the six labels of Louis et al. (2020) and Damgaard et al. (2021). *Lacking Context* replaces the *N/A* label from Damgaard et al. (2021). Although it sounds similar, it is different from *Neither Yes nor No* and *Other*: Answers labelled as *Lacking Context* have a clear polarity, unlike *Neither Yes nor No*, and does function as an answer the question, unlike *Other*, only the position on the scale from yes to no is unclear without further context.

For pragmatically difficult cases, like questions with negations (negative questions), we set additional guidelines to ensure uniform annotations. We follow polarity-based annotation (Holmberg, 2012, 2013), focusing on the grammatical form of the question (e.g., positive or negative wording) rather than the polarity of its meaning. For example, the question *Are you not Moby?* from InQA+ can be answered with either *(Yes) I am Moby* or *(No) I am not Moby*, leading to *Yes* or *No* annotations respectively. In contrast, the answer *Yes, I am not Moby* falls under truth-based polarity (Kuno, 1975; Holmberg, 2012, 2013). Appendix B.2 contains more detailed annotation guidelines and examples.

Label	InQA+	GenIQA		
Yes	183	400	250	218
No	100	132	210	182
Conditional Yes	18	319	274	306
Neither Yes nor No	79	477	464	468
Other	34	71	141	191
Lacking Context	24	101	161	135
Total	438	1.500		

Table 3: Label distributions of InQA+ and GenIQA. The distribution of InQA+ applies to all languages due to the parallelism.

Label set variations To investigate the effect of the label set complexity, we conduct experiments with two simplified versions of the label set.

In the **REMAPPED** version, we merge labels that are similar to each other: firstly, *Yes* and *Conditional Yes* (which both denote instances that signify gradients of *Yes*), and secondly, *Other* and *Lacking Context* (which contain samples that are unclear in their function as an answer). This creates sharper distinctions between clear and vague answers. REMAPPED consists of four labels.

The **YESNO** version only contains instances with the labels *Yes* and *No*. This is the most minimal version of IQA. For InQA+, we omit all instances with other labels (leaving 283 instances). For GenIQA, we generate a dedicated YESNO version in each language to preserve the dataset size of 1,500.

We report results for these label set variations in §6.3 and discuss their benefits and drawbacks in §7.

3.2. Inter-annotator agreement

IQA is a pragmatically hard task. We derive this not only from the statements of previous work (Damgaard et al., 2021; Müller and Plank, 2024), but also from our finding that humans and machines alike find it hard to classify indirect answers as the Inter-Annotator Agreements (IAA) shows.

In our hand-labelled InQA+ data, we obtain a moderate agreement with a Cohen's Kappa of 0.53 on a sample of 100 InQA+ EN instances which

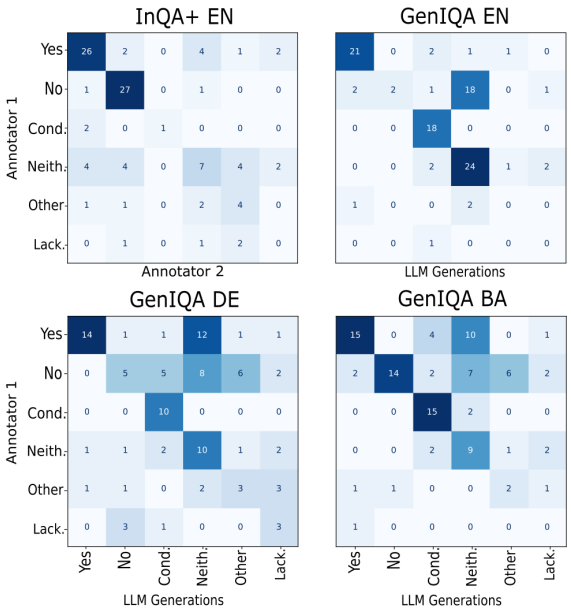


Figure 1: Confusion matrices between two annotators (top left) and the GENIQA labels as originally generated vs. re-annotated by the main annotator, respectively on 100 sentences from each dataset. Cond. = Conditional Yes; Neith. = Neither Yes nor No; Lack. = Lacking Context.

were double-annotated by expert annotators. In other IQA resources, we observe similarly moderate agreements, for example in IndirectQA (Müller and Plank, 2024) with a Cohen's Kappa of 0.54 on 205 double-annotated pairs and Circa (Louis et al., 2020) with a Fleiss' Kappa of 0.61.

Figure 1 (top left) shows that we have the most agreement for cases that are either clearly *Yes* or *No* and low agreement for every other label. Our IAA is $\kappa = 0.57$ for REMAPPED and $\kappa = 0.70$ for YESNO.

4. GENIQA: Artificial Training Dataset

As the availability of IQA data is limited, especially for low-resource languages, we experiment with LLM-generated training data. We create GENIQA, which consists of 1,500 question-answer pairs which we generate with GPT-4o-mini (OpenAI, 2024) in English, Standard German and Bavarian. The data statement can be found in Appendix A.2. All languages were generated independently and not translated. The pairs were annotated by the model at generation time with the same label set as INQA+ (§3.1). The label distribution per language is found in Table 3.

As preliminary experiments showed high language quality for generated English and Standard German texts for many LLMs, we focus on Bavarian for selecting an LLM to generate GENIQA, as it is a priori less clear which LLM works for Bavarian. We perform experiments on dialect generation with 9 LLMs[4] to find a model that can generate Bavarian dialect. GPT-4o-mini was capable of generating the best dialect output of all tested models, albeit still not of high quality as we discuss in the next Section. For comparability between languages, we use the same prompt to also use GPT-4o-mini to generate the English and Standard German datasets. The prompt includes six examples, one per label (more details in Appendix C). We generate 50 question-answer pairs per prompt until we reach 1,500. We remove duplicates and need 46 iterations for GENIQA EN, 34 for DE and 33 for BAR.

Dataset variations As with INQA+, we also experiment with reduced label set variations (REMAPPED, YESNO; §3.1). To compare the performance of IndirectQA EN (Müller and Plank, 2024) and GENIQA trained models on a more equal basis, we perform experiments with a sized down version of GENIQA EN: **GENIQA EN SMALL**. Like IndirectQA EN, it consists of 615 instances and exhibits a similar label distribution to create comparable conditions.

4.1. Bavarian GENIQA language quality

Generally, the quality of the generated dialect output is low and the LLMs we tested are not capable of generating authentic and natural-sounding dialect. We investigate the quality of the generated data both manually and by surveying a larger set of dialect speakers.

Manual inspections reveal that many instances of the Bavarian GENIQA dataset either only contain one dialect word, use auxiliary verbs incorrectly or choose the wrong genus of the determiner. Furthermore, the dialect is not generated consistently. In a sample of 100 instances, 70 % of questions and 6 % of answers were Standard German, as classified by a native dialect speaker. We speculate that a possible reason could be that questions are much rarer than indicative statements in the pre-training data for GPT-4o-mini, and questions in a low-resource dialect might be especially rare. The rest of the data contained at least some nuances of dialect. The generated dialect also often resembles colloquial German more than genuine Bavarian. Nevertheless, some instances contain generations of correct characteristic Bavarian gram-

[4] gemma-2-2b-it (Gemma Team, 2024), Llama-3.2-1B-Instruct (Meta, 2024), OLMo-2-1124-7B-Instruct (OLMo et al., 2024), GPT-4o-mini (OpenAI, 2024), LLäMmlein_1B (Pfister et al., 2024), Betzerl (CAIDAS Uni Würzburg, 2024) and two sizes of EuroLLM (Martins et al., 2025) (EuroLLM-1.7B-Instruct and EuroLLM-9B-Instruct), more details about the LLM dialect experiments and prompt tuning is provided in Appendix C.

mar phenomena like clitics, the merging of the word *es* (ENG. it) into the preceding words, like *is's* (DEU. ist es; ENG. is it). We provide more insights into the manually assessed quality in Appendix D.

We confirm the poor quality we observed by surveying native speakers. The survey was hosted via SoSci Survey (Leiner, 2025) and distributed via various social channels including Discord, Whatsapp and Reddit. 76 active dialect speakers, the majority stemming from the Upper and Lower Bavarian regions, answered the survey and assessed ten indirect answers on whether they include dialect and if they find the answer authentic in that they themselves would produce such a sentence; nine of those answers are LLM-generated, one is a hand-translated instance from InQA+.

When mapping the scales to numbers (0 = worst, 3 = best), the LLM-generated answers of GPT-4o-mini received a dialectness rating of 1.1 and a authenticity rating of 1.4 on average (compared to the hand-translated instance with ratings of 2.5 and 1.8),[5] confirming the low quality. However, we do not strife to create a high-quality artificial dataset. Instead our GenIQA experiments demonstrate the current state of LLMs with regards to the Bavarian dialect in a naive and intuitive approach. The full questionnaire, participant information and results can be found in Appendix D.

4.2. Labelling accuracy

We manually re-annotate 100 instances of GenIQA. Our GenIQA datasets only reach annotation accuracies of 65 % for EN, 45 % for DE and 55 % for BAR. The overall quality is low with the models performing best on English generations. LLMs are good at generating coherent texts that fulfil social expectations. However, as soon as more complex phenomena like indirectness or references are involved, they struggle (Hu et al., 2023; Ma et al., 2025). Qiu et al. (2023) found that ChatGPT does not process implicatures like humans do, because it cannot switch between pragmatic and semantic interpretations easily.

The label distributions differ across the language splits, as visualised in Figure 1. For each split, the labels *Yes*, *Conditional Yes* and *Neither Yes nor No* have the highest accuracy with respect to the manual annotations, unlike the IAA between two humans, who have high agreement for *No* as well.

5. Experimental Setups

We explore the effect of data quantity and quality with three fine-tuneable multilingual models in their base sizes for comparability: mBERT (Devlin et al., 2019) is a widely used baseline for many experiments in a lot of research papers, for example in the related Circa (Louis et al., 2020) and IndirectQA (Müller and Plank, 2024) research, sarcasm research of Jayaraman et al. (2022) and Zhang et al. (2021) and Bavarian tasks like slot and intent detection (van der Goot et al., 2021; Winkler et al., 2024). We complement it with XLM-RoBERTa (XLM-R) (Conneau et al., 2020) and mDeBERTa (He et al., 2020), which are used in other Bavarian research by Peng et al. (2024) and Krückl et al. (2025). mBERT is the only model which has seen Bavarian training data (from Wikipedia) during pre-training.[6] We refrain from using instruction-tuned GPT models as we saw low label generation quality (§4.2), and because fine-tuned encoder models often lead to better classification results than prompting instruction-tuned models (Mosbach et al., 2023; Qorib et al., 2024; Weller et al., 2025).

We have multiple training setups to test the effects of data quality, language and quantity in different scenarios. In the first setup (§6.1), we compare training on manually vs. artificially created data. In further experiments, we vary the size of the training split (§6.2) by adding Friends-QIA (Damgaard et al., 2021) to the training data and use the dataset versions with the reduced label sets (§6.3). The datasets' annotations are compatible without further processing. The data is stored in text files with one question, indirect answer and label per line, separated by a tab space.

For each training setup, we fine-tune the hyperparameters learning rate, batch size, warm-up ratio, weight decay, dropout rate and label smoothing in the scope of the search spaces in Table 4 individually. We choose hyperparameter values based on accuracy. The optimal hyperparameters differ across datasets, label set variations and language models. We explain the fine-tuning method and details in Appendix E. We employ grid and random search to find the best set of parameters. We observe that the grid-searched parameters yield better dev accuracy than the random-searched ones. Additionally, to avoid overfitting as much as possible,

Parameter	Grid Seach	Random Search
Learning Rate	[1e-4, 1e-5, 1e-6]	1e-6—5e-4
Batch Size	[4, 8, 16, 32]	[4, 8, 16, 32]
Warm-up Ratio	[0.1, 0.3, 0.5]	0.1—0.5
Weight Decay	[0.001, 0.01, 0.1]	0.001—0.1
Dropout Rate	[0.1, 0.3, 0.5]	0.1—0.5
Label Smoothing	[0.1, 0.3, 0.5]	0.05—0.5

Table 4: Hyperparameter search spaces for grid and random search.

[5] The ratings are higher when only considering answers by respondents from the region that the translator is from (refer to Figure 7 in Appendix D for more details).

[6] https://github.com/google-research/bert/blob/master/multilingual.md

we employ early stopping in training.

We train each model on three random seeds and report mean accuracy, macro-F1 and standard deviations, as the scores vary greatly per seed and single runs are unreliable and noisy. We transparently show the capabilities of SOTA models and discuss them with regards to high deviation.

6. Results and Analysis

Since the IQA task is challenging, we focus our experiments on the effects of data quality and quantity to see which is more influential to reduce overfitting. For the analysis of a difficult task, it is important to take multiple metrics into account, in our case: accuracy and F1 scores. The performance cannot always be read directly from the accuracy, and the combination with F1 scores reveals more model capabilities.

6.1. Data quality versus quantity

We first present our main investigation: comparing small high-quality training data (IndirectQA EN; Müller and Plank, 2024) against larger, but low-quality artificial training data (GenIQA) to research the trade-off of high quality versus high quantity.

Using manual versus generated data When analysing the results of mBERT in Table 5, we observe a major issue: all scores lie below the majority baseline accuracy (41.8 %) due to overfitting and the F1 scores are low, even if they almost always surpass the majority class baseline F1.

The highest accuracy score (38.6 %) is achieved by the model trained with IndirectQA EN when evaluated on the Bavarian InQA+ test set. This score is close to the majority class baseline due to a bias for the majority class, revealing that the training was not effective, which the low F1 score (13.4 %) also indicates. In previous work, we see similar behaviour and discuss this in §7.

From the mBERT results, the models trained with the artificial but larger GenIQA datasets have lower accuracy than the IndirectQA EN models, but the higher diversity in the data (stemming from the larger dataset size) benefits the generalisation performance on the in-language test data.

The best F1 scores for the test sets stem from the models trained on their respective GenIQA training dataset (16.7 % for InQA+ EN and 16.2 % for InQA+ BAR). The only exception to this is InQA+ DE, where the model trained with GenIQA BAR reaches the best performance. This is likely due to the generated "Bavarian" data largely resembling German (§4.1).

With mDeBERTa (He et al., 2020) and XLM-R (Conneau et al., 2020), results only approach the majority class baseline (highest score: 41.3 % accuracy with XLM-R). While accuracy of GenIQA trained models lags behind the performance of IndirectQA EN, mDeBERTa trained on GenIQA EN performs just as well (accuracy) or even better (F1) than IndirectQA EN. XLM-R produces the highest accuracies for all languages when fine-tuned with the small IndirectQA EN. Further, we see large increases in the F1 scores when fine-tuning with any GenIQA dataset, indicating that the larger dataset size might be beneficial for learning class representations beyond the majority class. Many of highest F1 scores are produced by models trained on the Bavarian version of GenIQA. This might be due to that split having a more uniform label distribution than the English and German splits (Figure 1).

As the overall performance is already unsatisfactory in English with no accuracy scores surpassing the majority class baseline accuracy, we see even worse performance for Standard German and the low-resource dialect Bavarian (Table 5). In the standard setups, the small yet high-quality training data IndirectQA EN yielded the best accuracy scores for Standard German (32.7–41.3 %) and Bavarian (36.5–40.2 %). However, the F1 scores are again better with the larger GenIQA datasets, especially GenIQA BAR (on InQA+ DE: 15.7–21.6 %; on InQA+ BAR: 16.2–19.3 %).

Because observed trends are similar across models, and to save compute, subsequent experiments are only with mBERT.

Performance on training and development sets Our greatest challenge is overfitting which occurred in each experimental setup. It becomes visible when comparing the accuracy scores of each model on the training and development sets.

All train (accuracy between 48.1–72.3 %) and dev scores (accuracy between 40.1–62.4 %) lie above the majority class baseline accuracy of 30.9–35.6 %, depending on the dataset. This shows that the training itself works. We provide the full results table and further discussion in Appendix F. That the scores are low even for English experiment setups underscores the task difficulty of IQA.

The gaps of 2.2–12.4 percentage points (pps.) between train and dev accuracies reveal the grade of overfitting: the larger the gap, the more overfitting took place. We observe the largest discrepancies for XLM-R models fine-tuned with GenIQA EN and GenIQA BAR with gaps of 12.4 and 11.0 pps. between the train and dev set accuracy.

6.2. Assessing the effect of dataset size

Since we are interested in the effect of data quality versus quantity, we next experiment with datasets of varying sizes to specifically target characteris-

	Accuracy			Macro F1		
	INQA+ EN	INQA+ DE	INQA+ BAR	INQA+ EN	INQA+ DE	INQA+ BAR
Majority Class Baseline	41.78	41.78	41.78	9.82	9.82	9.82
mBERT trained with						
IndirectQA EN	34.02 ±2.57	36.30 ±0.60	38.58 ±3.59	15.23 ±1.87	14.51 ±3.05	13.42 ±1.68
GENIQA EN	27.25 ±2.87	23.29 ±7.15	21.31 ±3.49	16.69 ±2.39	9.85 ±6.38	7.60 ±2.63
GENIQA DE	20.09 ±4.56	19.18 ±3.14	20.55 ±1.95	12.41 ±6.19	13.33 ±4.36	10.35 ±3.29
GENIQA BAR	16.59 ±0.92	15.37 ±1.99	18.87 ±1.99	12.01 ±0.84	15.67 ±2.48	16.16 ±2.08
mDeBERTa						
IndirectQA EN	33.26 ±7.42	32.65 ±8.58	36.45 ±2.12	12.93 ±2.67	12.83 ±2.84	11.71 ±0.98
GENIQA EN	33.11 ±9.76	32.04 ±10.97	30.21 ±11.56	19.27 ±3.93	18.56 ±4.44	16.61 ±4.44
GENIQA DE	25.57 ±7.49	25.42 ±7.05	22.98 ±2.42	16.90 ±2.70	17.40 ±3.99	13.11 ±4.13
GENIQA BAR	20.02 ±3.47	17.88 ±3.58	20.47 ±1.94	20.15 ±4.65	18.01 ±3.79	18.22 ±2.57
XLM-R						
IndirectQA EN	39.42 ±2.30	41.25 ±0.57	40.18 ±1.81	14.90 ±4.63	15.05 ±4.87	12.16 ±2.10
GENIQA EN	32.88 ±3.36	29.60 ±3.90	27.63 ±5.48	20.43 ±2.12	18.04 ±2.54	16.80 ±2.83
GENIQA DE	19.71 ±4.03	21.00 ±3.59	22.60 ±3.90	16.56 ±3.34	18.03 ±2.46	15.56 ±4.17
GENIQA BAR	22.15 ±3.84	22.37 ±1.95	22.83 ±2.20	21.35 ±4.44	21.55 ±2.49	19.29 ±0.68

Table 5: Average accuracy and macro F1 scores over three seeds and standard deviations of models trained on our datasets, evaluated on INQA+. The train/dev split follows a ratio of 80-20.

	Accuracy			Macro F1		
	INQA+ EN	INQA+ DE	INQA+ BAR	INQA+ EN	INQA+ DE	INQA+ BAR
Majority Class Baseline	41.78	41.78	41.78	9.82	9.82	9.82
Majority Class Baseline REMAPPED	45.89	45.89	45.89	15.73	15.73	15.73
Majority Class Baseline YESNO	64.66	64.66	64.66	39.27	39.27	39.27
mBERT trained with						
IndirectQA EN REMAPPED	27.02 ±4.53	23.36 ±8.95	23.44 ±11.46	17.10 ±2.54	14.01 ±6.49	13.48 ±7.04
IndirectQA EN + Friends-QIA REMAPPED	47.56 ±2.91	40.56 ±8.59	34.25 ±7.46	35.40 ±2.46	30.13 ±5.69	24.21 ±3.96
IndirectQA EN YESNO	63.96 ±3.59	60.42 ±8.28	60.42 ±8.28	51.26 ±10.45	44.47 ±5.88	44.38 ±5.74
GENIQA EN YESNO	62.78 ±3.34	68.43 ±0.82	64.78 ±0.74	50.25 ±7.85	58.01 ±6.83	46.42 ±7.32
GENIQA DE YESNO	51.59 ±10.41	63.02 ±4.42	59.60 ±4.22	48.81 ±12.30	61.68 ±3.48	55.25 ±2.60
GENIQA BAR YESNO	65.37 ±0.93	65.96 ±1.08	64.19 ±0.41	43.80 ±6.93	56.87 ±3.64	58.30 ±0.54
IndirectQA EN (repeated from Table 5)	34.02 ±2.57	36.30 ±0.60	38.58 ±3.59	15.23 ±1.87	14.51 ±3.05	13.42 ±1.68
Friends-QIA	43.07 ±4.80	35.77 ±8.17	30.29 ±6.90	21.92 ±1.41	17.75 ±3.81	14.16 ±1.29
IndirectQA EN + Friends-QIA	46.27 ±1.94	38.81 ±2.25	31.74 ±5.27	23.43 ±0.89	19.76 ±0.63	16.27 ±2.14
GENIQA EN SMALL	30.82 ±5.31	27.25 ±8.33	24.51 ±4.82	16.64 ±1.92	11.48 ±2.80	10.60 ±2.54

Table 6: Average accuracy and macro F1 scores over three seeds and standard deviations of models trained on experimental dataset variations. The REMAPPED and YESNO setups are evaluated on the corresponding REMAPPED and YESNO INQA+ test sets.

tics like the data quantity. Enlarging the dataset size by adding data from different datasets can reveal what the original training data is lacking if the performance is not as expected.

We see this first when training with Friends-QIA (Damgaard et al., 2021), which is much larger than IndirectQA (5,930 versus 615) and which yields large accuracy gains over IndirectQA EN (Müller and Plank, 2024) (Table 6) with an increase of 9.1 pps. on INQA+ EN. The transfer capabilities to Standard German and Bavarian display lower accuracy (losses of 0.5–8.3 pps.), but higher F1 scores as the increases of 0.7–6.7 pps. across all languages show. Training on a combined setup of Friends-QIA and IndirectQA EN pushes performance even further with increases of 1.5–3.2 pps. over Friends-QIA alone. Thus, more data is beneficial for IQA in each language. However, at the same dataset size, the quality of the data is still more crucial for the performance, as the accuracy loss of GENIQA EN SMALL (§4) of 3.2-14.1 pps. against IndirectQA EN reveals.

6.3. Reduced ambiguity in label set

Fusing and omitting labels can eliminate ambiguous labels like *Lacking Context* which in many cases, as we observed, could have been resolved by hearing the intonations or seeing the mimics of the speaking actor in the movie, because humans usually distinguish literal meaning from sarcasm through intonations (Glenwright et al., 2014). Without context, it remains ambiguous.

By modifying the label sets, we expect better per-

formance through task simplification and unambiguous labels, as the simplification already increased the human agreement (cf. §3.2). For this reason, we experiment with models fine-tuned on the REMAPPED and YESNO dataset variants presented in §3.1. We evaluate them on the corresponding REMAPPED and YESNO INQA+ test sets.

Altering the data can reduce label ambiguity. However, fusing the labels can also decrease the dataset size and obscure the granularity of the label set and lose nuances, which makes the annotations less informative. These operations can thus also have negative impact on the performance and need to be applied thoughtfully.

While we do not see improvements by training with REMAPPED alone (loss of 7 pps. accuracy in comparison to the full label variant, cf. Table 6), combining it with a REMAPPED version of Friends-QIA (Damgaard et al., 2021) improves the performance clearly with F1 scores twice as high. We inspect this aspect of dataset size further in §6.2

The accuracy of YESNO lies around the majority class baseline. The F1 scores, however, are considerably higher than for the full label and REMAPPED training dataset variants. We observe increases of 34.1–51.9 pps. over the majority class baseline F1, showing the simplicity of the minimalistic setup.

7. Discussion and Learnings

Both our experiments and previous work achieve results around the majority class baseline, confirming that IQA is highly challenging. Even in high-resource languages, the performance is low, and the gap widens in low-resource languages such as Bavarian. We thus share our learnings for IQA.

Data quantity is important. A fundamental challenge we faced was a dataset that was too small for the pragmatically hard IQA task. We can confirm this with successful experiments using the larger combination of IndirectQA EN (Müller and Plank, 2024) and Friends-QIA (Damgaard et al., 2021) (615 + 5,930 instances). Even GENIQA with a size of 1,500 instances is still too small to reach meaningful scores.

From previous research, we deduce that a dataset of at least a size between ∼6,000 (Friends-QIA; Damgaard et al., 2021) and ∼35,000 (Circa; Louis et al., 2020) might be necessary to learn enough pragmatic features from the training data.

Increasing the amount of data is not always possible, thus experimenting with various models is helpful as they each preprocess data differently. In our case, changing the model architecture alone does not improve the scores (§6.1), at least not without also changing the model size. We see this from the research of Sravanthi et al. (2024) who achieve results that even surpass human performance on response classification of answers with implied meaning with very large models like flan-t5-xxl (Chung et al., 2022) and llama-2-70b-chat (Meta, 2024). As we established in §5, smaller sized encoder-LMs do not work well with IQA, which reflects in their performances during the experiments.

Language interpretations and annotations are inconsistent. Indirectness is subjective and ambiguous. As we discussed in §3.2, the annotations for IQA depend on individual interpretations of the indirect answers and the agreement is moderate (0.53). It suffers even more when the label set is ambiguous, meaning that the labels do not have clearly distinguishable boundaries and detailed annotation guidelines. Like this, the same labels can be applied differently in different datasets, which may lead to discrepancies in the annotations between the training data and test data, causing unexpected results. §§3.1 and 3.2 explain that such cases raise the need for clear annotation guidelines that can also help with the annotation reconciliation between data sources.

Many annotations will inevitably remain ambiguous and dependent on different perspectives. We want to note here that differing annotations are not errors per se, but can fall into the scope of human label variation (Plank, 2022). Clear annotation guidelines help in avoiding ambiguities and annotation disagreements, which is why cleaning the data and correctly labelling it in the case of discrepancies may be important before the training and evaluation.

This is not a sole occurrence in IQA, but also persists in other pragmatically challenging tasks about indirect language, like sarcasm detection. The MUStARD (Multimodal Sarcasm Dataset; Castro et al., 2019) and CSC (Conversational Sarcasm Corpus; Jang and Frassinelli, 2024) datasets have similarly low to moderate agreements (0.23–0.59) for scripted language from TV shows and natural sarcasm utterances respectively, which underscores the annotation ambiguity of indirect language that also applies to IQA. Additionally, the perspectivism research of Casola et al. (2024) reveals that for irony detection, the results vary greatly between different demographic groups. It is important to track the annotators' backgrounds and meta information and take it into account for analyses, as the subjectiveness of different groups may influence the annotations.

8. Conclusions

We presented InQA+, a multilingual evaluation dataset, and GenIQA, an LLM-generated training

corpus for IQA in English, German, and Bavarian. Our experiments confirm that IQA is pragmatically hard for high- and low-resource languages and we find that a large dataset is beneficial for good performance. Nevertheless, data quality and annotation/label clarity are decisive factors. Simplified labels help mitigate ambiguities, but come at the cost of loss of informational nuances.

Therefore, we suggest conducting intensive pre-experiments to check that the training dataset leads to models of sufficient quality that do not overfit. Additionally, we recommend to invest in resources and ways to create sufficiently big datasets (also for languages other than English), as this is crucial for the performance.

We will release our new datasets, INQA+ and GENIQA, to enable further research on indirect question answering in high- and low-resource languages and on manually vs. automatically generated datasets.

Limitations

We experiment with a limited selection of two high-resource languages and one low-resource dialect, because IQA resources are sparse as §2 shows. For future work, it is possible to repeat the data acquisition method with the other languages in the opensubtitles (Lison and Tiedemann, 2016) corpus to create more multilingual IQA corpora.

The data we used stems from movie subtitles provided by opensubtitles v2018 (Lison and Tiedemann, 2016) and contains scripted language that is not necessarily naturally occurring. This means that the performance of real-world data might lead to different results.

Ethical Considerations

The Bavarian translation and annotation of the data was performed by an employed worker who is paid according to national standards. The data does not contain any personally identifiable information and is used according to the permissions of the dataset provider.

Acknowledgements

We thank the anonymous reviewers as well as the members of the MaiNLP research lab for their constructive feedback, especially Rob van der Goot and Felicia Körner.

The work presented in this paper was funded by ERC Consolidator Grant DIALECT 101043235.

9. Bibliographical References

Jamilu Awwalu, Saleh El-Yakub Abdullahi, and Abraham Eseoghene Evwiekpaefe. 2020. Parts of speech tagging: A review of techniques. *FUDMA JOURNAL OF SCIENCES*, 4(2):712–721.

Verena Blaschke, Barbara Kovačić, Siyao Peng, and Barbara Plank. 2024a. Maibaam annotation guidelines. *arXiv preprint arXiv:2403.05902*.

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024b. MaiBaam: A multi-dialectal Bavarian Universal Dependency treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10921–10938, Torino, Italia. ELRA and ICCL.

Verena Blaschke, Christoph Purschke, Hinrich Schuetze, and Barbara Plank. 2024c. What do dialect speakers want? a survey of attitudes towards language technology for German dialects. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 823–841, Bangkok, Thailand. Association for Computational Linguistics.

Verena Blaschke, Miriam Winkler, and Barbara Plank. 2026. Standard-to-dialect transfer trends differ across text and speech: A case study on intent and topic classification in German dialects.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

CAIDAS Uni Würzburg. 2024. Betzerl. https://huggingface.co/LSX-UniWue/Betzerl_1B_wiki_preview.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an *obviously* perfect paper).

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Cathrine Damgaard, Paulina Toborek, Trine Eriksen, and Barbara Plank. 2021. "I'll be there for you": The one with understanding indirect answers. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 1–11, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Pallavi R. Deore, Nita V. Patil, and Ajay S. Patil. 2025. Deep learning-based parts-of-speech tagging in marathi language. *Procedia Computer Science*, 258:3771–3780. International Conference on Machine Learning and Data Engineering.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Parul Dubey, Pushkar Dubey, and Pitshou N. Bokoro. 2025. Unpacking sarcasm: A contextual and transformer-based approach for improved detection. *Computers*, 14(3).

Libuše Dušková. 1981. Negative questions in english. *International Review of Applied Linguistics in Language Teaching*, 19(1-4):181–194.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size.

Melanie Glenwright, Jayanthi M. Parackel, Kristene R. J. Cheung, and Elizabeth S. Nilsen. 2014. Intonation influences how children and adults interpret sarcasm. *Journal of Child Language*, 41(2):472–484.

Ijazul Haq, Yingjie Zhang, and Intakhab Alam Qadri. 2025. Pos tagging of low-resource pashto language: annotated corpus and bert-based model. *Lang Resources & Evaluation 59*, pages 3243—-3265.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Anders Holmberg. 2012. The syntax of negative questions and their answers. *Proceedings of GLOW in Asia IX*.

Anders Holmberg. 2013. The syntax of answers to polar questions in english and swedish. *Lingua*, 128:31–50.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models.

Hyewon Jang and Diego Frassinelli. 2024. Generalizable sarcasm detection is just around the corner, of course! In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4238–4249, Mexico City, Mexico. Association for Computational Linguistics.

Ashok Kumar Jayaraman, Tina Esther Trueman, Gayathri Ananthakrishnan, Satanik Mitra, Qian Liu, and Erik Cambria. 2022. Sarcasm detection in news headlines using supervised learning. In *2022 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, pages 288–294.

Shumaila Khan, Iqbal Qasim, Wahab Khan, Khursheed Aurangzeb, Javed Ali Khan, and Muhammad Shahid Anwar. 2025. A novel transformer attention-based approach for sarcasm detection. *Expert Systems*, 42(1):e13686.

Xaver Maria Krückl, Verena Blaschke, and Barbara Plank. 2025. Improving dialectal slot and intent detection with auxiliary tasks: A multi-dialectal Bavarian case study. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 128–146, Abu Dhabi, UAE. Association for Computational Linguistics.

Xaver Maria Krückl, Verena Blaschke, and Barbara Plank. 2025. Improving dialectal slot and intent detection with auxiliary tasks: A multi-dialectal bavarian case study.

Susumu Kuno. 1975. The structure of the japanese language. *Foundations of Language*, 13(3).

Dominik J. Leiner. 2025. Sosci survey (version 3.7.06). https://www.soscisurvey.de.

Hongwei Li, Hongyan Mao, and Jingzi Wang. 2022. Part-of-speech tagging with rule-based data preprocessing and transformer. *Electronics*, 11(1).

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "I'd rather just go to bed": Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, Online. Association for Computational Linguistics.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. de Souza, Alexandra Birch, and André F.T. Martins. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62. Proceedings of the Second EuroHPC user day.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Aiom Minnette Mitri, Eusebius Lawai Lyngdoh, Sunita Warjri, Goutam Saha, Saralin A. Lyngdoh, and Arnab Kumar Maji. 2025. Probing a pretrained roberta on khasi language for pos tagging. *Natural Language Processing*, 31(2):230–249.

Abdukarim Mohammad, Mohammed Abdullahi, and Jerome Aondongu Achir. 2025. Improving part-of-speech tagging with relative positional encoding in transformer models and basic rules. *Indonesian Journal of Data and Science*, 6(1):10–19.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Christin Müller and Barbara Plank. 2024. IndirectQA: Understanding indirect answers to implicit polar questions in French and Spanish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9025–9035, Torino, Italia. ELRA and ICCL.

Tahira Naseem, Benjamin Snyder, Jacob Eisenstein, and Regina Barzilay. 2009. Multilingual part-of-speech tagging: Two unsupervised approaches. *Journal of Artificial Intelligence Research*, 36:341–385.

Meiling Ning, Zhongbao Zhang, Junda Ye, Jiabao Guo, and Qingyuan Guan. 2025. Better language model-based judging reward modeling through scaling comprehension boundaries.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. 2 olmo 2 furious.

OpenAI. 2024. Openai api documentation.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2017. Creating and characterizing a diverse corpus of sarcasm in dialogue.

Walter Paci, Alessandro Panunzi, and Sandro Pezzelle. 2025. They want to pretend not to understand: The limits of current llms in interpreting implicit content of political discourse.

Siyao Peng, Zihang Sun, Huangyan Shan, Marie Kolm, Verena Blaschke, Ekaterina Artemova, and Barbara Plank. 2024. Sebastian, Basti,

Wastl?! recognizing named entities in Bavarian dialectal data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14478–14493, Torino, Italia. ELRA and ICCL.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2024. Llämmlein: Compact and competitive german-only language models from scratch.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Zhuang Qiu, Xufeng Duan, and Zhenguang Cai. 2023. Does ChatGPT resemble humans in processing implicatures? In *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*, pages 25–34, Nancy, France. Association for Computational Linguistics.

Muhammad Reza Qorib, Geonsik Moon, and Hwee Tou Ng. 2024. Are decoder-only language models better than encoder-only language models in understanding word meaning? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16339–16347, Bangkok, Thailand. Association for Computational Linguistics.

Krishna Sanagavarapu, Jathin Singaraju, Anusha Kakileti, Anirudh Kaza, Aaron Mathews, Helen Li, Nathan Brito, and Eduardo Blanco. 2022. Disentangling indirect answers to yes-no questions in real conversations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4677–4695, Seattle, United States. Association for Computational Linguistics.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1041–1050, Honolulu, Hawaii. Association for Computational Linguistics.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2020. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches.

Khaled Mahmud Sujon, Rohayanti Hassan, Kwonhue Choi, and Md Abdus Samad. 2025. Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. *Journal of Big Data*, 12(268).

Dennis Ulmer, Elisa Bassignana, Max Müller-Eberstein, Daniel Varab, Mike Zhang, Rob van der Goot, Christian Hardmeier, and Barbara Plank. 2022. Experimental standards for deep learning in natural language processing research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2673–2692, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zijie Wang, Md Hossain, Shivam Mathur, Terry Melo, Kadir Ozler, Keun Park, Jacob Quintero, MohammadHossein Rezaei, Shreya Shakya, Md Uddin, and Eduardo Blanco. 2023. Interpreting indirect answers to yes-no questions in multiple languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2210–2227, Singapore. Association for Computational Linguistics.

Helmut Weiß. 1998. *Syntax des Bairischen: Studien zur Grammatik einer natürlichen Sprache*. Max Niemeyer Verlag.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025. Seq vs seq: An open suite of paired encoders and decoders.

Miriam Winkler, Virginija Juozapaityte, Rob van der Goot, and Barbara Plank. 2024. Slot and intent detection resources for Bavarian and Lithuanian: Assessing translations vs natural queries to digital assistants. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14898–14915, Torino, Italia. ELRA and ICCL.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Yazhou Zhang, Yaochen Liu, Qiuchi Li, Prayag Tiwari, Benyou Wang, Yuhua Li, Hari Mohan Pandey, Peng Zhang, and Dawei Song. 2021. Cfn: A complex-valued fuzzy network for sarcasm detection in conversations. *IEEE Transactions on Fuzzy Systems*, 29(12):3696–3710.

A. Data Statements

The information in §A.3 and §A.4 applies to both InQA+ (§A.1) andGenIQA (§A.2).

A.1. InQA+ Data Statement

- Dataset name and version: InQA+ (EN, DE and BAR) (version 0.1)
- Dataset curators: Miriam Winkler, Verena Blaschke, Barbara Plank
- Dataset and data statement citation: Please cite this paper.
- Data statement version: 1.0

Executive summary and curation rationale. This dataset extends the work of Müller and Plank (2024) and adds a new Indirect Question Answering (IQA) resource. It contains 438 indirect question-answers per language in English, Standard German and Bavarian, a German dialect. It serves as an evaluation dataset for IQA.

Documentation for source datasets. We use data from the opensubtitles v2018 corpus (Lison and Tiedemann, 2016). The subtitles stem from the website `http://www.opensubtitles.org/`, which made the data available for research purposes. The subtitles are aligned between different languages.

The dataset consists of questions and indirect answers from movie scripts with numeric labels indicating the polarity of the indirect answers. They are separated by tabspaces with one question-answer pair and label per line.

Date. The movies stem from different years and are identifiable through the meta data provided in the opensubtitles v2018 corpus (Lison and Tiedemann, 2016). The movies' release dates range from the 19th to 21st century.

Modality. Written (movie subtitles).

Genre. Questions and indirect answers from movie scripts.

Dataset statistics. The dataset consists of 438 English sentences and a German and Bavarian translation for each. It consists only of a test split for each language.

Labels. Each question-answer pair carries a numerical label indicating the polarity of the answer, ranging from the gradients of *Yes* and *No* to middle and out-of-context labels. The meaning of the labels is explained in Table 2 and the annotation details can be found in §3.1.

Language varieties and annotator demographics. The dataset contains data in English (en-EN), Standard German (de-DE) and a dialectal variant of German, Bavarian (de-BAR). The Bavarian translation depicts the dialectal variant of the border area between rural Upper and Lower Bavaria. The translations were carried out by a native speaker of Standard German and Bavarian in her 20s.

Linguistic situation and data quality. The English and German language as it occurs in the dataset is scripted, as it stems from movie scripts of various genres, e.g., comedy or crime. The Bavarian translation aims at being as natural sounding in the dialect as possible.

We worked with the movie subtitles as included in the opensubtitles v2018 corpus (Lison and Tiedemann, 2016). The sentences vary greatly in quality, as some are missing (several) words and are clearly incomplete. InQA+ does keep some of these instances if the incomplete sentence makes sense as an answer by itself.

Data preprocessing. We pre-filter the opensubtitles v2018 corpus (Lison and Tiedemann, 2016) by searching the raw German corpus for questions that are followed by a line that does not contain the direct answer particles *Yes* and *No*. We then go over the potential candidates to find question answer pairs where the answer is indirect and plausible with regard to the question. To create the English corpus, we find the parallel English sentences of the chosen German instances. The Bavarian translation is based on the English corpus as to not

be influenced by the Standard German wording.

A.2. GenIQA Data Statement

- Dataset name and version: GenIQA (EN, DE and BAR) (version 0.1)
- Dataset curators: Miriam Winkler, Verena Blaschke, Barbara Plank
- Dataset and data statement citation: Please cite this paper.
- Data statement version: 1.0

Executive summary and curation rationale. This dataset is a training dataset for Indirect Question Answering (IQA) resource. It contains 1.500 LLM-generated indirect question-answer pairs per language in English, Standard German and Bavarian, a German dialect.

Date. The data was generated in the time period of March to July 2025 with GPT-4o-mini (OpenAI, 2024).

Modality. Written (LLM-generated question-answer pairs).

Genre. Questions and indirect answers.

Dataset statistics. The dataset contains of 1.500 English, Standard German and Bavarian sentences each. It consists only of a train split for each language.

Labels. Each question-answer pair carries a numerical label indicating the polarity of the answer, ranging from the gradients of *Yes* and *No* to middle and out-of-context labels. The meaning of the labels is explained in Table 2 and the annotation details can be found in §3.1. The labels were generated by GPT-4o-mini jointly with the question–answer pairs.

Language varieties. The dataset contains data in English (en-EN), Standard German (de-DE) and a dialectal variant of German, Bavarian (de-BAR). Each language was generated natively and there is no information about the regional variety of Bavarian that the model generated.

Linguistic situation and data quality. The language in the dataset is scripted, as the LLM generations do not depict natural text. The generations for English and Standard German are high-quality. However, as detailed in §4.1, the quality of the Bavarian dialect is lacking. A large part of instances was generated in either Standard German or unnatural pseudo-dialect.

A.3. Distribution, use, and maintenance

Distribution and usage terms. We release the datasets in the linked GitHub repository.[7] InQA+ can be used under the terms of opensubtitles v2018 of mentioning the corpus paper (Lison and Tiedemann, 2016) and the source website http://www.opensubtitles.org/.

Maintenance and updates. We do not plan further updates of this dataset. If you find errors, please open an issue on GitHub or contact Verena Blaschke or Barbara Plank.

[7] https://github.com/mainlp/Multilingual-IQA

A.4. Limitations, acknowledgements, and further information

Limitations. The data format and label set corresponds to the IndirectQA (Müller and Plank, 2024) dataset in order to ensure compatibility.

The Bavarian translation reflects the dialect of a single speaker.

Disclosures and acknowledgements. There are no conflicts of interest. This research is supported by European Research Council (ERC) Consolidator Grant DIALECT 101043235.

About this data statement. A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software.

This data statement was written based on the template for the Data Statements Version 3 Schema. The template was prepared by Angelina McMillan-Major and Emily M. Bender and can be found at http://techpolicylab.uw.edu/data-statements.

B. InQA+ Details

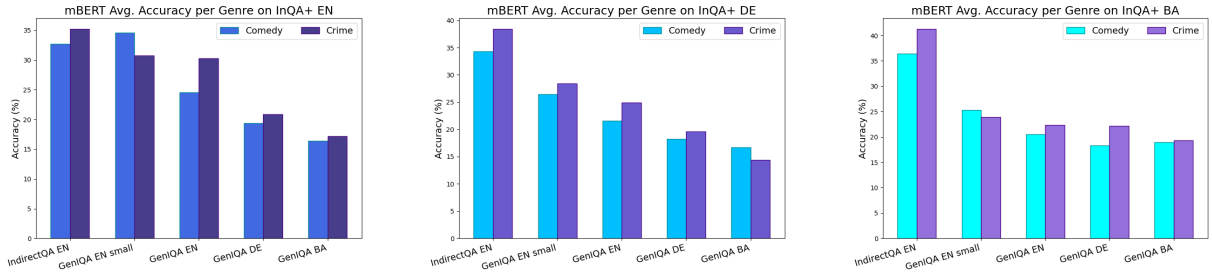
B.1. Data Genres

Müller and Plank (2024) divide their dataset IndirectQA into two separate genres: comedy and crime to research the effect of domain They found that comedy sentences, that contain complex pragmatic structures like sarcasm more frequently than crime yield lower accuracy scores.

We do not see distinctive differences in performance when dividing InQA+ into comedy and crime test sets and evaluating our trained models on the separate genres. However, the accuracy of the crime test set is generally slightly higher (Figure 2), fitting into the observation by Müller and Plank (2024). The same trend applies to InQA+ yesno (Figure 3), even if the disparities are even lower.

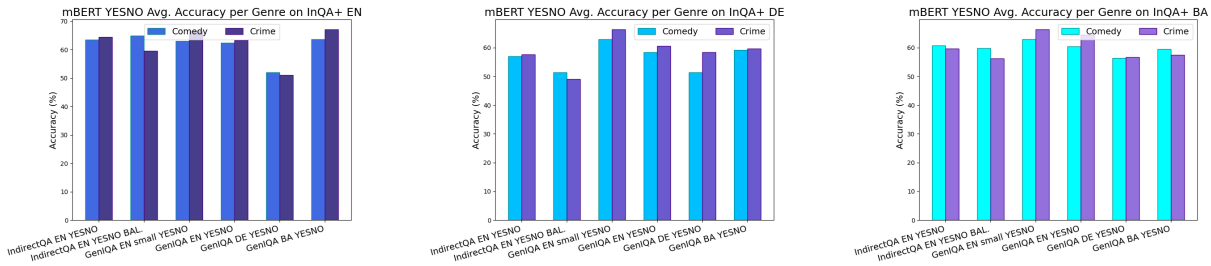
B.2. Annotation Details

In §3.1, we explain the annotation of InQA+. The label definitions are the main annotation guidelinewith an additional rule for negative questions



(a) English. (b) German. (c) Bavarian.

Figure 2: Average accuracy scores per genre over three seeds of mBERT models, evaluation on InQA+.



(a) English. (b) German. (c) Bavarian.

Figure 3: Average accuracy scores per genre over three seeds of mBERT models, evaluation on InQA+ YESNO.

that are a challenge for human and machine annotations. With the additional rule, negative questions can be disambiguated and annotated uniformly across larger datasets. However, while this method untangles the polarity for humans to understand, the uncertainty of the answers' meanings still is a key factor to the difficulty of the IQA task. We interpret the labels as following:

- ***Yes***: **A clear yes or all gradients of yes (including weaker forms, e.g., maybe yes).** Every instance that answers the question positively and confirms the statement. Clear *Yes* answers include instances like *I sure do* or *Of course*, while the gradients are more varied, for example *I guess so* or *Oh, well, maybe*.

- ***No***: **A clear no or all gradients of no.** Every instance that answers the question negatively and contradicts the statement. Clear *No* answers include sentences like *Oh, I can't* or *Course not*, while the gradients consist of instances like *Not yet, but I will* or *Probably not*.

- ***Conditional Yes***: **A yes that only holds if certain conditions are true.** Every instance that answers the question positively, but there is a direct or indirect condition. For example, *I will be if you want me to be* is a direct condition. Indirect conditions are inferred from the answer, like in *You going by Audrey? – I can.* – meaning yes, the responder can, if the questioner wants.

- ***Neither Yes nor No***: **A neutral answer that lies in the middle of yes and no.** Every instance that does answer the question, but the polarity is not clearly positive or negative. In the majority of cases the responder is thinking (e.g., *Let me think about that for a second*) or doesn't know the answer (*I don't know*).

- ***Other***: **The sentence does not match the questions as an answer.** Every instance that does not fit the question, because there is no clear connection without knowing further context, for example: *Another trick question, right? –Open the window*. We also label short phrases like *What?* or *Hm?* as *Other*, as the answers do not carry informative meaning.

- ***Lacking Context***: **Without further context, the answer cannot be clearly categorized.** Every instance where the polarity depends on the context around the question-answer pair, e.g., *Does he want us to go? – He was very quiet on the phone*. Without knowing the person referenced, it is unclear whether *he* wants the attendance or not. This label also includes sarcastic answers like *You would've taken the subway? – What do you think?*. For more information on the difference between *Neither Yes nor No*, *Other* and *Lacking Context*, please refer to §3.1.

Complex phenomena: negative questions In INQA+, 31 IQA pairs (7 % of the total instances) contain negative questions, which raises the need to standardize the annotation of such cases. The interpretation is context- and speaker dependent (Dušková, 1981).

As explained in §3.1, we resolve negative questions with polarity-based interpretations (Holmberg, 2012, 2013), as it is clearer to understand and annotate for various annotators. When calculating the IAA on a very small sample of double-annotated instances, we get an agreement of 0.35 Cohen's Kappa score before setting the additional guidelines, which is interpreted as no or only slight agreement. After streamlining the interpretation of negative questions as polarity-based, we raise the agreement to 0.54, falling in line with the agreement on the whole dataset.

The question *Are you not Moby?* has three different possible answers: two *Yes* answers – *(Yes) I am Moby* (polarity-based) or *(Yes) I am not Moby* (truth-based) – and one *No* answer – *(No) I am not Moby*. For German speakers, the *Yes* answers can be disambiguated with the the word *Doch*, an adverb that has no English equivalent, as a help for decision making. In our example case, it carries the meaning of *Yes, I am (indeed) Moby* and cannot be confused with *Yes, I am not Moby*.

Further, in the case of IQA, the answer additionally might not always clearly reveal the polarity like in the *Moby* example. Looking at the question-answer pair *Can't you hear the bell ringing? - That's Lilly's job*, the annotators additionally have to decode the indirectness in the answer before annotating. The answer implies that the bell was heard, but that the responding person does not feel responsible to take any action. With this inferred knowledge, the more direct answer can be decoded as *(Yes) I can hear the bell ringing, but that is Lilly's job*, leading to a *Yes* annotation.

C. LLM Generation Tests

The selection of the right model is crucial for the generation quality which is why we took great care in testing the models. As explained in §4, we focus our testing on the generation of Bavarian, a German dialect.

C.1. LLM dialect generation testing

The generation of authentic, natural sounding Bavarian is not trivial. We first test the models' dialect generation capabilities with the prompt *Write a sentence in Bavarian dialect* before presenting the models with the complex IQA task (**Dialect** criterion in Table 7).

The only free model capable of producing natural sounding, comprehensible Bavarian is EuroLLM-9B (Martins et al., 2025). It returns sentences like *I bin do in Land, wo's Bier so g'scheit is* and also provides the unprompted translation *I am in the land where beer is so good*. However, its run time of a minimum of 5+ hours on the computational resources we had available for this study was too long to be feasible for producing a whole dataset. Other than EuroLLM-9B, GPT-4o-mini (OpenAI, 2024) as a paid API also generated natural sounding dialect. The other LLMs generated mostly "gibberish" sentences.

We continue with zero- and few-shot prompts tailored to the IQA task and our label set with IQA examples (for few-shot). From the generation results, we analysed if the sentences were coherent and complete (**Correct** criterion), if the answer made sense with regards to the question (**Logical** criterion) and if the answers were indirect (**Indirect** criterion) in Table 7. In general, the best results were acquired within the few-shot scenario. The results of the zero-shot prompting were often arbitrary and did not follow the given instructions.

The highest quality results were generated by GPT-4o-mini. It follows the prompt in detail and generates coherent dialect. Even though OpenAI is not open about their training data sources, we can see from the generation performance that GPT-4o-mini has seen Bavarian text in the supposed assortment of openly available corpora, scraped internet text and reinforced post-learning through human feedback (OpenAI, 2024).

Even though we only see unfinished generations with EuroLLM-9B due to interruptions, the experiments with a few-shot IQA prompt show the good dialect quality: *Hast denn da Frida schon am Bus-Halt gwunn* (translation with assumptive completion in []: *Have you [waved] to Frida at the bus stop yet?*). The article *da* in front of the proper name and the omitted *e* in between *g* and *w* in *gwunn* – in contrast to Standard German prefix *ge-* – is typical for the Bavarian dialect. While EuroLLM-9B was very promising, the smaller version with faster run times, EuroLLM-1.7B (Martins et al., 2025), did neither produce dialect nor follow the prompt instructions.

OLMo-2 (OLMo et al., 2024) had a similarly low performance: the generated answers did to the most parts contain no dialect and instead of following the prompt, it either translates or just returns parts of the requested question-answer-label format, e.g., the label. Gemma-2 (Gemma Team, 2024) produced pseudo-Bavarian. The sentence structure and question-answer logic is correct, but the dialect does not make any sense. Llama's (Meta, 2024) production is worse than Gemma's as it only produced gibberish with the same sentence structure of *Do you have [...] - I have [...]*.

Unfortunately, we could not test generating with

Model Name	Instruction-tuned?	Dialect?	Correct?	Logical?	Indirect?
EuroLLM-9B-Instruct	✅	✅	✅	✅	✅
EuroLLM-1.7B-Instruct	✅	❌	❌	❌	❌
OLMo-2-1124-7B-Instruct	✅	⚠️	✅	⚠️	✅
gemma-2-2b-it	✅	⚠️	✅	✅	⚠️
Llama-3.2-1B-Instruct	✅	❌	❌	⚠️	⚠️
gpt-4o-mini	✅	✅	✅	✅	✅
LLäMmlein_1B	❌	/	/	/	/
Betzerl_1B_wiki_preview	❌	/	/	/	/

Table 7: Assessment matrix of relevant features for IQA and Bavarian for various LLMs that are true (✅), partially true with some restrictions (⚠️) or false (❌).

the German LLM LLäMmlein_1B (Pfister et al., 2024) and the Bavarian adapter Betzerl (CAIDAS Uni Würzburg, 2024) as they are not instruction-tuned and thus not suitable for our purpose.

C.2. Prompt wording testing

For the generation of the full GenIQA datasets, we fine-tuned the prompt to better capture the beneficial features provided by OpenAI's best practices (OpenAI, 2024). The following prompt is the final one we used for the generation process:

> *Write 50 polar questions and sentences that answer the questions. Both should be in English and formatted as the given format. The answer should be indirect, meaning that it does not contain direct answer particles like Yes, No or similar indicators. The answer should be labelled with one of the given labels from the label set and therefore be suitable for the explanation of the label. Focus on the labels 1 to 4, but also include examples for the labels 5 and 6. Do not number the examples, just write them in the given format.*
>
> *Desired format: question answer label*
>
> *Label set:*
> *1 = Yes: A clear yes or all gradients of yes (meaning weaker forms of yes like a maybe yes).*
> *2 = No: A clear no or all gradients of no.*
> *3 = Conditional Yes: A yes that only holds if certain conditions are true.*
> *4 = Neither yes nor no: A neutral answer that lies in the middle of yes and no.*
> *5 = Other : The sentence does not match the question as an answer.*
> *6 = Lacking context: Without further context, the answer cannot be categorized as yes or no.*
>
> *Example for each label:*
> *This is Long Tieng, right? Right. 1*
> *You hungry? Just black coffee for me. 2*
> *Will you be home tonight? I will be if you want me to be. 3*
> *Is anyone still in there? I don't know. 4*
> *Another trick question, right? Open the window. 5*
> *- Hey, you guys want sangria? - It's Guys' Night! 6*

The prompt contains more fine-grained instructions than our testing prompts and explains the meaning of the label sets clearer. As we saw the best results with few-shot prompting, we added more examples (one for each label) from InQA+. Additionally, we tell the model to focus on labels *Yes*, *No*, *Conditional Yes* and *Neither Yes nor No*, all of which bear the most interesting information for the task. Without this addition, the model had the tendency to label most generations as *Other*.

Prompt language testing Since we experiment with the data creation for EN, DE and BAR, we also try out in-language prompts, meaning a German and a Bavarian prompt for the creation of Bavarian data. With the idea that the in-language prompts "prime" the model for a closer language to the one we want to generate, we translate our IQA generation prompt above accordingly. For the German split of GenIQA, the prompt language did not make a big difference: the labelling accuracy of 100 sentences generated with the German prompt is 44 % – comparable to the 45 % reached with the English prompt. The Bavarian prompt did not work well and yielded poor generations.

C.3. Generation temperature testing

After fine-tuning the prompt, we experiment with the generation temperature. It is an important parameter for the quality of generations as a high temperature makes the generation more random and diverse and a lower temperature makes them more deterministic with the chance of it being more repetitive (OpenAI, 2024). We want creative and varied sentences for the GenIQA datasets but keep the labelling accuracy of the generated annotations as high as possible. Thus, we compare 100 generations with the temperatures 0.2 and 0.8 respectively and manually calculate the labelling accuracy by hand-annotating the sentences. For the low temperature, we find an agreement of 48 %; with the high temperature, it lies at 59 %. Higher temperature thus gives more create freedom and higher

labelling accuracy to the model, which is why we choose this value.

D. Generated Dialect Quality

We provide more details on the dialect quality of the LLM-generated Bavarian data (GenIQA BAR) which we discussed in §4.1. We first present a more in-depth look into our manual quality assessment and provide the details of our native speaker survey afterwards.

D.1. Manual dialect quality assessment of GenIQA BAR

We want to analyse the quality of GenIQA BAR as generated by GPT-4o-mini (OpenAI, 2024) and differences between artificial and natural data for low-resource languages in more detail. The model does not use Bavarian grammar, dialect vocabulary (words specific in the dialect that does not exist/is not used in the standard language) or dialect spellings that match the pronunciation consistently.

We noticed two prominent positive features. Firstly, in dialect speech, clitics are characteristic and common (Weiß, 1998; Blaschke et al., 2024a), e.g., merging the word *es* (it) into the preceding words. The model applies this for example in the sentences *I find die Berge schön, aber am Meer* ***is's*** *auch nett* (I think the mountains are beautiful, but it's also nice by the sea). Secondly, the best sounding artificial dialect sentences contain words with elisions (e.g., missing vowels, especially prevalent in pre- and suffixes), like *g'wesn* instead of *gewesen* in *I bin da schon oft* ***g'wesn*** (I have been there a lot).

However, the most generations contain erroneous or pseudo dialect (incorrectly imitated dialect characteristics). We classify three prominent error classes: wrong usage of verbs, incorrect omissions and determiners of false genus. Wrong usage of verbs refers to verbs that do not fit into the grammatical context. For example, the answer *I* ***hätt*** *gern, aber es gibt viel zu tun* (I would like to, but there is a lot do do) uses *hätt* (subjunctive II form of the verb *haben* (to have)) that does not suit its question *Kommst du mit uns zum Grillen?* (Are you coming with us to the barbecue?). Subjunctive II of *werden* (would) would be correct instead.

Then, GPT-4o-mini replicates clitics and elisions, but produces incorrect letter omissions. In the sentence *I hab a paar Blüm' und Kräuter, aber ned viel Platz* (I have a few flowers and herbs, but not much space), the word **Blüm'** (flower) does not exist in this form in the dialect of the annotator and is wrongly cut short.

False genus generations are also common in GenIQA BAR. For example, *I ess sie, wenn's an Fest gibt* (I eat them when there's a party) uses *an*, the accusative masculine determiner. As *Fest* is grammatically neuter, the correct dialect determiner is *a*. This behaviour is not restricted to dialect sentences and also sometimes occurs with Standard German.

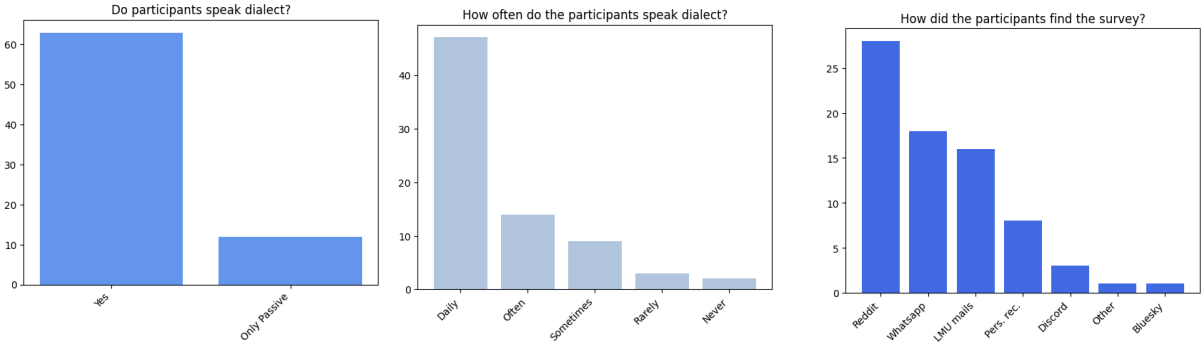
Additionally, we observed in the sample of 100 instances we analysed in §4.1 that the kind of vocabulary GPT-4o-mini generates in Bavarian differs from Standard German. Many GenIQA BAR sentences contain stereotypical Bavarian words that each have zero occurrences in GenIQA DE but multiple in the whole Bavarian data, for example *Bier* (beer), *Wirtshaus* (traditional Bavarian restaurant) and traditional Bavarian food like *Leberkäse* (meat loaf) and *Brezen* (pretzels) are a selection of the words we found. Since OpenAI does not disclose their training data in detail, we assume that the dialect data GPT-4o-mini encountered strongly covered Bavarian traditions to build the connection of this vocabulary and Bavarian.

D.2. Native speaker survey

Bavarian dialect is a broad and varied spectrum with many regional varieties. We conduct an anonymous survey and ask multiple native Bavarian speakers from different regions in Bavaria to rate the quality of the AI-generated dialect, as the perception of the dialect may vary depending on the participants' own spoken dialect variety. For this, we ask the participants to disclose the administrative region (Upper Bavaria, Lower Bavaria, Upper Palatinate, Upper Franconia, Middle Franconia, Lower Franconia or Swabia) they live in as the only personal information. The online survey was conducted with SoSci Survey (Leiner, 2025) and the data was collected between 29.05.2025 and 11.06.2025. In total, 76 active dialect speakers of varying degrees (Figure 4a) answered the survey, with the majority of them stemming from Upper and Lower Bavarian (Figure 5) and speaking dialect on a daily basis (Figure 4b). Out of the seven distribution channels, the most participants found the survey on Reddit (Figure 4c).

The design of the survey is independent from the IQA task to assess only the linguistic quality of the generated dialect and to allow participants without linguistic knowledge to partake. The participants' task is to judge ten question-answer pairs on if the answer contains dialect and if it is authentic (meaning that they would use such an answer in their daily life). We select the examples in a way that each represents a different generation phenomena and degree of quality (low to high):

1. Kann ich bei dir lernen? - So lange du ned sabbelst, klar.
   Can I learn at your place? - As long as you



(a) Information about *if* participants speak dialect.

(b) Information about *how often* participants speak dialect.

(c) Distribution channels where the participants found the survey.

Figure 4: Personal disclosures of the participants in the dialect quality survey.

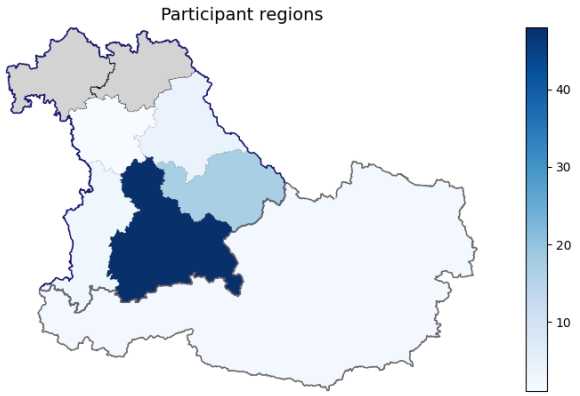


Figure 5: Participant origin regions of the dialect quality survey: Bavaria (per administrative region) and Austria.

don't blabber, sure.
The word *sabbeln*[8] (blabber) stems from Northern German dialects.
Quality: Low

2. Hast du den Film schon gesehen? - I hob nur die Werbung gseh.
Have you seen the film yet? - I've only seen the ad.
Contains pseudo-dialect (*gseh*, eng. seen).
Quality: Low

3. Ist der Chef im Büro? - Möglicherweise is er unterwegs.
Is the boss in the office? - He may be on the move.
Only contains one dialect word (*is*, eng. is).
Quality: Medium

4. Wirst du morgen aufstehen? - I schau ma amoi, wies mir geht.
Will you get up tomorrow? - I'll see how I'm doing.
Wrong Bavarian grammar (*ma* (Bavarian version of *we*) exposes the sentence as fake dialect as it is incorrect here).
Quality: Low

5. Ist die Pizza fertig? - Es riecht schon ganz lecker!
Is the pizza ready? - It already smells delicious!
Standard German.
Quality: Low

6. Habt ihr einen Tisch reserviert? - I hab's vergessen.
Did you book a table? - I forgot.
Only contains one dialectal word, but sounds authentic nonetheless.
Quality: High

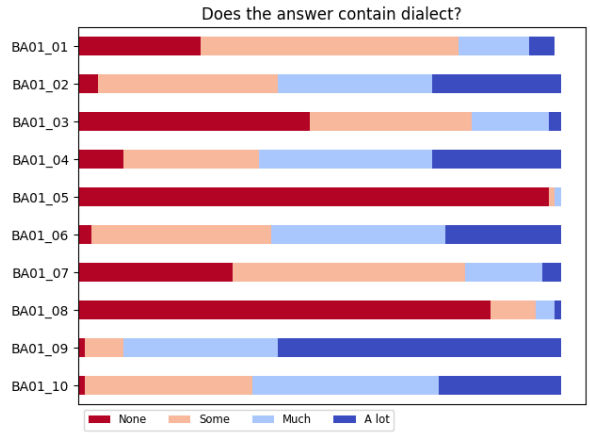
7. Gehst du oft ins Kino? - Ich schau lieber Filme doheim.
Do you often go to the cinema? - I prefer to watch films at home.
Wrong dialect spelling (*doheim*, std. ger. daheim, eng. at home).
Quality: Low

8. Ist das Bier kalt? - Es steht's im Kühlschrank.
Is the beer cold? - It's in the fridge.
No expression of the intended meaning. The answer is Bavarian, but only with the interpretation of *You (plural) are in the fridge*. For the question, it does not hold the correct meaning.
Quality: Low

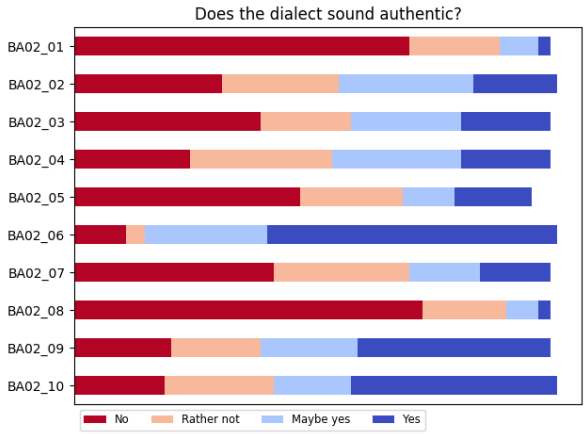
9. Hast du schonmal Ente gekocht? - I kimm mid de Entn klar. Entspann di.
Have you ever cooked duck? - I can handle the ducks. Relax.
Manually translated answer by the annotator
Quality: High

10. Isst du etwas Gesundes? - I hätt grad a Lust auf a Stück Pizza.

[8] https://www.dwds.de/wb/sabbeln

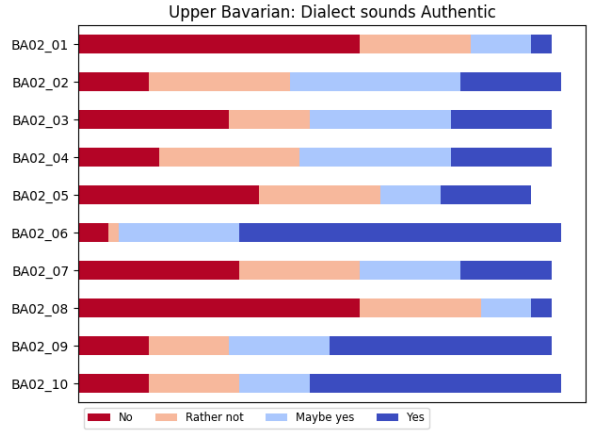


(a) Dialect presence.

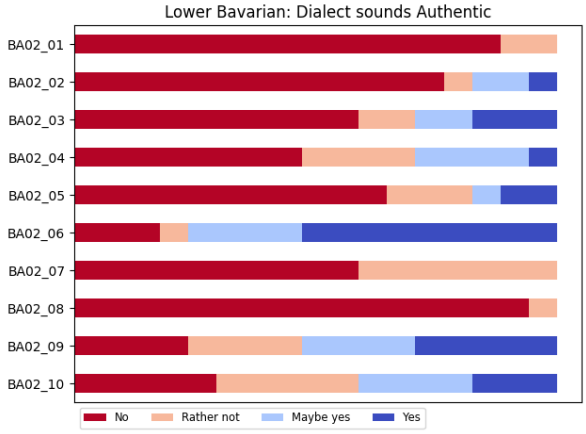


(b) Dialect authenticity.

Figure 6: Results of the dialect questions.



(a) Authenticity rating in Upper Bavaria (author's region).



(b) Authenticity rating in Lower Bavaria.

Figure 7: Results of the dialect questions per region.

Are you eating something healthy? - I crave a slice of pizza right now.
Uncommon Bavarian grammar (using an article with *Lust* (craving) is not wrong per se, but unusual).
Quality: Medium

As the ratings in Figure 6a reveal, the participants recognized the Standard German and natively translated answer from INQA+ BAR. For answers 1, 3 and 7, the participants' opinions were rather negative. Answer 8's rating is very negative. Only 5 out of 10 answers (2, 4, 6, 9 and 10) were rated to contain dialect. As we already touched upon in Section 4.1, the ratings from dialect speakers stemming from Upper Bavaria – the author's region – are higher than from Lower Bavaria (cf. Figure 7). We distinguish only Upper and Lower Bavaria due to not having enough participants from other regions to achieve representative and comparable results.

From this selection, we see that the answers received uniformly mixed ratings for the authenticity of the dialect, except for answer 6, as visualised in Figure 6b. As we expected, this answer is mainly rated as authentic, meaning that the participants would use this sentence in their daily lives.

E. Hyperparameter Fine-tuning

To get the best possible performance for each model and training dataset, we train individual hyperparameter sets of learning rate, batch size warm-up ratio, weight decay, dropout rate and label smoothing for each setup (cf. Table 8 and 9) as mentioned in §5. For this, we train the parameters in an order of importance to and influence on the model. This prioritization is necessary due to our available computational resources that do not allow testing all combinations. Thus, we divided the parameters into two fine-tuning groups: optimisation (learning rate, batch size and warm-up ratio) and regularisation (weight decay, dropout rate and label

Model Train Data	LR	Batch Size	Warm-up Ratio	Weight Decay	Dropout Rate	Label Smoothing
mBERT						
IndirectQA EN	1e-5	16	0.1	0.1	0.1	0.0
GENIQA EN	1e-5	16	0.5	0.01	0.5	0.1
GENIQA EN SMALL	1e-5	8	0.3	0.01	0.5	0.1
GENIQA DE	1e-5	8	0.1	0.001	0.5	0.3
GENIQA BA	1e-5	4	0.1	0.01	0.5	0.1
IndirectQA EN YESNO	1e-5	16	0.1	0.1	0.1	0.0
GENIQA EN YESNO	1e-5	4	0.1	0.001	0.3	0.1
GENIQA DE YESNO	1e-4	16	0.3	0.1	0.1	0.0
GENIQA BAR YESNO	1e-5	16	0.3	0.1	0.1	0.0
mDeBERTa						
IndirectQA EN	1e-4	32	0.1	0.1	0.1	0.0
GENIQA EN	1e-4	32	0.1	0.01	0.1	0.3
GENIQA DE	1e-5	32	0.1	0.001	0.1	0.1
GENIQA BAR	1e-4	4	0.5	0.1	0.1	0.0
XLM-R						
IndirectQA EN	1e-5	32	0.5	0.1	0.5	0.3
GENIQA EN	1e-5	32	0.3	0.1	0.1	0.0
GENIQA DE	1e-5	8	0.1	0.01	0.1	0.1
GENIQA BA	1e-5	16	0.3	0.1	0.1	0.0

Table 8: Grid-searched parameters per model and training dataset after fine-tuning three seeds each for extensive experiments.

Model Train Data	Trials	LR	Batch Size	Warm-up Ratio	Weight Decay	Dropout Rate	Label Smoothing
mBERT							
IndirectQA EN REMAPPED	30	4.4274e-4	32	0.1587	0.0015	0.4748	0.1929
IndirectQA EN + Friends-QIA REMAPPED	30	1.6414e-6	4	0.1747	0.0524	0.3930	0.0753
Friends-QIA	30	2.1156e-6	8	0.4961	0.0065	0.3877	0.0931
IndirectQA EN + Friends-QIA	30	4.2071e-6	32	0.1554	0.0003	0.1300	0.0706

Table 9: Random-searched parameters over 30 trials for mBERT per training dataset for exploratory experiments. Rounded to four digits.

smoothing) parameters. We extract the best configuration of optimisation parameters depending on the accuracy score and use them to fine-tune the regularisation parameters to maximise the generalisation capabilities of the model in training. The final choice of parameters was again made depending on the best average accuracy score. Through the regularisation fine-tuning, we boost the accuracy on the dev data by an average of 0.5 % as opposed to only training the optimisation parameters.

Grid-searching is heavy on our computational resources. Finetuning mBERT (Devlin et al., 2019) takes ∼5 hours for small datasets like IndirectQA and ∼10 hours for large datasets like GENIQA, while fine-tuning mDeBERTa (He et al., 2020) and XLM-R (Conneau et al., 2020) take ∼18 hours for IndirectQA and ∼24 hours for GENIQA. All training runs and experiments were conducted with a Windows-based system with an 8GB VRAM Nvidia GeForce RTX 3070 GPU with and a Linux-based system with an Nvidia GeForce RTX 2060 SUPER with 8GB VRAM.

Random hyperparameter search saves resources due to more efficient exploration of the hyperparameter space (Ulmer et al., 2022), which only takes ∼0.5 hours with IndirectQA and 30 trials, but the resulting parameters yielded slightly worse results with a deficit of -3 pps. for accuracy. We use random search for our experimental variants due to the training efficiency.

Each model has its own set of fine-tuned hyperparameters for the best performance because the parameters do not generalise well to other models with different architectures and parameter sizes. We observed this in pre-experiments where we trained models on IndirectQA EN with parameters

Dataset Training Setup	Train Accuracy	Dev Accuracy	Gap	Majority Class Baseline
mBERT trained with				
IndirectQA EN	53.71 ±5.71	43.90 ±4.53	9.81	35.61
GENIQA EN	60.78 ±2.85	58.56 ±4.60	2.22	31.80
GENIQA DE	56.82 ±8.09	51.22 ±2.52	5.60	30.93
GENIQA BAR	61.31 ±1.00	52.78 ±3.24	8.53	31.20
IndirectQA EN REMAPPED	53.50 ±4.40	51.76 ±2.86	1.73	37.56
IndirectQA EN + Friends-QIA REMAPPED	56.65 ±1.40	56.03 ±1.49	0.62	47.54
IndirectQA EN YESNO	82.27 ±7.84	74.44 ±2.55	7.83	73.24
GENIQA EN YESNO	88.18 ±4.84	84.89 ±0.84	3.29	67.73
GENIQA DE YESNO	82.79 ±3.70	78.96 ±1.50	3.82	59.76
GENIQA BAR YESNO	87.43 ±2.68	82.72 ±3.99	4.71	74.22
Friends-QIA	60.06 ±1.04	56.66 ±1.22	3.40	46.61
IndirectQA EN + Friends-QIA	61.89 ±1.61	55.67 ±1.70	6.22	45.40
GENIQA EN SMALL	63.31 ±8.67	57.72 ±5.69	5.58	35.45
mDeBERTa trained with				
IndirectQA EN	48.08 ±5.77	40.11 ±2.86	7.97	35.61
GENIQA EN SMALL	68.13 ±8.73	58.81 ±4.62	9.32	35.45
GENIQA EN	65.11 ±3.68	61.67 ±3.61	3.44	31.80
GENIQA DE	66.64 ±1.08	56.56 ±2.34	10.09	30.93
GENIQA BAR	65.58 ±2.72	55.33 ±1.00	10.24	31.20
XLM-R trained with				
IndirectQA EN	49.97 ±12.80	45.53 ±6.35	4.44	35.61
GENIQA EN SMALL	72.30 ±5.84	59.89 ±4.01	12.41	35.45
GENIQA EN	70.31 ±0.92	62.44 ±3.34	7.87	31.80
GENIQA DE	59.22 ±5.16	53.67 ±2.65	5.56	30.93
GENIQA BAR	67.16 ±5.95	56.11 ±4.11	11.04	31.20

Table 10: Average accuracy scores over three seeds on development and train datasets. The gap denotes the generalisation difference from the train accuracy to the dev accuracy.

fine-tuned on GENIQA BAR which resulted in lower accuracy scores for every evaluation dataset.

F. Training and Development Dataset Results

We present the results of our setups evaluated on their respective train and development datasets mentioned in §6.1 in Table 10. The comparison of the accuracy scores and the gap between them on the train and dev sets reveals the overfitting. Usually, this generalisation error should be small (Zhang et al., 2016). However, for our trained models, we see that there is a large gap between each score-pair (cf. Table 10 with an average accuracy drop of -6 pps. from train to dev on mBERT (Devlin et al., 2019) models and -8 pps. on mDeBERTa (He et al., 2020) and XLM-R (Conneau et al., 2020)). The larger the gap between train and development results, the more overfitting took place.

All train and dev scores lie above the majority class baseline, but are low overall, even for English. We carefully inspected the training logs and confirm that the training loss goes down steadily. Due to overfitting, the results are not as expressive as we expected and need great care when analysing. High accuracy scores can be deceiving, as they often stem from an overprediction of the majority class and resemble the majority class baseline distribution. Further, high F1 scores are also not completely reliable. They can indicate that the model has learned good generalisation despite the low accuracy, but in some cases, we observed that the high F1 scores are caused by model biases for minority classes, which boosts the macro-F1 through high recall scores. This is why we necessarily inspect both scores and check the predicted label distributions to inspect the specific model behaviour.

High deviation scores for both accuracy and F1 scores (e.g., Tables 5, 6 and 10), make the training and evaluation across various seeds a necessity. Otherwise, the performance would be too unreliable for analyses, as the scores vary greatly.