Annotating Conversational Phases and Communication Techniques: A Corpus of German Teacher-Parent Counseling Conversations

Tobias Hallmen[1], **Kathrin Gietl**[2], **Karoline Hillesheim**[2], **Annemarie Friedrich**[3], **Elisabeth André**[1]
[1]Chair for Human-Centered Artificial Intelligence
[2]Chair of Primary School Pedagogy and Primary School Didactics
[3]Chair of Computational Linguistics
University of Augsburg
{tobias.hallmen, kathrin1.gietl, karoline.hillesheim, annemarie.friedrich, elisabeth.andre}@uni-a.de

Abstract

Teacher-parent conversations are critical for student success, yet teachers often lack structured training in counseling communication skills. We present the first annotated corpus of teacher-parent counseling conversations consisting of 59 German dialogues (approximately 6k sentences, 21k annotations) simulated by prospective elementary school teachers, peers, and professional actors. The corpus features theory-grounded annotations for conversational phases (Beginning, Informational, Argumentative, Decision-Making, Concluding) and communication techniques (Paraphrasing, Verbalizing, Structuring). We provide detailed annotation guidelines operationalizing established counseling pedagogy frameworks for computational analysis. Inter-annotator agreement analysis reveals substantial agreement (Fleiss' κ = 0.669 to 0.724, Krippendorff's α = 0.666 to 0.735). Our analysis reveals confusion patterns, providing insights into counseling discourse structure. Baseline experiments with BERT-based models and open-source LLMs achieve F1 scores of up to 71% depending on task and model. The corpus, guidelines, and baseline code are publicly available under CC BY-NC-SA 4.0 license, enabling research on automated dialogue analysis and AI-based training tools for teacher education.

Keywords: teacher-parent counseling, counseling competence, teacher training, educational dialogue, dialogue annotation, conversational phases, communication techniques, German corpus

1. Introduction

Teacher-parent conversations are critical for student success, with research showing that effective parent-teacher collaboration improves academic performance and reduces behavioral problems (Cox, 2005; Epstein and Van Voorhis, 2001). However, these conversations are often reported as stressful and unsatisfactory by both parties (Landert et al., 2009; Sacher, 2005). To mitigate this situation, structured training in counseling communication skills is increasingly taught in teacher education programs.

Professional counseling frameworks identify specific conversational phases (e.g. information gathering, argumentation, decision-making) and communication techniques (e.g. paraphrasing, verbalizing emotions, structuring) as essential for effective counseling (Benien, 2003; Gerich, 2016). Such frameworks are used as the theoretical basis in teacher education. Natural language processing (NLP) methods can help to create automated feedback systems. In this paper, we address the lack of annotated conversational data to enable the computational analysis of counseling skills.

We present a German corpus of 59 teacher-parent counseling conversations consisting of approximately 6k sentences annotated for conversational phases and communication techniques (ca. 21k annotations) according to established counseling pedagogy frameworks. The conversations are human-enacted simulations between prospective elementary school teachers, peers, and trained actors playing parental roles. They cover diverse topics such as school transitions, behavioral concerns, and learning difficulties.

As main contributions we present a richly annotated corpus of teacher-parent counseling conversations with gold standard labels for 5 conversational phases (with subcategories) and 3 communication technique types (with subtypes). We develop detailed annotation guidelines operationalizing counseling theory. We analyse inter-annotator agreement carefully, finding substantial agreement (0.669 for phases, 0.724 for techniques; Landis and Koch 1977). Our discussion highlights confusion patterns that occur in the data in the case of disagreements. Our baseline experiments on techniques with BERT-based models and large language models (LLMs) achieve per class up to 71% F1, yielding insights to optimal context length and class difficulties.

This dataset enables research on automated dialogue analysis, counseling skill assessment, and can inform the development of training tools for prospective teachers.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6361–6372
11–16 May 2026.

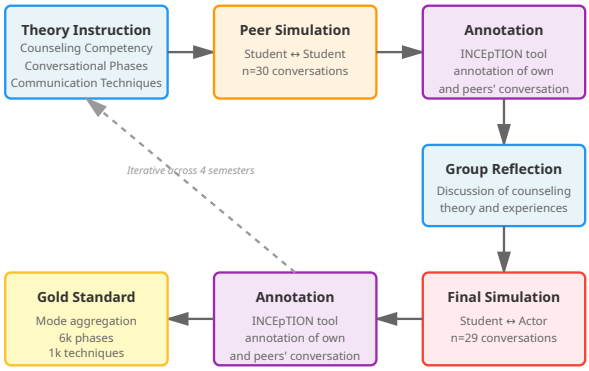

Figure 1: Overview of the data collection and annotation process. Students progress from theory instruction through peer simulations with self-annotation and reflection, culminating in final simulations with professional actors. The process is repeated across four semesters (n=59 total conversations).

2. Related Work

Our work intersects three main research areas: annotated dialogue corpora for professional communication, corpora specific to teacher-parent communication, and automated analysis of counseling skills. We position our contribution within each of these domains.

2.1. Dialogue Corpora for Professional Communication

The landscape of annotated professional dialogue datasets has expanded significantly over the past decade across medical, mental health, and educational domains. The most mature areas include motivational interviewing (MI), with major annotated datasets such as AnnoMI (Pérez-Rosas et al., 2019), MI-TAGS (Cohen et al., 2024), and BiMISC (Sun et al., 2024), all employing standardized coding schemes like MISC (Motivational Interviewing Skills Code) and MITI (Motivational Interviewing Treatment Integrity).

In mental health counseling, HOPE (Sharma et al., 2020) contributed the first dialogue-act annotated dataset with approximately 12,900 utterances from 202 counseling sessions, featuring 12 hierarchical dialogue-act labels. ESConv (Liu et al., 2021) established theoretical foundations for emotional support with annotations of 8 support strategies grounded in Helping Skills Theory, including question, paraphrasing, reflection of feelings, and providing suggestions. More recently, MESC (Chu et al., 2025) extended this work multimodally with annotations covering emotion categories and therapeutic support strategies.

Medical communication datasets have primarily focused on MI and task-oriented medical history-taking. MediTOD (Saley et al., 2024) stands out as the only English medical dataset with comprehensive slot-attribute annotations developed by medical professionals, while IMCS-21 (Chen et al., 2023) offers multi-level annotation sophistication for pediatrics with token-level named entity recognition and utterance-level dialogue act classification.

2.2. Teacher-Parent Communication

To the best of our knowledge, no annotated datasets exist for teacher-parent conversations or parent-teacher conferences. This represents a significant research gap given the critical role of family-school partnerships in education (Cox, 2005; Epstein and Van Voorhis, 2001). The closest available resources focus exclusively on teacher-student interactions, including the Teacher-Student Chatroom Corpus (TSCC) (Caines et al., 2020) with 260 lessons annotated for pedagogical discourse, and TalkMoves (Suresh et al., 2022) with 567 K-12 mathematics lesson transcripts.

This gap likely stems from multiple barriers: privacy concerns around sensitive student information, consent challenges for recording family-school interactions, and methodological difficulties in collecting naturalistic conversations. Our dataset addresses this critical gap by providing the first annotated corpus of teacher-parent counseling conversations with theory-grounded (Benien, 2003; Gartmeier, 2018; Gerich et al., 2015) annotations for both conversational phases and communication techniques. While simulated with prospective teachers, peers, and professional actors, it captures the communicative challenges teachers encounter and enables research on a domain with distinct characteristics: institutional power dynamics, focus on child development and learning, and collaborative rather than clinical relationships.

2.3. Automated Analysis of Counseling Skills

Automatic detection of counseling skills has matured significantly for MI behaviors. Tanana et al. (2016) compare discrete sentence feature models against recursive neural networks across 341 MI session transcripts, achieving Cohen's κ (Cohen, 1960) greater than 0.6 for open questions, closed questions, affirmations, and giving information—representing agreement comparable to human-to-human reliability. Reflection detection reached impressive performance with maximum entropy Markov modeling achieving 93% recall and 73% precision for detecting counselor reflections (Xiao et al., 2016).

Dataset	Docs	Expert	Lay	Sentences
1st	6	2	0	862
2nd	20	3	10	10,937
3rd	17	1	8	6,185
4th	16	2	5	2,912
Total	**59**	**8**	**23**	**20,896**

Table 1: Distribution of documents, annotators, and sentences across data collection phases.

Empathy detection has evolved from binary classification to nuanced multi-dimensional assessment. Sharma et al. (2020) introduced a theoretically-grounded computational framework based on three empathy communication mechanisms: emotional reactions, interpretations, and explorations. Their multi-task RoBERTa-based model jointly learns empathy detection and rationale extraction. Jiang et al. (2023) advanced this with joint empathy detection and empathy intent recognition across 8 intent categories.

However, these methods focus primarily on MI and mental health counseling contexts. Our work establishes baseline performance for teacher-parent counseling using both BERT-based models and LLMs, revealing domain-specific challenges.We provide comprehensive inter-annotator agreement analysis including confusion patterns offering linguistic and pedagogical insights for future annotation efforts in professional dialogue.

3. Dataset

Our annotated German corpus of teacher-parent counseling conversations fills a critical gap in professional dialogue research. The dataset uniquely combines pedagogically authentic scenarios with controlled simulations, enabling open distribution while capturing the communicative challenges prospective teachers encounter in practice.

3.1. Data Collection

Our corpus consists of simulated teacher-parent counseling conversations conducted by pairs of humans as part of a teacher education seminar on counseling competence in German elementary school pedagogy. The conversations were recorded over four semesters. For an overview, see Figure 1.

Participants are prospective elementary school teachers. In peer training, they act either as teachers or parents during the seminar in 30 simulations. At the end of the seminar, two professional actors (1 male, 1 female) portrayed parents presenting realistic counseling scenarios for each student as finale, resulting in 29 simulations.

The conversations address diverse counseling situations commonly encountered in elementary education: Secondary school transition and grade-based eligibility disputes, behavioral concerns during lessons, participation refusal (e.g. swimming lessons), learning difficulties and homework completion, effects of family situations (e.g. parental divorce), and mobile device usage and classroom distractions.

An initial 6 conversations were recorded in a professional studio with high-quality but shared audio (48kHz) and multi-camera video (2160p25). To better reflect modern digital communication and reduce setup complexity, later conversations were conducted via Zoom, with separate audio (16kHz) and video streams (360p25 or 540p25) per participant.

For extraction of text, we used DISCOVER[1] (Hallmen et al., 2025b): All conversations were automatically transcribed using the transcription module with manual corrections for errors. Voice activity detection was used to segment transcriptions at the sentence level. DISCOVER's speaker diarization (assigning utterances in a shared audio to the respective speakers) module was also applied on the 6 initial recordings, but was not necessary for the following Zoom recordings with each participant having their own device.

For further details on the seminar's nature, application, and multimodality, see Hallmen et al. (2025a). For the pedagogical perspective on counseling competence development, see Gietl et al. (2026).

3.2. Dataset Statistics

The corpus is organized into four batches collected across semesters, containing 30 + 29 = 59 conversations with 6k sentences total. Multiple annotators (23 lay annotators who were students in the seminar, and 8 trained expert annotators from the teaching team or with experience in either annotation or counseling) labeled the conversations, producing approximately 21k sentence-level annotations. Through adjudication using mode (most frequent label), a gold standard was created containing: 6k conversational phase annotations across 5 main categories and 1k communication technique annotations across 3 main categories. Table 1 shows the distribution of documents and annotations across the four data collection phases and Table 2 shows the distribution of labels in the condensed gold standard.

[1] https://github.com/hcmlab/discover

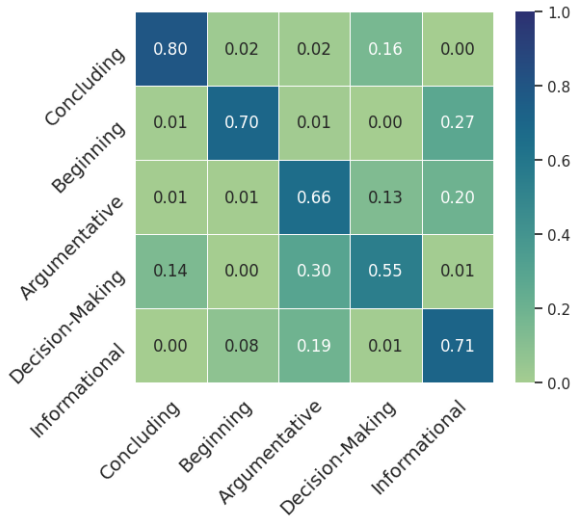

Figure 2: Coincidence matrix of the annotations for the five conversation phases. Unlike confusion matrices, coincidence matrices are not necessarily symmetric, as annotator disagreements are directional. If the sentence was annotated as row, it was also annotated as column in x% of the time.

4. Annotation Methodology

We annotate at sentence or sub-sentence level, as techniques can shift within a single utterance, whereas the sentence always belongs as a whole to a phase. Only teacher utterances are annotated for techniques, while phases apply to the entire conversation flow. Annotation is done in INCEpTION (Klie et al., 2018). To account for slightly different annotation styles, e.g. including or excluding white space or punctuation, each annotation is extrapolated to the whole sentence.

4.1. Annotation Scheme

Our annotation scheme focuses on teachers' verbal communication skills that are central to effective counseling conversations. We adapt established frameworks from counseling pedagogy (Benien, 2003; Gartmeier, 2018; Gerich et al., 2015) into a hierarchical annotation structure suitable for NLP tasks. Hence, the conversations are annotated from the teacher's point of view. The indentation differentiates supercategories from subcategories. All examples below are translated from German for readability. The corpus and annotation guidelines are entirely in German. Lines starting with "T:" stand for an utterance by the teacher, "P:" respectively for parent.

4.1.1. Conversation Phases

Following the conversation phase model of Benien (2003), we annotate five main phases that structure counseling conversations, each with specific sub-phases:

The **Beginning phase** establishes contact and clarifies the conversation framework both in content and time:

- **Greeting**s cover the exchange of greeting formulas:
  T: Good morning, Mrs. X. Great, that you have come.
  P: Good morning, Mrs. Y.
- **Small talk** includes light conversation to establish rapport for a comfortable atmosphere:
  T: Did you find your way here easily?
  P: Yes, it was fine. There was a bit of traffic, but that's okay.
- **Time frame** is used for the specification and utterance of available time:
  T: I think it's good that we're taking the time to talk about this for the next five minutes.
- **Content frame** relates to interlocutors formulating the conversation's contents and concerns: *T: [...] to talk about Emil's transfer today.*
- **Other** covers utterances in the Beginning phase that do not fall under one of the other subcategories: *T: Shall I close the window?*

The **Informational phase** involves the exchange of situation-relevant information to create a common ground without yet developing solutions:
P: We practice a lot with Emil at home. Nevertheless, he gets poor grades at school.
T: I observe that Emil works well at school. During rehearsals, he seems very nervous to me.

The **Argumentative phase** focuses on developing, gauging, and discussing potential solutions, with occasional introduction of new information:
T: There are different ways to obtain a high school diploma. Emil could also go to secondary school first.
P: I imagine that you also talk to the children about these options at school.

The **Decision-Making phase** specifies and translates proposals from the argumentative phase into concrete action steps with defined goals:
P: I plan to talk less about the transition with Emil over the next three weeks.
T: I think that's a good idea.

Finally, the **Concluding phase** wraps up the conversation, an appointment is made, and interlocutors say their goodbyes:

- The **Appreciative reflection** covers benevolent recapitulation:
  T: Thank you for your openness and for taking

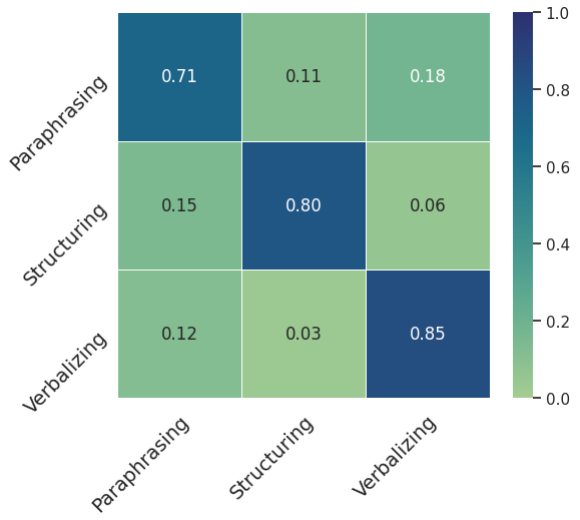

Figure 3: Coincidence matrix of the annotations for the three communication techniques. Unlike confusion matrices, coincidence matrices are not necessarily symmetric, as annotator disagreements are directional. If the sentence was annotated as row, it was also annotated as column in x% of the time.

the time. I feel that together we have found some good ways to support your child even better.

- The label **Appointment** includes scheduling for follow-up meetings:
  T: Right, shall we make an appointment right away or just call again?
  P: I'll get back to you after I've spoken to my wife.

- **Farewell** covers the exchange of parting formulas:
  P: Thank you very much, goodbye.
  T: Goodbye, Mrs. X.

- **Other** is used for all utterances belonging to the concluding phase without falling under one of the other subcategories:
  T: From my point of view, we've got it now.
  P: Yes, that's fine.

4.1.2. Communication Techniques

We annotate three core communication techniques adapted from Gerich et al. (2015), operationalized through definitions by Bay (2021), Gartmeier (2018), and Hertel and Schmitz (2010):

The technique **Verbalizing** focuses on emotions perceived in utterances:

- **Undefined attention reaction**s are brief verbal signals, keyword repetitions, or prompts:
  P: I would like to talk to my wife about discussing this again at home.
  T: Yes, okay.
  P: And I think there's another exam or rehearsal coming up next week.

- A **Statement** is directly naming of a feeling:
  P: Yeah, sorry, but I just don't understand. So, at home he can do things, yeah. And then he goes to school and gets another bad grade.
  T: Exactly, I can definitely hear a lot of anger in your voice.

- **Clarifying question**s are seeking precision about vague emotional hints:
  T: You say you always practice with him?
  P: Yes.
  T: Really always?
  P: All the time. It's all about this transfer now.

- **Further question**s are in-depth emotional statements or open questions:
  T: I'm just wondering what worries you so much when you imagine Emil going to middle school.

The **Paraphrasing** technique involves restating the factual core message in one's own words, potentially condensing or clarifying the original statement:
P: Apparently, [my son] is not good enough for high school. Well, in my opinion, that can't be true. He is intelligent and he tries hard. I practice with him every afternoon. My wife practices with him in the evenings. When we drive, we practice the multiplication tables, repeating them over and over again. And I really have to say, it seems that you are not managing to teach the children to read and write properly. [..]
T: So, as I understand it, you think that Emil actually has what it takes to go to high school?

The **Structuring** technique represents meta-communicative utterances to control conversation timing and/or content, making the conversation flow transparent through summaries or transitions:
T: So, what would be important for you to clarify in the eight minutes we have today? (refers to the succeeding conversation)
T: Okay, let me summarize. (references the preceding conversation)

4.2. Inter-Annotator Agreement

To establish annotation quality, we calculate agreement between annotators using two metrics: Fleiss' kappa (Fleiss, 1971) for overall agreement considering chance, and Krippendorff's alpha (Krippendorff, 2004) for reliability with our ordinal categories.

Type	Label	Count	Mass
Phases	Concluding	684	11.5%
	Beginning	777	13.1%
	Argumentative	2080	35.0%
	Decision-Making	591	10.0%
	Informational	1802	30.4%
Techniques	Paraphrasing	376	6.3%
	Structuring	251	4.2%
	Verbalizing	714	12.0%

Table 2: Corpus statistics for gold standard annotations.

4.2.1. Conversation phases

For conversation phases, we achieved Fleiss' κ = 0.669 ($p < 0.001$) and Krippendorff's α = 0.666 ($p < 0.001$). These values indicate substantial agreement (Landis and Koch, 1977), as can be seen in the coincidence matrix in Figure 2 (we report coincidence rather than confusion matrices, as disagreements are asymmetric).

The achieved agreement levels indicate that annotators can reliably distinguish the core of the conversational phases in teacher-parent counseling. The Beginning and Concluding phases show particularly strong agreement (observed in coincidence matrix diagonal values), reflecting their clear linguistic markers such as greetings, farewells, and explicit appointment-making language. The Informational phase is also well-defined, with annotators consistently identifying information exchange segments.

Main confusion occurs along phase boundaries and by misinterpretation of definitions – latter can also be caused by poorly executed phase transitions in the underlying conversation:

(1) **Beginning** $\leftrightarrow$ **Informational**: Transitions between establishing context and information exchange can be ambiguous, particularly when content frame setting includes substantial information

(2) **Argumentative** $\leftrightarrow$ **Informational**: Distinguishing between pure information exchange and early solution exploration requires careful interpretation

(3) **Decision-Making** $\leftrightarrow$ **Argumentative**: Final synthesis of decisions can resemble continued argumentation

These confusions reveal fundamental properties of counseling discourse rather than annotation scheme deficiencies. Phase boundaries in natural conversation are often fluid rather than discrete, and certain utterances serve multiple functions, e.g. summarizing information while proposing next steps. Phase transitions also depend on the counselor's communication skills, thereby these transitional areas may represent the most pedagogically significant moments in counseling.

4.2.2. Communication Techniques

For communication techniques, we achieved higher, substantial agreement: Fleiss' κ = 0.724 ($p < 0.001$) and Krippendorff's α = 0.735 ($p < 0.001$). The coincidence matrix (Figure 3) shows, that Verbalizing emerges as the most reliably annotated technique, benefiting from explicit emotional language and specific verbal markers. Structuring also achieves strong recognition, likely due to its meta-communicative nature and distinctive phrases like "let me summarize" or "we have X minutes remaining." Paraphrasing demonstrates solid agreement, indicating that annotators successfully identify when teachers restate parental concerns objectively in their own words. The overall strong performance suggests these techniques have sufficient linguistic distinctiveness for both manual annotation and future automated detection. Main confusion occurs between:

(1) **Paraphrasing** $\leftrightarrow$ **Verbalizing**: Both techniques involve reflecting the parent's message, but paraphrasing focuses on factual content while verbalizing targets emotions

(2) **Structuring** $\leftrightarrow$ **Paraphrasing**: Meta-summaries (structuring) often include paraphrased content

The higher agreement for techniques compared to phases aligns with our expectations: techniques have more explicit linguistic markers (e.g. "What I'm hearing is..." for paraphrasing, "Let me summarize..." for structuring), while phase boundaries depend on broader discourse context.

4.2.3. Gold Standard Creation

During the seminar, the students were taught in the specific conversation theory including phases and techniques. Guided by the instructor, they performed an exemplary annotation of a conversation. Afterwards, they annotated their own and two peers' conversations as a method of self-reflection and theory consolidation. They could always ask instructors when in doubt. After selecting the mode for each sentence to create a gold standard from the 21k annotations, 6k annotated conversation phase and 1k technique instances remain. The thus resulting gold standard benefits from multiple perspectives while maintaining high quality through expert oversight and guided student annotation. The label distribution is shown in Table 2.

Context (k)	Accuracy	F1 Score	Precision	Recall
k=1	86.4% ± 3.0%	**55.4% ± 4.2%**	**54.5% ± 5.5%**	**59.2% ± 7.9%**
k=2	**86.9% ± 2.2%**	51.7% ± 5.9%	54.0% ± 8.8%	53.8% ± 6.9%
k=3	85.1% ± 1.8%	50.8% ± 4.8%	50.2% ± 6.3%	54.4% ± 7.3%
k=4	85.7% ± 2.0%	49.4% ± 8.0%	50.6% ± 9.6%	53.7% ± 11.8%
k=5	86.9% ± 3.7%	51.9% ± 7.7%	51.6% ± 7.4%	55.0% ± 13.1%

Table 3: Impact of context window size (k preceding sentences) on technique detection with reported mean ± standard deviation for model DistilBERT. All results dataset with 4-fold stratified cross-validation.

5. Baseline Experiments

We establish baseline performance using transformer-based language models and open-weight LLMs on communication technique classification. Given the limited training data and class imbalance, we focus on multi-class classification with all three technique types plus a background class ("Other").

5.1. Experimental Setup

We fine-tune and evaluate five pre-trained transformer models representing different architectures and training objectives: DistilBERT (Sanh et al., 2019), DeBERTa-v3 (He et al., 2021), ELECTRA (Clark et al., 2020), ModernBERT (Warner et al., 2024), and XLM-RoBERTa (Conneau et al., 2019).

For the LLMs, we evaluated pretrained models in a zero-shot fashion, only providing task and technique definitions in the prompt: Llama3.2 (AI, 2024), Phi3.5 (Microsoft, 2024), Phi4 (Abdin et al., 2024), Qwen2.5 (Team, 2025), and Gemma3 (Team and Google, 2025).

For training, we apply a 4-fold stratified cross-validation using the semester in which the data was collected as the split-defining variable. We use AdamW (Loshchilov and Hutter, 2019) optimizer with learning rate 2×10^{-5}. Batch size varied from 64 to 128, depending on model size and hardware limits (24GB VRAM), with sequences truncated at 512 tokens. Training continued until early stopping with patience of 3 epochs on validation split.

To determine the optimal context length we systematically vary the amount of preceding sentences included (k = 1,…, 5 previous sentences) to assess the importance of discourse history.

5.2. Communication Techniques Classification

5.2.1. Impact of Context Length

Table 3 shows the impact of discourse context on technique classification performance using DistilBERT. We formulate the task as 4-way classification: Paraphrasing, Structuring, Verbalizing, and Other (background class including non-technique utterances).

Counter-intuitively, there is no clear winner and bigger context does not help – the confidence intervals all overlap, therefore no context length is significantly better than any other. For simplicity and computational efficiency we stick to single-sentence context ($k = 1$) as best performance achieving. Additional context increases standard deviation by up to 6 percentage points, suggesting that: (1) techniques have strong local linguistic markers, and (2) longer contexts introduce noise or create more complex learning problems given our limited training data. Testing the different context lengths for the five LLMs yielded similar results with k = 1 being up to 5 percentage points better at F1 than k = 2, …, 5.

5.2.2. Model Comparison

Table 4 compares all five BERT-like models using single-sentence context ($k = 1$) on the 4-way classification task.

XLM-RoBERTa substantially outperforms other models, achieving approximately 63% F1 score. This up to 13 point improvement over other variants suggests that: (1) larger capacity and more sophisticated pre-training objectives help with this nuanced task, and (2) cross-lingual pre-training (XLM-R) may benefit from exposure to counseling-related concepts across languages or just more exposure to German in general.

Performance of pretrained LLMs is underwhelming with F1 scores ranging from 3% to 38%, with the best model being Gemma3. We then did an experiment targeting model size: Gemma3 in variations 4B, 12B, and 27B parameters. 4B was the worst, 12B performed best, with 27B already showing signs of degradation. Maybe more pretraining knowledge can lead to over-interpretation of concepts.

5.2.3. Per-Class Performance

Table 5 shows detailed per-class performance for the best-performing model (XLM-RoBERTa) averaged across cross-validation folds.

Model	Parameters	Accuracy	F1 Score	Precision	Recall
XLM-RoBERTa	0.27B	**88.7% ± 3.9%**	**62.8% ± 5.2%**	**65.7% ± 3.7%**	**62.8% ± 9.8%**
DeBERTa-v3	0.18B	87.4% ± 4.1%	59.3% ± 6.3%	59.3% ± 7.9%	61.8% ± 9.0%
ModernBERT	0.15B	87.1% ± 3.0%	50.5% ± 5.8%	54.4% ± 9.3%	50.7% ± 8.0%
DistilBERT	0.07B	86.4% ± 3.0%	55.4% ± 4.2%	54.5% ± 5.5%	59.2% ± 7.9%
ELECTRA	0.11B	85.3% ± 2.5%	50.6% ± 8.3%	50.2% ± 9.1%	56.4% ± 13.4%

Table 4: Model comparison for multi-class technique classification ($k = 1$). Macro-averaged F1, Precision, and Recall across all four classes (Paraphrasing, Structuring, Verbalizing, Other).

Performance varies substantially across classes: Verbalizing achieves the highest F1 (70.6%), likely due to explicit emotional language markers, Structuring proves most difficult (37.0% F1), possibly due to its relative rarity and overlap with other techniques, and Paraphrasing settles in between with 50.3% F1. Clearly the dominance of the background class Other with 77.5% mass is trained well, thereby easily distinguished (93.3% F1). High standard deviations reflect cross-fold variability and the challenges of the dataset.

Additionally, some model "errors" may reflect pedagogically significant moments: when the model predicts a technique should be used based on context but the prospective teacher did not employ it, this discrepancy could identify missed counseling opportunities valuable for training feedback.

6. Conclusion

This work provides an initial annotated corpus for teacher-parent counseling competence training, addressing a critical gap in both educational technology and dialogue corpus research. The substantial inter-annotator agreement achieved demonstrates that theory-grounded counseling concepts can be reliably operationalized for computational analysis, despite the inherent complexity of professional dialogue. We established a baseline showing how AI analysis can enhance professional communication training for prospective teachers.

6.1. Summary of Contributions

We developed a German corpus of 59 simulated teacher-parent counseling conversations containing approximately 6k sentences, annotated for conversational phases and communication techniques based on established counseling pedagogy frameworks. The annotation scheme operationalizes theoretical constructs into computationally tractable categories, achieving substantial inter-annotator agreement with Fleiss' κ of 0.669 for phases and 0.724 for techniques, Krippendorff's α of 0.666 and 0.735 respectively. Analysis of confusion patterns revealed interpretable disagreements at phase boundaries and between related techniques, highlighting the inherent complexity of counseling discourse annotation.

Our experiments training transformer-based models to detect communication techniques established baseline performance and revealed task-specific challenges. The best model, XLM-RoBERTa, achieved 62.8% F1 score for multi-class technique classification, with substantial variation across classes. Surprisingly, single-sentence context outperformed longer discourse contexts, suggesting that techniques have strong local linguistic markers but that our limited training data makes longer contexts problematic.

6.2. Future Directions

Several promising directions emerge from our work. Collecting additional conversations across diverse topics, participant demographics, and conversation settings would strengthen statistical analyses and improve model robustness. Longitudinal data tracking the same students across multiple simulations could reveal learning trajectories and identify effective feedback mechanisms. The confusion patterns in our annotations suggest opportunities for refinement through clearer operational definitions, additional annotator training, or allowing multi-label annotations for utterances serving multiple functions.

Improving automated technique detection could be achieved through data augmentation using LLMs validated by experts, multi-task learning that jointly learns phase detection and technique classification, or developing domain adaptation methods to transfer models across conversation settings.

While students perceived the theory and application thereof in conversation and annotation positively, we have not yet assessed whether it improves their counseling competence over time. Building on acceptance studies of AI-based communication feedback in comparable simulation settings (Bauermann et al., 2025), future work should compare learning outcomes between students receiving AI feedback based on automatic annotation of transcripts versus traditional feedback only, track skill development across multiple simulations

Class	F1 Score	Precision	Recall
Paraphrasing	50.3% ± 11.6%	59.1% ± 7.8%	46.6% ± 20.6%
Structuring	37.0% ± 12.3%	43.3% ± 15.8%	33.5% ± 11.1%
Verbalizing	70.6% ± 15.4%	68.1% ± 22.8%	76.9% ± 10.8%
Other	93.3% ± 2.4%	92.5% ± 4.3%	94.1% ± 2.3%

Table 5: Techniques per-class performance of XLM-RoBERTa on dataset (4-fold CV, $k = 1$). Mean $\pm$ standard deviation.

and into authentic teaching practice, and investigate which feedback modalities most effectively support learning.

7. Limitations

Several limitations must be acknowledged. With only 59 conversations, our dataset is small compared to other professional dialogue corpora, limiting the statistical power of our analyses and the generalizability of our models. Our conversations are simulations between prospective teachers, peers, and professional actors rather than authentic teacher-parent interactions, which may not fully capture the complexity, emotional intensity, and power dynamics of real counseling situations. We achieved substantial inter-annotator agreement (Fleiss' κ = 0.669 to 0.724), demonstrating reliable operationalization of counseling theory for computational analysis. While our values are lower than agreement observed in simpler dialogue tasks, they align with agreement rates reported for other professional dialogue annotation involving nuanced pragmatic judgments (κ = 0.67; Cohen et al. 2024). Currently, verbal features require manual annotation as our baseline models' performance of 62.8% F1 remains insufficient for reliable automated feedback. Finally, our dataset is entirely in German within the German elementary school context, and the generalizability to other languages and educational systems remains unexplored.

We additionally evaluated transfer of (trained) models to a doctor-patient communication dataset to assess cross-context generalizability, but found substantially lower performance (F1: 19 to 43%), likely due to domain differences (one-sided vs. working together), incomplete conversations (only information phase), and severe class imbalance (heavily verbalizing).

Data Availability

The complete dataset, including transcripts, annotations, and annotation guidelines, will be publicly available on GitHub at https://github.com/saveli/tpcc, under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. The dataset includes: 59 conversation transcripts in German with sentence-level segmentation, gold standard annotations for conversational phases and communication techniques, complete annotation guidelines (in German), and baseline model code and evaluation scripts.

Ethics Statement

All conversations are simulated role-plays between prospective teachers and professional actors. No authentic teacher-parent conversations involving real families were recorded. All participants are adults and provided informed consent for recording and research use. The seminar was a non-mandatory elective within the teacher education program, offering academic credits upon completion. Transcripts were reviewed to ensure complete anonymization with no identifying information remaining.

Acknowledgements

We thank all students who participated in the seminar simulations and annotation work, whose engagement made this corpus possible.

AI tools (Claude, DeepL) were used to assist with rephrasing and translation suggestions. All final decisions on wording and content are the authors' responsibility. Translations of German examples and definitions are provided for readability only. The guidelines and corpus themselves are in German.

This work was partially funded by the Stiftung Innovation in der Hochschullehre (KodiLL project, FBM2020, grant no. 120).

Bibliographical References

Marah I. Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R.

Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report.

Meta AI. 2024. The llama 3 herd of models.

Moritz Bauermann, Thomas Rotthoff, Tobias Hallmen, Miriam Kunz, Elisabeth André, and Ann-Kathrin Schindler. 2025. Medical students' perceptions of ai-based feedback and feedforward on communication skills in doctor–patient consultation-an acceptance study in a video-based simulation. *Medical Education Online*, 30(1):2592414.

Rolf H Bay. 2021. *Erfolgreiche Gespräche durch aktives Zuhören*. expert verlag.

Karl Benien. 2003. Schwierige gespräche führen. *Modelle für Beratungs-, Kritik-und Konfliktgespräche im Berufsalltag. Rowohlt, Reinbeck bei Hamburg*.

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. The teacher-student chatroom corpus. *arXiv preprint arXiv:2011.07109*.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2023. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1):btac817.

Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. 2025. Towards multimodal emotional support conversation systems. *IEEE Transactions on Multimedia*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Ben Cohen, Moreah Zisquit, Stav Yosef, Doron Friedman, and Kfir Bar. 2024. Motivational interviewing transcripts annotated with global scores. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11642–11657.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Diane D Cox. 2005. Evidence-based interventions using home-school collaboration. *School Psychology Quarterly*, 20(4):473.

Joyce L Epstein and Frances L Van Voorhis. 2001. More than minutes: Teachers' roles in designing homework. *Educational psychologist*, 36(3):181–193.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Martin Gartmeier. 2018. *Gespräche zwischen Lehrpersonen und Eltern*. Springer.

Mara Gerich. 2016. *Teachers 'Counseling Competence in Parent-Teacher Talks: Modeling, Intervention, Behavior-Based Assessment*. Springer.

Mara Gerich, Simone Bruder, Silke Hertel, Monika Trittel, and Bernhard Schmitz. 2015. What skills and abilities are essential for counseling on learning difficulties and learning strategies? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*.

Kathrin Gietl, Karoline Hillesheim, Moritz Bauermann, Tobias Hallmen, and Andreas Hartinger. 2026. Förderung der beratungskompetenz von studierenden des lehramts an grundschulen durch simulierte elterngespräche und ki-basiertes feedback. erste ergebnisse aus einem interdisziplinären projekt. In *Bezugsnotwendigkeiten der Grundschule. Pädagogik und Fachdidaktik in der Grundschulbildung*, pages 321–330.

Tobias Hallmen, Kathrin Gietl, Karoline Hillesheim, Moritz Bauermann, Annemarie Friedrich, and Elisabeth André. 2025a. Ai-based feedback in counselling competence training of prospective teachers. *arXiv preprint arXiv:2505.03423*.

Tobias Hallmen, Dominik Schiller, Antonia Vehlen, Steffen Eberhardt, Tobias Baur, Daksitha Withanage, Wolfgang Lutz, and Elisabeth André. 2025b. Discover: a data-driven interactive system for comprehensive observation, visualization, and exploration of human behavior. *Frontiers in Digital Health*, 7:1638539.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Silke Hertel and Bernhard Schmitz. 2010. *Lehrer als Berater in Schule und Unterricht*. Kohlhammer Verlag.

Liting Jiang, Di Wu, Bohui Mao, Yanbing Li, and Wushour Slamu. 2023. Empathy intent drives empathy detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6290.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Charles Landert, Martina Brägger, Dachverband Schweizer Lehrerinnen, and LCH Lehrer. 2009. Lch arbeitszeiterhebung 2009 (aze'09): Bericht zur erhebung bei 5'000 lehrpersonen im zeitraum oktober 2008-september 2009: im auftrag des dachverbandes schweizer lehrer lch; charles landert und martina brägger.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Microsoft. 2024. Phi-3.5-mini-instruct. https://huggingface.co/microsoft/Phi-3.5-mini-instruct.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

Werner Sacher. 2005. *Erfolgreiche und misslingende Elternarbeit: Ursachen und Handlungsmöglichkeiten; erarbeitet auf der Grundlage der Repräsentativbefragung an bayerischen Schulen im Sommer 2004*. Lehrstuhl für Schulpädagogik, Friedrich-Alexander-Univ.

Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, et al. 2024. Meditod: An english dialogue dataset for medical history taking with comprehensive annotations. *arXiv preprint arXiv:2410.14204*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Xin Sun, Jiahuan Pei, Jan de Wit, Mohammad Aliannejadi, Emiel Krahmer, Jos TP Dobber, and Jos A Bosch. 2024. Eliciting motivational interviewing skill codes in psychotherapy with llms: A bilingual dataset and analytical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5609–5621.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H Martin, and Tamara Sumner. 2022. The talkmoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. *arXiv preprint arXiv:2204.09652*.

Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of substance abuse treatment*, 65:43–50.

Gemma Team and Google. 2025. Gemma 3: A new generation of open models. Technical report, Google.

Qwen Team. 2025. Qwen2.5 technical report.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Bo Xiao, Dogan Can, James Gibson, Zac E Imel, David C Atkins, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2016. Behavioral coding of therapist language in addiction counseling using recurrent neural networks. In *Interspeech*, pages 908–912.