LLM-based Classification of Grounding Acts in German

Milena Belosevic* **and Hendrik Buschmeier†**
* German Linguistics, Faculty of Linguistics and Literary Studies,
Bielefeld University, Bielefeld, Germany
†Digital Linguistics Lab, Faculty of Linguistics and Literary Studies,
Bielefeld University, Bielefeld, Germany

Abstract

While the capabilities of Large Language Models (LLMs) to identify and classify grounding acts, such as clarification requests or acknowledgments, have been recently tested for English, there is still little research in this area for other languages. This paper investigates whether LLMs can reliably classify grounding acts in German by creating two balanced datasets of *advancing* and *non-advancing* grounding acts from institutional counseling and classroom conversations. We first apply five-shot instruction tuning with QLoRA to four models trained either only on German (LLäMlein and BübLeLM) or on multilingual data (Teuken-7B and EuroLLM). Since this strategy fails to generalize reliably, we fine-tune the same models using a classifier head, again with QLoRA, which yields substantially better results. All models are trained on the counseling dataset and evaluated both in-domain and on the unseen data from the classroom domain. We compare the classifier fine-tuned LLMs against two BERT-based baselines (GBERT-large and Google RemBERT). Results show that the classifier-head BübLeLM outperforms the best-performing baseline (GBERT-large) in in-domain settings. At the same time, GBERT-large achieves the best across-domain performance, making it the most robust overall model for grounding act classification in German.

1 Introduction

Understanding the communicative intent behind utterances – commonly referred to as Dialogue Act (DA) classification – is a core task in Natural Language Understanding and crucial for building responsive conversational agents (Ahmadvand et al., 2019). While extensively studied in human interactions, DA classification has been less explored in Large Language Model (LLM)-powered systems, especially for languages other than English. One particularly important type of dialogue acts are grounding acts – utterances through which interlocutors manage mutual understanding by providing, requesting, or acknowledging evidence of understanding (Traum and Allen, 1992). Grounding is a collaborative process involving alignment and coordination of cognitive states between participants (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989). A special type of grounding focused on in this paper is conversational grounding that involves a common understanding of shared knowledge within a conversation (Traum and Allen, 1992; Traum and Hinkelman, 1992). This paper investigates how well German-only LLMs (i.e., those trained only on German data) and multilingual LLMs can classify grounding acts in German, distinguishing between advancing (e.g., acknowledgments, follow-ups) and non-advancing acts (e.g., clarification requests) (Shaikh et al., 2025). We fine-tune the LLMs using instruction tuning and classifier-head approaches across two domains. The models' in-domain and out-of-domain performance is evaluated using macro-F1, accuracy, and confusion matrices. GBERT and Google RemBERT (both BERT-based models) serve as baselines for the classifier-head fine-tuning experiments. Our results show that instruction tuning is ineffective for this task, while classifier-head fine-tuning leads to better performance. LLMs can outperform BERT models in-domain, but GBERT remains the most robust for out-of-domain classification. The paper is structured as follows: After reviewing relevant previous work in Section 2, Section 3 describes dataset construction and splitting as well as human annotation. We outline the methodology, including details about model selection and evaluation, in Section 4. The results of the instruction-tuning experiments are presented in Section 5, while Section 6 describes the results of the classifier-head fine-tuning approach. We discuss and summarize our findings in Section 7.

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 112–127, 2025

2 Related work

This section discusses the most recent work on applying LLMs for text classification (see Chae and Davidson 2025 for an overview), specifically the classification of grounding acts in languages other than English.

Prior work has primarily focused on whether LLMs can generate and classify grounding acts in English (Shaikh et al., 2024, 2025; Schneider et al., 2024). Regarding the generation of grounding acts, it has been demonstrated that LLM-powered conversational agents mainly fail to generate appropriate grounding acts (e.g. Shaikh et al. 2024). Prompting (Kuhn et al., 2023; Chen et al., 2023), fine-tuning (Andukuri et al., 2024), or a combination of both methods (Mohapatra et al., 2024b) are used to elicit grounding acts from models. For example, Tack and Piech (2022) and Hicke et al. (2023) simulate model responses in educational settings via fine-tuning to assess the model's grounding behavior. Similarly, Tack and Piech (2022) explore how well state-of-the-art conversational agents, such as Blender and GPT-3, reply to students in educational dialogues. Shaikh et al. (2024) use data from the educational domain, among others, to test whether LLMs can produce grounding acts in a human-like manner.

While these studies are concerned with LLM-based generation of grounding acts in English, we test their capabilities to classify German grounding acts accurately. LLM-based classification of grounding acts in English is based on prompting (Schneider et al., 2024) or on a small amount of human-annotated grounding acts (Jokinen et al., 2024). These grounding acts often serve as a validation set for testing different prompting strategies. Then LLMs are prompted to annotate new data (test set) using the final prompt, which was formed based on a validation set (Shaikh et al., 2024, 2025). In contrast, we compare the performance of fine-tuned LLMs with human annotations of German grounding acts. Additionally, work on LLM-based classifications of grounding acts is based on existing datasets in English (Mohapatra et al., 2024a). We compiled two new datasets of naturally occurring conversations in German from two domains.

As for the effectiveness of LLMs in text classification tasks, Nowacki et al. (2025) propose to modify decoder-style LLMs by adding appropriate layers to improve the classification parameters and build LLM classifiers for discovering mental disorders from texts. The enhanced models are evaluated against encoder-based BERT-based classifiers and standard zero-shot or few-shot LLM approaches. The results demonstrate that the LLMs fine-tuned with a classifier head outperform these baselines, achieving higher F1 scores and providing more precise classifications. In contrast, Fatemi et al. (2025) provide evidence for the effectiveness of instruction fine-tuning and model merging for adapting LLMs to financial text classification tasks in English.

Jadhav et al. (2025) evaluate fine-tuned BERT models and LLMs as annotators for classification tasks in a low-resource language, Marathi, including sentiment analysis, news classification, and hate speech detection. The results revealed that many LLMs still fall significantly short of the baseline performance achieved by BERT-based models and are not yet capable of replacing human annotators.

Plakidis and Rehm (2022) present a German dataset of 600 tweets annotated for speech acts, derived from the GermEval 2019 corpus on offensive language. The authors apply a multi-level annotation scheme – coarse- and fine-grained speech act categories and sentence types – to analyze pragmatic patterns in offensive versus non-offensive tweets. This work demonstrates the potential of pragmatic annotation for improving classification tasks for German data.

More recently, Leitner and Rehm (2025) explore the performance of various LLMs on German text classification tasks across five small and medium-sized datasets, focusing on both prompting and fine-tuning via QLoRA. The study evaluates multilingual and German models, some of which are also used in the present paper on classification tasks with imbalanced data, ranging from speech act labeling to offensive language detection. Results are not reported, but one of the goals is to investigate why experiments failed and how to improve the model's performance.

3 Dataset construction and annotation

3.1 Domains and sources

We focus on two domains, classroom conversations and institutional counseling, to test LLM capabilities to classify grounding acts. Both domains were selected based on the findings from previous studies of LLM capabilities to identify and generate grounding acts in English in educational and counseling domains (Shaikh et al., 2024). Since, to our knowledge, datasets comprising naturally occurring conversations in German with annotated ground-

ing acts do not exist, we compiled new training and evaluation datasets based on publicly available conversation transcripts from these domains[1]. For institutional counseling, we use several datasets listed in Table 3. This table also details the number of conversations and annotated grounding acts from each source. Most conversations and the majority of annotated grounding acts (38.5%) come from the conversations in the German immigration office collected by Herzberger (2013). They are followed by sales conversations from the database of spoken German, the largest database of the German spoken language, specifically from the FOLK subcorpus (The Research and Teaching Corpus of Spoken German), which covers a diverse range of verbal communication in private, institutional, and public settings (Schmidt, 2014). We used conversations from the institutional service encounters (e.g., service encounters in a department store, counseling sessions on student financial aid, and coaching sessions).

In the classroom domain, the majority of transcripts and annotated grounding acts (60%) were extracted from the DIALLS (Dialogue and Argumentation for Literacy Learning in Schools) project dataset (Rapanta et al., 2021), followed by the oral exams at university and classroom conversations from the FOLK database (see Table 4 for details).

For fine-tuning, the dataset was split into 960 training examples (80%), 120 validation examples (10%), and 120 test examples (10%), resulting in a total of 1200 labeled instances. The in-domain data was balanced, with 600 advancing and 600 non-advancing utterances. Additionally, a separate equally balanced out-of-domain evaluation set consisting of 780 utterances (390 advancing and 390 non-advancing) was used to assess generalization to unseen domains (see Table 5 in the Appendix).

To better characterize the grounding phenomena captured in our dataset, we provide several examples from both domains. Each utterance was annotated as either ADVANCING or NON-ADVANCING.

- **Advancing**:
  - Counseling: *Haben wir das Verfahren ruhen lassen? 'Did we stay the proceedings?'*
  - Classroom: *Okay, Gruppenarbeit beendet, wer möchte anfangen? 'Okay, the group work is finished. Who wants to begin?'*
- **Non-Advancing**:
  - Counseling: *Was behandelt belastet Sie am meisten jetzt? 'What is bothering troubling you the most right now?'*
  - Classroom: *Ich wollte sowieso letzte woch nächste woche noch mal reinkommen, wegen eines thesenpapiers. 'I was planning to come by again last week next week anyway, because of a thesis paper.'*

We observed that advancing utterances tended to be longer and more syntactically complex, while non-advancing acts often consisted of shorter or elliptical utterances.

3.2 Annotation procedure

Two linguists manually annotated the counseling and classroom datasets based on the full conversation transcripts, following the typology of grounding acts proposed in prior work on conversational grounding (Clark and Schaefer, 1989; Traum and Allen, 1992; Traum and Hinkelman, 1992). Specifically, we modify the typology of grounding acts used in Shaikh et al. (2025). Shaikh et al. (2025) distinguish between three types of grounding acts: advancing, addressing, and ambiguous. Advancing acts contribute to the progression of the conversation by confirming understanding or elaborating on the prior utterance through acknowledgments (e.g., "Got it"), follow-ups (e.g., "Why do you think that?"), direct answers to questions, opinion expression, or apologizing. In contrast, addressing acts reflect grounding problems expressed by self-repairs, restarts, or reformulations. Ambiguous acts, such as clarification requests, are used to prevent potential misunderstandings. In this paper, we consider both addressing and ambiguous acts as non-advancing grounding acts, since they do not actively move the dialogue forward but instead signal or stem from a breakdown in mutual understanding.

The total number of advancing cases was much higher than the number of annotated non-advancing grounding acts. To generate balanced datasets, the total number of advancing utterances used for training and evaluation was adjusted to the total number of non-advancing grounding acts (600 in the in-domain dataset/counseling and 390 in the out-of-domain/classroom dataset). In cases of disagreement that usually apply to questions such as *Sie*

[1]All datasets, annotation guidelines, and evaluation scripts are available at: https://github.com/milenabelosevic/ClassificationGA.

haben bereits einen deutschen Pass oder? ('You already have a German passport, right?'), a third annotator was consulted, and the final label was determined by majority vote (2 out of 3). The annotation was performed on the sentence level. Short utterances like *yes*, *ok*, *no*, and non-verbal signals (e.g., *mhm*) were excluded *only* when they occurred as isolated turns without surrounding verbal content. If embedded in a meaningful sequence or co-occurring with other verbal acts, such utterances were annotated as usual. Only in some cases more than one sentence was considered (e.g., other-repair: *I was looking for my pass. You mean your child's pass?*). The overall inter-annotator agreement (the IAA for both datasets) was almost perfect (Cohen's $\kappa = 0.844$, $p < 0.001$), with a raw agreement rate of 92.17%. Cohen's κ for the counseling dataset was 0.817 ($p < 0.001$, agreement rate: 90.83%) and for the classroom dataset: 0.885 ($p < 0.001$, agreement rate: 94.23%).

4 Methodology

We focus on monolingual German and multilingual language models to assess how language specialization and cross-lingual training affect models' classification performance.

We evaluate four decoder-only LLMs: two German-only: LLäMlein (Pfister et al., 2024) and BübLeLM (Delobelle et al., 2024), and two multilingual models: EuroLLM-9B (Martins et al., 2024) and Teuken 7B Instruct (Ali et al., 2024). LLäMlein 7B Chat is tuned on chat-style instruction datasets (e.g., Alpaca, Guanaco) to support both instruction following and conversational ability. Teuken Instruct follows a similar approach in a multilingual setting, covering all 24 European Union languages while retaining a decoder-only architecture. EuroLLM-9B is instruction-tuned on 35 languages, with the aim of generalizing across typologically related languages. BübLeLM is a small model trained exclusively on curated German data, such as contemporary web content, legislative documents, news data, and Wikipedia sources.

As baselines, we include GBERT-large (Chan et al., 2020) and Google RemBERT (Chung et al., 2020), two BERT-based encoder models fine-tuned for binary classification. GBERT-large is trained exclusively on German corpora, while RemBERT supports over 110 languages. These models serve as strong reference points for assessing whether instruction-tuned LLMs can match or exceed encoder-based sentence classifiers. Note that we initially attempted to fine-tune XLM-RoBERTa but encountered incompatibilities with 4-bit quantization. Therefore, we substituted it with RemBERT, a QLoRA-compatible multilingual encoder with similar coverage.

While encoder models like GBERT are suitable for classification, decoder and encoder–decoder LLMs were adapted by replacing their language modeling heads with a lightweight classifier trained using QLoRA (Dettmers et al., 2023). This setup enables sentence-level classification by extracting the final hidden state of the last token, aligning with recent work on LLM adaptation for text classification tasks (Nowacki et al., 2025).

All models were fine-tuned on the same balanced dataset (600 advancing acts, 600 non-advancing acts) using identical training parameters (e.g., learningrate = 2×10^{-4}, batchsize = 8, weightdecay = 0.01) unless model-specific adjustments were required. We used the original tokenizer associated with each model checkpoint to ensure compatibility during fine-tuning and evaluation. Each model was trained for up to three epochs, with early stopping based on validation loss trends. Instruction-tuned models were prompted uniformly. Training was conducted on a Google Colab A100 GPU using 4-bit quantization to reduce memory requirements.

Evaluation was conducted on a held-out test set using macro-F1, accuracy, and confusion matrices. Additionally, we used a balanced out-of-domain test set (390 advancing, 390 non-advancing) to evaluate generalization across domains.

5 Instruction tuning

The LLMs were fine-tuned using QLoRA on the same dataset for three epochs. We tested several prompting strategies, including 3-shot prompting and variations in example order. Since only five-shot prompting yielded a noticeable improvement, we adopted it as our final strategy (see the prompt formulation in the Appendix). The examples used in the five-shot prompts were selected from the in-domain (counseling) dataset, as the classroom dataset was reserved for out-of-domain evaluation. LLäMlein showed stable training. Training loss declined from 2.651 to 2.619, while validation loss dropped from 2.662 to 2.635. Although the reduction in loss was consistent, the overall loss values remained relatively high, and no significant diver-

gence between training and validation loss occurred. This suggests that the model was learning from the data without overfitting. However, the model failed to generate valid predictions on the unseen test set during in-domain evaluation. Precision, recall, and F1-scores were 0.00 for both classes, highlighting their inability to follow instruction-tuning prompts effectively. Teuken-7B-Instruct exhibited stable training and moderate convergence. Across three training epochs, the training loss steadily decreased from 0.4572 to 0.3908, indicating effective learning. Validation loss remained relatively stable (ranging from 0.4670 to 0.4744), suggesting that the model generalized reasonably well and did not overfit. However, on the held-out test set, recall and F1 for the advancing class were 1.00, and the model completely failed to identify non-advancing cases, resulting in a precision of 0.50 for advancing and 0.00 for non-advancing grounding acts. BübLeLM-2B showed stable but modest improvements in loss during training: training loss decreased from 0.4664 in epoch 1 to 0.4415 in epoch 3, while validation loss decreased slightly from 0.4996 to 0.4868. During in-domain evaluation on unseen data, it consistently predicted the non-advancing class for all test examples. This resulted in a final accuracy of 0.50 and a macro-averaged F1 score of 0.33, with an F1 of 0.00 and zero recall for the advancing class. EuroLLM-9B-Instruct outperformed other models in recall for advancing (0.82) but still failed on non-advancing predictions, leading to a micro-F1 of 0.46 and macro-F1 of 0.32.

Given that instruction tuning did not yield sufficient classification ability even for in-domain settings, we did not proceed with out-of-domain testing. These findings also support our shift to classifier-head fine-tuning as a more reliable strategy for this task.

6 Classifier-head fine-tuning

We fine-tuned all models for up to three epochs and monitored training and validation loss to assess generalization and convergence. LLäMlein showed consistent improvement, with training loss decreasing from 0.386 to 0.105 and validation loss from 0.394 to 0.356, indicating moderate generalization without signs of overfitting. BübLeLM converged rapidly, with training loss dropping from 0.51 to 0.08 and validation loss stabilizing around 0.38, suggesting effective learning and generalization across both classes.

Teuken-7B reached near-zero training loss, but exhibited a high final validation loss (1.12), indicating overfitting despite strong in-domain accuracy (85.8%). EuroLLM also showed sharp training loss reduction (0.30 to < 0.01), while validation loss decreased gradually from 0.06 to 0.02, suggesting stable convergence and slight generalization.

Among BERT-based baselines, GBERT-large demonstrated decreasing training loss (0.59 to 0.49) but unstable validation loss (ending at 0.675), which indicates a limited generalization. Google RemBERT exhibited modest training and validation loss reductions. However, the relatively high validation loss suggests that the model struggled to clearly separate the two classes, probably due to the limited training data.

6.1 In-domain evaluation

To identify the model that best distinguishes between advancing and non-advancing utterances, we evaluated all models on a balanced held-out test set from the same domain as training data (counseling domain, 60 grounding acts per class). As outlined in Section 4, macro-F1 is used as the primary metric for in-domain evaluation, supported by precision, recall, and confusion matrices to analyze class-specific behavior for in-domain settings.

Baseline models. GBERT served as a strong monolingual baseline. It achieved an accuracy of 88% and a macro-F1 score of 0.88, indicating that the model handled both classes with balanced performance. Performance was well balanced, with F1-scores of 0.87 for non-advancing and 0.89 for advancing grounding acts. These results confirm GBERT's ability to reliably distinguish between both classes, with no major bias toward either. In contrast, the multilingual RemBERT baseline achieved only 56% accuracy and a macro-F1 of 0.55. It showed moderate performance for the non-advancing category (F1: 0.61) and struggled with advancing utterances (F1:0.50), suggesting a bias toward the majority-negative class. This indicates that multilingual pretraining alone is insufficient for robust classification of grounding acts in this domain.

LLMs. Three of the four tested models – Teuken-7B, BübLeLM, and LLäMlein – outperformed both baselines in terms of macro-F1 and class-balanced performance. Specifically, BübLeLM achieved the highest in-domain accuracy (91%) and macro-F1 score (0.91), with perfectly balanced F1-scores of

	Model	Acc.	Macro		
			Prec.	Rec.	F1
1	BübLeLM	0.91	0.91	0.91	**0.91**
2	Teuken-7B	0.90	0.91	0.90	**0.90**
3	GBERT-large (BL)	0.88	0.89	0.88	**0.88**
4	LLäMlein	0.85	0.85	0.85	**0.85**
5	RemBERT (BL)	0.56	0.56	0.56	**0.55**
6	EuroLLM	0.50	0.25	0.50	**0.33**

Table 1: In-domain performance of LLMs and baseline (BL) models, ranked from best to worst based on macro-F1 score. Macro-averaged precision and recall are included to reflect class-balanced performance.

0.91 for both classes. Therefore, it was the most reliable model for binary grounding act classification. Teuken-7B also performed strongly, with a macro-F1 of 0.90 and slightly imbalanced scores for advancing (F1: 0.91) and non-advancing (F1: 0.89) acts. The confusion matrix reveals that both classes were predicted with high precision and recall (for advancing precision = 0.85, recall = 0.97, and for non-advancing precision = 0.96, recall = 0.83). The LLäMlein model achieved an overall accuracy of 85% and a macro-F1 of 0.85. It performed robustly across both classes, with an F1-score of 0.86 for advancing and 0.84 for non-advancing. These results place it just below the GBERT-large baseline (macro-F1: 0.88), but above RemBERT (macro-F1: 0.55). Despite converging during training, EuroLLM showed weak in-domain generalization, achieving only 50% accuracy and a macro-F1 of 0.33. It correctly identified all advancing utterances (recall: 1.00) but failed completely on non-advancing (F1: 0.00). The confusion matrix shows a severe class bias, indicating that the model performed far below both baseline models.

Table 1 presents the in-domain performance of all tested models, ranked by macro-F1 score (our primary evaluation metric for measuring how balanced the model performance was across classes, i.e., class-balanced reliability).

As shown in Table 1, the baseline models were outperformed by three LLMs in the in-domain-performance evaluation. A complete summary of per-class performance for each model is included in Table 6 in the Appendix. It shows a more fine-grained comparison of classification behavior across models.

6.2 Out-of-domain evaluation

To assess the generalization ability of all models, we evaluated their performance on a held-out out-of-domain dataset consisting of 780 labeled grounding acts equally distributed between the advancing and non-advancing classes. Following our methodology, we report accuracy, precision, recall, and macro-F1 scores to measure class-balanced reliability. Confusion matrices were used to examine whether the model was unfairly biased toward predicting one class over the other.

Baseline models. The GBERT-large classifier demonstrated robust out-of-domain performance, achieving an accuracy of 79% and a macro-F1 of 0.79. It performed reliably across both classes, with F1-scores of 0.78 (non-advancing) and 0.79 (advancing class), indicating that the model generalized well over data it was not trained on. In contrast, RemBERT showed weaker generalization, with an accuracy of 56% and macro-F1 of 0.55. Its performance favored non-advancing type (F1: 0.60) over advancing acts (F1: 0.49), suggesting a mild class bias.

LLMs. BübLeLM showed moderate generalization with an accuracy of 66% and a macro-F1 of 0.65. Contrary to its well-balanced in-domain behavior, the model performed slightly better on non-advancing (F1: 0.69) than on advancing grounding acts (F1: 0.62).

LLäMlein also achieved moderate out-of-domain performance, with an accuracy of 65% and macro-F1 of 0.64. It retained high recall for advancing grounding acts (86.4%) but struggled with non-advancing cases (recall = 44.4%), resulting in class imbalance. These results indicate that the model was only partially able to generalize, as it showed a constant bias toward advancing predictions.

Teuken-7B presents a more complex case. While the model achieved F1 for advancing acts (0.64), it completely failed to predict any non-advancing cases (recall = 0.00). This extreme class bias led to a macro-F1 of only 0.32, despite excellent recall for the advancing class. This pattern suggests that the model is unreliable for balanced classification across domains.

EuroLLM failed to generalize effectively. It achieved an accuracy of 49.7% and a macro-F1 of only 0.33. Like in its in-domain performance, the model predicted almost exclusively advancing utterances (advancing recall = 0.99, non-advancing recall = 0.00). These findings confirm the model's persistent class bias.

It can be concluded that only GBERT-large maintained high and balanced classification ability in

	Model	Acc.	Macro		
			Prec.	Rec.	F1
1	GBERT-large (BL)	0.79	0.79	0.79	**0.79**
2	BübLeLM	0.66	0.66	0.66	**0.65**
3	LLäMlein	0.65	0.69	0.65	**0.64**
4	RemBERT (BL)	0.56	0.56	0.56	**0.55**
5	EuroLLM	0.50	0.25	0.50	**0.33**
6	Teuken-7B	0.89	0.45	0.50	**0.32**

Table 2: Out-of-domain performance of LLMs and baseline (BL) models, ranked from best to worst based on macro-F1 score. Macro-averaged precision and recall indicate class-balanced generalization to unseen data.

the out-of-domain setting (see Table 2). While BübLeLM and LLäMlein generalized moderately well, their macro-F1 scores (0.65 and 0.64, respectively) remained below that of GBERT (0.79). Both models showed class-specific weaknesses, favoring either non-advancing or advancing acts, but avoided permanently predicting only one class. By contrast, Teuken-7B and EuroLLM failed to recognize one class altogether, leading to very low macro-F1 scores (0.32 and 0.33, respectively). Despite its overall poor performance, RemBERT still showed a minimal ability to differentiate between the two classes (macro-F1: 0.55).

Note that Macro-F1 was the key criterion for reliable class distinction and not raw accuracy. Therefore, despite high accuracy, Teuken-7B ranks fifth because it failed to predict the non-advancing class entirely. Detailed per-class performance metrics, including precision, recall, and F1-scores for both classes, are reported in Table 7 in the Appendix.

6.3 In-domain error analysis

To better understand the limitations of the evaluated models, we conducted a separate error analysis for both in-domain and out-of-domain settings. This includes an analysis of both advancing and non-advancing classes, with attention to common error types (false positives and false negatives) and systematic class biases.

Baseline models. The two baseline models, GBERT-large and Google RemBERT, exhibited distinct error patterns despite both achieving moderate in-domain performance (see Appendix Table 1. GBERT produced only 14 misclassifications (11.67% error rate), from which 12 errors were false positives – non-advancing utterances incorrectly labeled as advancing. These cases often involved longer utterances. Notably, false negatives (missed advancing utterances) were rare. In contrast, RemBERT showed a stronger imbalance, with 59 total errors, including 43 false negatives that usually occurred in utterances lacking explicit markers of conversational progress.

LLMs. Both qualitative and quantitative error analysis show that among the LLM-based classifiers, performance varied widely. BübLeLM showed the highest robustness, with only 11 out of 120 instances (9.17% error rate). Most of its errors overlapped with disagreements among human annotators. Teuken-7B also demonstrated strong performance, misclassifying twelve examples. LLäMlein showed a higher error rate (15%), with a tendency to overpredict advancing for non-advancing utterances, probably due to their brevity or ambiguity. In contrast, EuroLLM misclassified all 60 non-advancing utterances as advancing, producing a 50% error rate and revealing a huge class prediction bias (see Appendix Table 1.

Overall, these results imply that while multilingual models like RemBERT struggle with subtle classification, classifier-head fine-tuning on German-only LLMs (Teuken-7B and BübLeLM) can yield models that perform competitively with strong baselines like GBERT in domain-specific dialogue classification (for a summary of in-domain errors across all models, see Table 8 in the Appendix).

6.4 Out-of-domain error analysis

In the out-of-domain setting, we examine model errors for both advancing and non-advancing classes, focusing on the distribution of false negatives and false positives. This allows us to assess class-specific weaknesses and whether models exhibit biases when generalizing to unseen domains.

Baseline models. The error analysis of Google RemBERT revealed difficulties in distinguishing between advancing and non-advancing utterances across domains, with a high false negative rate (56%) for advancing utterances and frequent misclassifications of utterances lacking overt markers of advancing acts, such as *yes* or *okay* for acknowledgments. Conversely, the model also misclassified 127 non-advancing utterances as advancing mainly in cases where questions are formally similar to advancing acts but require semantic interpretation to be correctly classified as advancing acts (e.g., *Kannst du bitte deinen Namen wiederholen?* '*Can you please repeat your name?*').

The confusion matrix (see Appendix Figure 2) shows a nearly symmetric distribution of errors produced by GBERT-large, underscoring the advantage of language-specific pretraining: 75 advancing utterances were misclassified as non-advancing, and 90 non-advancing utterances were predicted as advancing. This suggests that GBERT was able to reliably distinguish between both categories and responded well to a domain shift. A qualitative error analysis reveals that many false negatives (advancing acts misclassified as non-advancing) are short utterances (e.g., *Okay, wer möchte anfangen? 'OK, who wants to start?'*) or those comprising ambiguous question tags, such as *right* (e.g., *Vielleicht noch mal vor der mündlichen noch mal vorbeikommen, so dass man dann da bespricht, oder? 'Maybe you could come again before the oral exam so we can discuss everything then, right?'*). On the other hand, false positives, where non-advancing utterances were misclassified as advancing, included cases like *Können Sie die Frage nochmal wiederholen bitte? 'Can you please repeat your question?'* and *Ja, das heißt also ich brauche an den Seminaren nicht extra teilnehmen. 'Okay, you mean I don't need to attend the course again?'*. In these cases, the model incorrectly interpreted redirecting attention to another speaker or an initial short acknowledgment (e.g., "okay") as indicators of advancing acts. These results suggest that GBERT's generalization across domains often relies on formal properties of utterances rather than on recognizing pragmatic cues and context interpretation.

LLMs. Among the four LLMs, BübLeLM demonstrated the most balanced out-of-domain performance, correctly identifying 59% of advancing and 71% of non-advancing utterances. LLäMlein showed a notable bias toward the advancing class, misclassifying 217 out of 390 non-advancing utterances (55.6%). However, the qualitative error analysis shows that both models exhibit similar error patterns, as advancing acts were often misclassified in utterances comprising the teacher's self-repair and question reformulations (e.g.,*Was könnt, möchtet ihr ergänzen? 'What do you want to add? What can you add?'*). EuroLLM and Teuken performed the worst, failing to identify a single non-advancing utterance, thereby misclassifying all instances of the non-advancing class as advancing and generating 100% false positives for those examples. These results show differences between LLMs regarding their ability to generalize the distinction between advancing and non-advancing grounding acts across domains. Compared to the two baseline models, the LLM-based classifiers exhibited more varied error patterns in the out-of-domain setting (see Table 9). Among the LLMs, BübLeLM most closely resembled GBERT in behavior, with moderately balanced predictions across both classes.

These results highlight that both classes pose challenges under domain shift, though models vary in which class they tend to overpredict. While some models overgenerate advancing acts, others show difficulty recognizing non-advancing behaviors such as clarification or self-repair.

7 Discussion and conclusions

To systematically assess the usefulness of various German and multilingual models for binary classification of grounding acts (advancing vs. non-advancing), we conducted fine-tuning experiments using both instruction-tuning and classifier-head training strategies. Four large language models: LLäMlein, Teuken, EuroLLM-Instruct, and BübLeLM, were evaluated alongside two BERT-based baselines.

Instruction tuning vs. classification head. The instruction tuning approach (using five-shot prompts) proved ineffective for binary classification of grounding acts: both LLäMlein and EuroLLM-Instruct failed to predict either class correctly, yielding F1-scores of 0.0 and producing invalid outputs despite prompt familiarity. These results indicate that instruction-tuned generation is unsuitable for tasks that require strict adherence to fixed target labels. In contrast, fine-tuning a classifier head led to successful classification of both classes and substantially improved performance, demonstrating the superiority of this method for the application of LLMs in classification tasks.

The performance of classifier-head fine-tuned models. Among the classifier-head fine-tuned models, GBERT-large demonstrated the most reliable overall performance. While Teuken-7B achieved the highest in-domain accuracy and stability, it failed to generalize to out-of-domain data, entirely neglecting the non-advancing class. In contrast, GBERT-large maintained strong and balanced performance in both in-domain (macro-F1: 0.88) and out-of-domain settings (macro-F1: 0.79), making it the most reliable model under domain shift. Despite minor differences in training dynamics

and error rates, GBERT's consistent class-level predictions across domains make it a better candidate for robust binary classification of grounding acts compared to recent German-only and multilingual LLMs. Therefore, a BERT-based classifier can replace expensive fine-tuning of LLMs for the classification task discussed in this paper.

While this paper focused on classifying single utterances as either advancing or non-advancing, a promising direction for future research lies in modifying the granularity of the task. Specifically, models could be trained to assess short multi-turn dialogue segments and determine whether any non-advancing utterances are present. This would require models not only to recognize grounding acts in the context but also to reason over conversational context to detect breakdowns such as hesitation, ambiguity, or misunderstanding. The results of this approach could be beneficial for real-time dialogue monitoring and the development of automated tutoring systems.

8 Limitations

While our study shows that BERT-based models generalize better across domains than LLMs for grounding act classification in German, it is limited by the size and scope of the training data. Our datasets are balanced but relatively small and restricted to two institutional domains (education and counseling). As a result, the generalizability of the findings to broader conversational contexts remains an open question. Additionally, our fine-tuning experiments focus on sentence-level utterance classification, which may oversimplify cases where grounding is context-dependent or extend across multiple turns. Finally, the observed performance advantage of GBERT may not generalize to earlier BERT models, such as German BERT, which were not evaluated in this study and may differ in performance due to older training data and architecture.

9 Ethics statement

All datasets used in this study are derived from publicly available, anonymized transcripts of counseling and classroom dialogues. No private or sensitive personal information is included. Human annotation was performed following a transparent and replicable annotation schema. Our work does not involve any deployment of models in real-world end-user applications. The goal is strictly theoretical: to understand the capabilities of LLMs in classifying grounding acts. We acknowledge that automated classification of communicative intent in institutional settings may carry risks, such as misinterpretation or bias, if applied without oversight.

References

Sara Abiri. 2022. *Beratung in interkultureller Kommunikation*. Ph.D. thesis, Universität Hamburg.

Ali Ahmadvand, Jason Ingyu Choi, and Eugene Agichtein. 2019. Contextual dialogue act classification for open-domain conversational agents. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1273–1276, New York, NY, USA. Association for Computing Machinery.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Jan Ebert, Alexander Arno Weber, Richard Rutmann, Charvi Jain, Max Lübbering, Daniel Steinigen, Johannes Leveling, Katrin Klug, Jasper Schulze Buschhoff, Lena Jurkschat, Hammam Abdelwahab, Benny Jörg Stein, Karl-Heinz Sylla, Pavel Denisov, Nicolo' Brandizzi, Qasid Saleem, Anirban Bhowmick, Lennard Helmer, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Alex Jude, Lalith Manjunath, Samuel Weinbach, Carolin Penke, Oleg Filatov, Shima Asaadi, Fabio Barth, Rafet Sifa, Fabian Küch, Andreas Herten, René Jäkel, Georg Rehm, Stefan Kesselheim, Joachim Köhler, and Nicolas Flores-Herr. 2024. Teuken-7b-base & teuken-7b-instruct: Towards european llms.

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2024. Star-gate: Teaching language models to ask clarifying questions.

Sylvia Bendel. 2007. *Sprachliche Individualität in der Institution. Telefongespräche in der Bank und ihre individuelle Gestaltung*. Francke, Tübingen.

Wolfgang Boettcher, Anika Limburg, Dorothee Meer, and Vera Zegers. 2005. *„ich komm (0) weil ich wohl etwas das thema meiner hausarbeit etwas verfehlt habe," – Sprechstundengespräche an der Hochschule. Ein Transkriptband*. Verlag für Gesprächsforschung.

Ines Bose, Katja Bößhenz, Judith Pietschmann, and Ingmar Rothe. 2012. „°hh hh° also von kundenfreundlich halt ich da nicht viel bei ihnen;" – Analyse und optimierung von Callcenterkommunikation am Beispiel von telefonischen Reklamationsgesprächen. *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion*, 12:143–195.

Evald Johannes Brunner. 1996. *Grundfragen der Familientherapie: Systemische Theorie und Methodologie*. Springer, Berlin.

Gisela Brünner. 2000. *Wirtschaftskommunikation: Linguistische Analyse ihrer mündlichen Formen*. Niemeyer, Berlin.

Youngjin Chae and Thomas Davidson. 2025. Large language models for text classification: From zero-shot learning to instruction-tuning. *Sociol. Methods Res.*

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023. Controllable mixed-initiative dialogue generation through prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models.

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cogn. Sci.*, 13(2):259–294.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Pieter Delobelle, Alan Akbik, et al. 2024. BübleLM: A small German LM.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLORA: Efficient finetuning of quantized LLMs. *Advances in neural information processing systems*, 36:10088–10115.

Sorouralsadat Fatemi, Yuheng Hu, and Maryam Mousavi. 2025. A comparative analysis of instruction finetuning large language models for financial text classification. *ACM Trans. Manage. Inf. Syst.*, 16(1).

Jennifer Hartog. 1996. *Das genetische Beratungsgespräch. Institutionalisierte Kommunikation zwischen Experten und Nicht-Experten*. Narr, Tübingen.

Gesine Herzberger. 2013. *Das sprachliche und kommunikative Verhalten von Behördenmitarbeitern – Agenten-Klienten-Gespräche in einer Ausländerbehörde*. Ph.D. thesis, Univ. Würzburg.

Yann Hicke, Abhishek Masand, Wentao Guo, and Tushaar Gangavarapu. 2023. Assessing the efficacy of large language models in generating accurate teacher responses. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 745–755, Toronto, Canada. Association for Computational Linguistics.

Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2025. On limitations of LLM as annotator for low resource languages.

Kristiina Jokinen, Phillip Schneider, and Taiga Mori. 2024. Towards harnessing large language models for comprehension of conversational grounding.

Katharina König. 2016. Fragen in universitären sprechstundengesprächen: Gesprächsanalyse und authentisches gesprochenes Deutsch im DaF-Unterricht. *Informationen Deutsch als Fremdsprache*, 43(1):55–88.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. CLAM: Selective clarification for ambiguous questions with generative language models.

Elena Leitner and Georg Rehm. 2025. Exploring the limits of LLMs in German text classification: Prompting and fine-tuning strategies across small and medium-sized datasets. Presented at the LLM-Fails Workshop, IDS Mannheim. Accessed: 2025-04-26.

Petra Löning and Jochen Rehbein. 1993. *Arzt-Patienten-Kommunikation: Analysen zu interdisziplinären Problemen des medizinischen Diskurses*. De Gruyter, Berlin.

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. 2024. EuroLLM: Multilingual language models for Europe.

Biswesh Mohapatra, Seemab Hassan, Laurent Romary, and Justine Cassell. 2024a. Conversational grounding: Annotation and analysis of grounding acts and grounding units. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3967–3977, Torino, Italia. ELRA and ICCL.

Biswesh Mohapatra, Manav Nitin Kapadnis, Laurent Romary, and Justine Cassell. 2024b. Evaluating the effectiveness of large language models in establishing conversational grounding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9767–9781, Miami, Florida, USA. Association for Computational Linguistics.

Arkadiusz Nowacki, Wojciech Sitek, and Henryk Rybiński. 2025. LLM-based classifiers for discovering mental disorders. *J. Intell. Inf. Syst.*

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

Melina Plakidis and Georg Rehm. 2022. A dataset of offensive German language tweets annotated for speech acts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4799–4807, Marseille, France. European Language Resources Association.

Chrysi Rapanta, Cláudia Gonçalves, João Rui Pereira, Dilar Cascalheira, Beatriz Gil, Rita Morais, Anna Čermáková, Julia Peck, Benjamin Brummernhenrich, Regina Jucks, Mercè Garcia-Milà, Andrea Miralda-Banda, José Luna, Maria Vrikki, Maria Evagorou, and Fabrizio Macagno. 2021. Multicultural classroom discourse dataset on teachers' and students' dialogic empathy. *Data Brief*, 39(107518):107518.

Thomas Schmidt. 2014. The research and teaching corpus of spoken German — FOLK. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 383–387, Reykjavik, Iceland. European Language Resources Association (ELRA).

Phillip Schneider, Nektarios Machner, Kristiina Jokinen, and Florian Matthes. 2024. Bridging information gaps in dialogues with grounded exchanges using knowledge graphs. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 110–120, Kyoto, Japan. Association for Computational Linguistics.

Peter Schröder. 1985. *Beratungsgespräche. Ein kommentierter Textband*. Narr, Tübingen.

Thomas Schubert. 2003. *Wissenstransfer im telefonischen Beratungsgespräch*. Ph.D. thesis, Martin-Luther-Universität Halle-Wittenberg.

Omar Shaikh, Kristina Gligoric, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. 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 6279–6296, Mexico City, Mexico. Association for Computational Linguistics.

Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2025. Navigating rifts in human-LLM grounding: Study and benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20832–20847, Vienna, Austria. Association for Computational Linguistics.

Anaïs Tack and Chris Piech. 2022. The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues.

David R. Traum and James F. Allen. 1992. A speech-act approach to grounding in conversation. In *Proceedings of the Second International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 137–140, Banff, Canada. University of Alberta.

David R Traum and Elizabeth A Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. *Comput. Intell.*, 8(3):575–599.

A Appendix

Five-shot prompt used for instruction tuning (translated from German)

German original:
Anweisung: Deine Aufgabe ist es zu entscheiden, ob die folgende Äußerung das Gespräch voranbringt (Label: ADVANCING) oder nicht (Label: NON-ADVANCING). Du musst immer ausschließlich mit ADVANCING oder NON-ADVANCING antworten – keine weiteren Textelemente.

Beispiel 1:
Äußerung: Darf ich Sie bitten, ein Foto mitzubringen?
Label: ADVANCING

Beispiel 2:
Äußerung: Ich werde es, ähm, also ich nehme es mit.
Label: NON-ADVANCING

Beispiel 3:
Äußerung: Das ist ein kl... kleines Blatt Papier.
Label: NON-ADVANCING

Beispiel 4:
Äußerung: Ja, da muss ich noch nachfragen.
Label: ADVANCING

Beispiel 5:
Äußerung: Sie möchten Ihr Visum erneuern, also zum Beispiel den Pass...
Label: NON-ADVANCING

English translation:
Instruction: Your task is to decide whether the following sentence moves the conversation forward (Label: ADVANCING) or not (Label: NON-ADVANCING). You must always respond with ADVANCING or NON-ADVANCING only—no other text.

Example 1:
Sentence: May I ask you to bring a photo?
Label: ADVANCING

Example 2:
Sentence: I will put it I mean take it with me.
Label: NON-ADVANCING

Example 3:
Sentence: This is a sm... small piece of paper.
Label: NON-ADVANCING

Example 4:
Sentence: Yes, I have to ask about it.
Label: ADVANCING

Example 5:
Sentence: You want to renew your visum e.g., your passport..
Label: NON-ADVANCING

Topic	Transcripts	GA	Reference
German immigration office	92	462	Herzberger 2013
academic, psychological, legal, building society, private, counseling for the homeless	9	131	Schröder 1985
telephone conversations at the bank	20	104	Bendel 2007
call-center conversations	1	8	Bose et al. 2012
sales and complaint conversations, customer service calls, negotiations, meetings	6	11	Brünner 2000
family therapy sessions	1	37	Brunner 1996
genetic counseling conversations	1	6	Hartog 1996
conversation with a breast cancer patient	1	6	Löning and Rehbein 1993
telephone counseling sessions about home financing	12	138	Schubert 2003
counseling in the refugee center	2	42	Abiri 2022
FOLK database of spoken German, sales advisory service	39	255	DGD database

Table 3: Overview of primary sources used to compile the counseling dataset with the number of used transcripts, the number of annotated grounding acts (GA), and references for each source.

Topic	Transcripts	GA	Reference
Multilingual Corpus of the DIALLS (Dialogue and Argumentation for Literacy Learning in Schools) dataset	28	467	Rapanta et al. 2021
Office hour conversations at the university	23	117	Boettcher et al. 2005
Office hour conversations at the university	1	20	König 2016
FOLK database of spoken German, classroom conversations	22	176	DGD database

Table 4: Overview of primary sources used to compile the classroom dataset with the number of used transcripts, the number of annotated grounding acts (GA), and references for each source.

Split	Count	Proportion
Training	960	80%
Validation	120	10%
Test	120	10%
In-domain advancing	600	—
In-domain non-advancing	600	—
In-domain total	**1200**	**100%**
OOD eval (advancing)	390	—
OOD eval (non-advancing)	390	—
OOD total	**780**	—

Table 5: Data splits and class distributions used for fine-tuning and out-of-domain evaluation.

	Advancing			Non-Advancing		
Model	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**nonRecall**
BübLeLM	0.91	0.93	0.88	0.91	0.89	0.93
Teuken-7B	0.91	0.85	0.97	0.89	0.96	0.83
GBERT-large	0.89	0.83	0.97	0.87	0.96	0.80
LLäMlein	0.86	0.82	0.90	0.84	0.89	0.80
RemBERT	0.50	0.58	0.43	0.61	0.55	0.68
EuroLLM	0.67	0.50	1.00	0.00	0.00	0.00

Table 6: Per-class evaluation metrics for all models on the in-domain test set. Values reflect each model's ability to distinguish both classes reliably.

	Advancing			Non-Advancing		
Model	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**nonRecall**
GBERT-large	0.79	0.78	0.81	0.78	0.80	0.77
BübLeLM	0.62	0.69	0.56	0.69	0.63	0.75
LLäMlein	0.71	0.61	0.86	0.56	0.77	0.44
RemBERT	0.49	0.57	0.44	0.60	0.54	0.67
Teuken-7B	0.64	0.50	0.89	0.00	0.00	0.00
EuroLLM	0.66	0.50	0.99	0.00	0.00	0.00

Table 7: Per-class evaluation metrics for all models on the out-of-domain test set. Values reflect each model's ability to distinguish both classes in unseen domains.

Model (Ranked Best to Worst)	**Misclassified Examples (out of 120)**	**Error Rate (%)**
BübLeLM	11	9.17
Teuken-7B	12	10
GBERT-large	14	11.67
LLäMlein	18	15.00
EuroLLM	60	50.00

Table 8: In-domain classification error rates for all models on the held-out test set (120 utterances), ranked from best to worst. Lower values indicate better classification reliability.

Model	**FN: ADV → NON**	**FP: NON → ADV**	**Class Bias**
GBERT	75 (19.2%)	90 (23.1%)	Low
BübLeLM	159 (40.8%)	112 (28.7%)	Moderate
LLäMlein	53 (13.6%)	217 (55.6%)	High ADV bias
RemBERT	221 (56.7%)	127 (32.6%)	Mild NON bias
EuroLLM	0	390 (100%)	Extreme ADV bias
Teuken	0	390 (100%)	Extreme ADV bias

Table 9: Out-of-domain misclassification patterns: false negatives (FN), false positives (FP), and class prediction bias. Ordered from best to worst model.

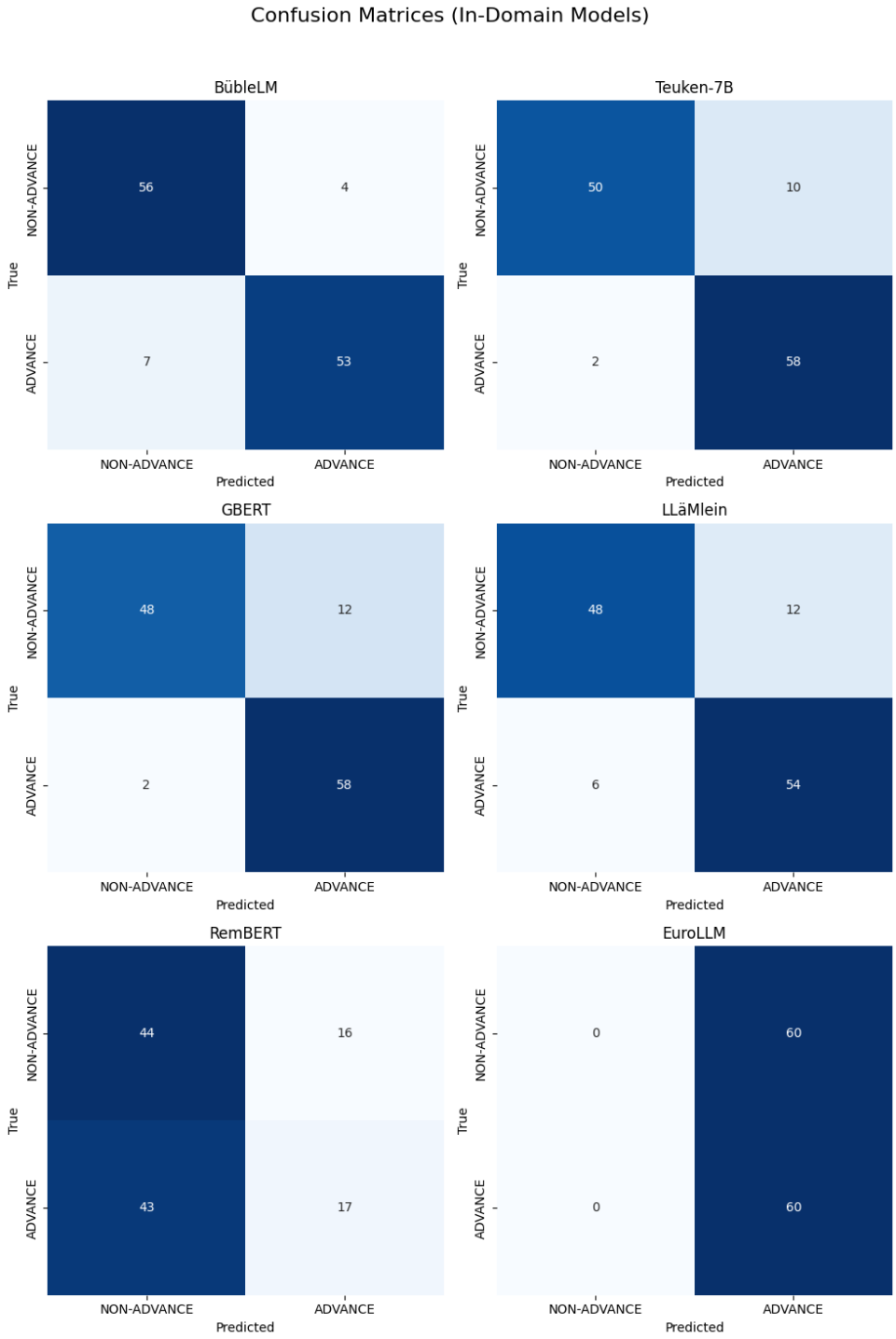

Figure 1: Confusion matrices for all models (in-domain)

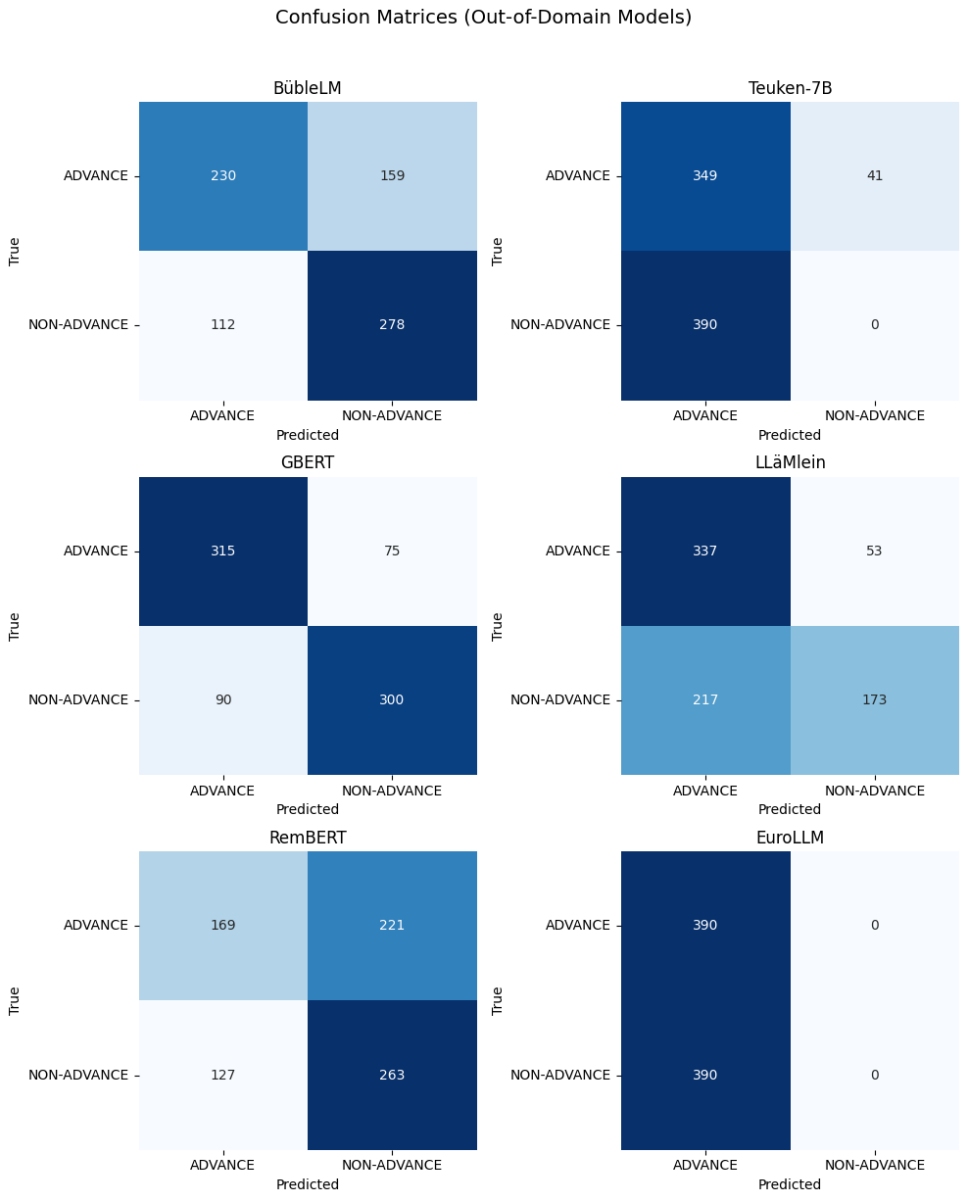

Figure 2: Confusion matrices for all models (out-of-domain)