Transition-Matrix Regularization for Next Dialogue Act Prediction in Counselling Conversations

Anonymous ACL submission

Abstract

This paper studies how empirical dialogue-flow statistics can be incorporated into Next Dialogue Act Prediction (NDAP). A KL regularization term is proposed that aligns predicted act distributions with corpus-derived transition patterns. Evaluated on a 60-class German counselling taxonomy using 5-fold cross-validation, this improves macro-F1 by 9–42% relative depending on encoder and substantially improves dialogue-flow alignment. Cross-dataset validation on HOPE suggests that improvements transfer across languages and counselling domains. In systematic ablations across pretrained encoders and architectures, the findings indicate that transition regularization provides consistent gains and disproportionately benefits weaker baseline models. The results suggest that lightweight discourse-flow priors complement pretrained encoders, especially in fine-grained, data-sparse dialogue tasks.

1 Introduction

NDAP forecasts the communicative function of the *upcoming* utterance from the dialogue history. Although the task has a long tradition in dialogue research (Nagata and Morimoto, 1994; Stolcke et al., 2000; Reithinger et al., 1996), it has received less attention in the era of large language models (LLMs). Yet it offers a structured and interpretable mechanism for steering LLM behavior, complementing prompting (Brown et al., 2020), instruction tuning (Wei et al., 2022), and reward-based approaches (Ouyang et al., 2022). By anticipating the next dialogue action, systems can condition prompts or constraints to encourage more stable, coherent, and goal-directed behavior (Chen et al., 2023).

In counselling and other high-structure domains, next acts follow consistent pragmatic patterns: greetings typically precede problem statements, exploration precedes intervention, and closing behaviors follow resolution (Bickmore et al., 2013; Althoff et al., 2016). Classical dialogue managers explicitly modeled these transitions through Markov or CRF-based structures (Stolcke et al., 2000; Zimmermann, 2009). Modern neural systems, however, largely abandon symbolic transition models and instead rely on end-to-end architectures to infer discourse structure implicitly (Ultes et al., 2017; Ravuru et al., 2022).

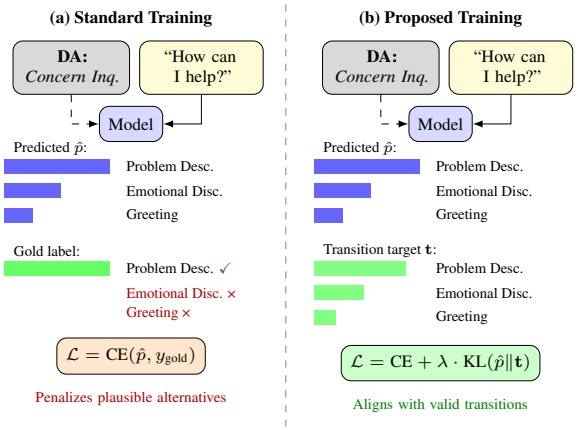

Figure 1: Comparison of standard NDAP training (a) vs. transition-matrix regularized training (b). Standard cross-entropy penalizes all non-gold predictions equally, even when multiple next acts are plausible. The proposed regularizer aligns predictions with empirical transition patterns from the corpus.

This shift removes an inductive bias. Neural models see only a single gold next-act label per instance, providing limited signal when several next acts are plausible—a common situation in counselling (Wu et al., 2022b; Demasi et al., 2020). The gold label in NDAP is inherently under-specified: it represents one observed continuation among many valid possibilities. Standard cross-entropy supervision thus penalizes the model for predicting other plausible acts. Consequently, models may struggle to capture the multi-modal distribution of plausible next actions (Zhao et al., 2017).

This limitation is addressed by incorporating an

empirical transition matrix directly into the loss. The **transition-matrix KL regularizer** encourages the predicted next-act distribution to align with observed transition statistics, injecting pragmatic discourse-flow information as a soft, differentiable constraint (Figure 1). This preserves the flexibility of neural encoders while reinstating a structural prior reminiscent of classical systems.

This idea is evaluated in German text-based counselling, where communicative actions are fine-grained and governed by psychosocial norms. The dataset uses a five-level taxonomy with 60 dialogue act categories (Albrecht et al., 2025). NDAP is performed across all speaker transitions. To exploit the taxonomy structure, category history augmented architectures are introduced.

The results show that transition-based regularization provides consistent gains and disproportionately benefits weaker models. The regularizer is lightweight, model- and architecture-agnostic, and can be integrated as a drop-in objective without architectural modifications or sequence-level decoding. Improvements span both predictive metrics (F1, Top-3) and structural measures of dialogue-flow alignment.

These results show that domain-specific structural priors can enhance NDAP and offer a mechanism for steering LLM-driven systems in specialized domains (see Appendix A.2.1 for client-side-specific evaluation relevant to client simulation). The annotated counselling dataset is released to support further research on controllable and structured dialogue modelling.

2 Related Work

Dialogue act (DA) prediction is a well-established task that assigns communicative functions to observed utterances. Classical systems model discourse structure using stochastic grammars, HMMs, or CRFs (Stolcke et al., 2000; Geertzen, 2009), while recent neural approaches employ hierarchical encoders, contextual attention, and multimodal cues (Colombo et al., 2020). These models capture local and long-range structure but operate entirely on *observed* utterances.

In contrast, our work addresses the fundamentally harder task of NDAP: forecasting the communicative function of the *upcoming* utterance without access to its surface form. Early statistical work showed that DA sequences exhibit strong structural regularities, with n-gram models improving prediction through explicit transition constraints (Reithinger et al., 1996; Geertzen, 2009). Neural NDAP research remains limited. Prior work integrates multi-turn history using hierarchical attention (Tanaka et al., 2019), incorporates multimodal features, or employs semi-supervised consistency objectives (He et al., 2022). Latent-variable architectures (Ji et al., 2016) regularize discourse trajectories, yet rely on implicitly learned transitions. Recent studies in specialized domains (e.g., talk-move prediction, counselling) forecast future strategies or acts as auxiliary signals (Ganesh et al., 2021; Wu et al., 2022b), but do not incorporate explicit transition-based priors. Overall, existing NDAP approaches condition on DA history but do not constrain predictions using empirical dialogue-flow statistics.

Explicit modeling of DA transitions is well explored in *sequence labeling*, where the goal is to assign a DA label to every utterance in a conversation. Classical systems encode transitions through n-gram dialogue grammars or HMMs (Stolcke et al., 2000), and neural architectures commonly add CRF layers to impose label-transition structure (Chen et al., 2018; Shang et al., 2020). These methods learn transition potentials or decode full DA sequences with structural constraints. However, they do not address NDAP, where only the *next* act must be predicted and no sequence decoder is used. Critically, none incorporate an *empirical* transition matrix directly into the loss for distribution-level regularization.

Distribution-level constraints have been introduced through posterior regularization, which biases models toward constraint-satisfying priors using KL divergence (Ganchev et al., 2010). Such techniques have improved dialogue understanding and state tracking (Jin et al., 2018), and future-aware constraints have been applied to generation models (Feng et al., 2020). These works demonstrate the utility of KL-based structural guidance, but they do not incorporate DA-transition statistics nor target NDAP.

Recent approaches increasingly rely on LLMs for dialogue generation, typically steered via prompting or instruction tuning. However, growing evidence suggests that LLMs do not reliably acquire human-like discourse behavior, particularly with respect to pragmatic sequencing and role consistency (Shukuri et al., 2023; Wagner and Ultes, 2024; Rudolph et al., 2025). Fluent surface realization therefore provides limited leverage for

structured dialogue control, motivating the use of explicit structural priors.

Our approach fills a gap between these research threads. Unlike NDAP models that rely on implicitly learned transitions, we introduce an explicit, data-derived *transition-matrix regularizer* that aligns the predicted next-act distribution with empirical dialogue-flow patterns. Unlike CRF or HMM models, our method does not require sequence decoding. And unlike posterior-regularization approaches, our structural prior is grounded directly in observed DA transitions. To our knowledge, this is the first instance of integrating empirical DA-transition constraints into the optimization objective for NDAP.

3 Method

3.1 Problem Formulation

Given a conversation history consisting of n utterances $\{u_1, u_2, \ldots, u_n\}$, the goal is to predict the category c_t of the next utterance. All speaker transitions are considered (Counselor→Counselor, Counselor→Client, Client→Client, Client→Counselor), performing NDAP based on the conversation history. Categories belong to a five-level hierarchy with 60 leaf-level dialogue acts.

The task assumes access to gold DA labels in conversation history, appropriate for post-hoc analysis, training simulations, and constrained LLM generation.

To ground the task in a concrete taxonomy, the full hierarchical structure is adopted from the OnCoCo dataset, which organizes 60 leaf-level categories into progressively more abstract semantic groups (Albrecht et al., 2025). The hierarchy captures both conversational function and pragmatic counseling flow: high-level groups distinguish phases such as greetings, problem exploration, motivation building, and closing, while lower levels specify fine-grained communicative functions such as factual problem descriptions, emotional disclosures, resource activation, or evaluation of solution attempts.

OnCoCo provides finer granularity (60 categories vs. 3–15 in alternatives such as Anno-MI (Wu et al., 2022a), HOPE (Malhotra et al., 2022), or MITI (Moyers et al., 2016)) and integrates multiple counselling paradigms rather than a single therapeutic approach, critical for controllable client simulation.

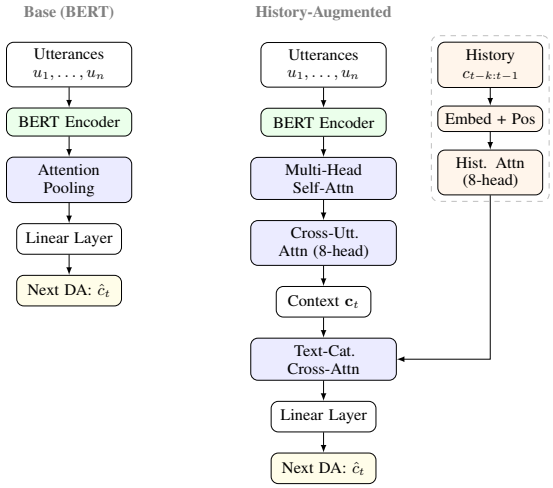

Figure 2: Model architectures for NDAP: BERT + attention pooling **(left)** and a history-augmented variant that incorporates conversational context and previous dialogue-act labels **(right)**.

3.2 Architectures

The base model encodes utterances via a pretrained BERT encoder and aggregates token representations through learned attention pooling. The pooled representation is passed to a linear classifier for NDAP.

The enhanced variant replaces attention pooling with multi-head self-attention for utterance encoding and adds 8-head cross-utterance attention to model conversation flow. A sliding window of the previous h categories is embedded with positional encodings and processed through 8-head self-attention. A text-category cross-attention mechanism integrates the historical context with the conversation representation, and a 4-head category transition attention layer models dependencies between consecutive acts (Figure 2).

3.3 Training

Models combine cross-entropy loss with transition matrix loss regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{tm}} \mathcal{L}_{\text{TM}}. \quad (1)$$

Cross-entropy provides the primary supervision signal based on the single annotated next category, while the transition-matrix loss introduces a soft prior reflecting the empirical multi-modal distribution of plausible next acts.

3.4 Transition Matrix and Dialogue Flow Regularization

A central component in our approach is the *transition matrix loss*, which encourages predictions

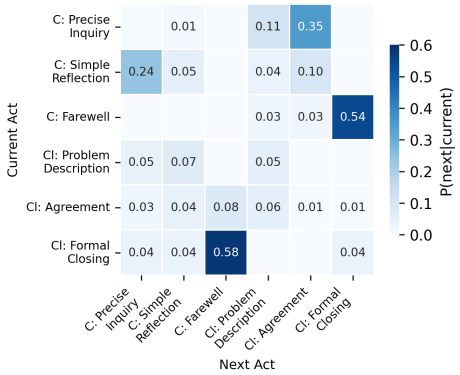

Figure 3: Example 6x6 subset (Fold 0) of the empirical transition matrix. C = Counselor, Cl = Client. High probabilities show pragmatic patterns: closings trigger closings, precise inquiry lead to agreement.

to respect observed category transition patterns in conversation. Although prior dialogue-act models have encoded transitions implicitly through sequential architectures or explicitly through statistical models, to the authors' knowledge no prior work incorporates an empirical transition matrix directly into the training objective of a neural NDAP model.

3.4.1 Computing the Empirical Transition Matrix

A transition matrix encodes the probability distribution over next states given the current state—here, the likelihood of each dialogue act following another. From the training corpus, we compute a normalized empirical transition matrix $\mathbf{T}$ where T_{ij} represents the probability of transitioning from category i to category j. Sparse transitions are smoothed with additive smoothing to avoid zero probabilities. In social counselling conversations, the transition matrix exhibits clear pragmatic patterns as can be seen in figure 3.

3.4.2 Transition Matrix Loss

The transition matrix loss uses KL divergence to measure alignment between model predictions and empirical category transitions. Given the previous category c_{t-1}, the target transition distribution is:

$$\mathbf{t}^{(t)} = \mathbf{T}[c_{t-1}] \in \mathbb{R}^C \tag{2}$$

The model's predicted distribution is:

$$\hat{\mathbf{p}}^{(t)} = \text{softmax}(\hat{\mathbf{y}}_{\text{final}}) \tag{3}$$

The transition matrix loss measures divergence:

$$\mathcal{L}_{\text{TM}} = \text{KL}(\hat{\mathbf{p}}^{(t)} \| \mathbf{t}^{(t)}) = \sum_{j=1}^{C} \hat{p}_j^{(t)} \log \frac{\hat{p}_j^{(t)}}{t_j^{(t)}} \tag{4}$$

KL divergence is asymmetric, penalizing impossible transitions more heavily than missing low-probability valid ones, making it well-suited for enforcing dialogue structure.

Our transition-matrix regularizer differs from label smoothing (Szegedy et al., 2015), which distributes probability mass uniformly across non-target classes. In contrast, our approach distributes mass according to empirically observed transition patterns, providing domain-specific rather than uniform smoothing (Table 4; see Appendix A.2.2 for additional comparisons on the client simulation subset).

The weight $\lambda_{\text{tm}} \in \{0.0, 0.2, 0.5, 1.0, 1.5\}$ is explored systematically (Section 4.4).

4 Experiments

4.1 Dataset

Models are evaluated on a corpus of German social counselling dialogues consisting of 76 conversations with 5,457 utterances. The data originate from structured counselling role-play sessions conducted by social science students in a university course on online text-based counselling. Participants acted in predefined client and counsellor roles, and no real clients or personal information were involved. All conversations were anonymized and collected with consent for research use.

Each utterance is annotated with a category from the OnCoCo taxonomy (Albrecht et al., 2025), which defines 60 leaf-level dialogue act categories organized in a five-level hierarchy. While the taxonomy has been previously described, this work contributes the first publicly available corpus of complete counselling conversations annotated with this scheme, enabling sequential modeling of dialogue flow. Annotations were produced by trained social science students and subsequently reviewed by a domain expert. This sequential review process precludes formal inter-rater agreement but ensures annotation quality through expert oversight.

The human-annotated conversational dataset, along with metadata describing the annotation schema, is released to support reproducibility and further research on hierarchical dialogue-act pre-

diction and counselling dialogue modeling.[1]

For experimentation, NDAP is performed across all speaker transitions, yielding instances across 60 categories. Evaluation uses 5-fold cross-validation with conversation-level partitioning to prevent data leakage. For each fold, training is conducted on 80% of conversations with evaluation on the held-out 20%. Mean performance ± standard deviation across all five test folds is reported. Rather than selecting optimal hyperparameters on a separate validation set (which would further reduce already limited training data), results across the full hyperparameter grid are reported (Section 4.4), allowing readers to assess performance at each regularization strength. The original class distribution exhibits substantial imbalance. LLM-based synthetic data augmentation was also explored to address this imbalance; however, it did not improve BERT-based models (see Appendix A.2.3 for details).

4.2 Baselines and Models

The history-augmented architecture is compared against several baseline approaches: (1) a transition-matrix baseline that uses only empirical dialogue-flow patterns without learning, (2) a Simple RNN baseline, and (3) the architecture proposed by Tanaka et al. (2019), which uses hierarchical attention to integrate multi-turn dialogue history for NDAP.

To validate the robustness of the findings across different pretrained language models, all neural baselines and history-aware models were tested using 7 different German BERT variants: EuroBERT-210m, EuroBERT-610m, G-BERT-base, G-BERT-large, GELECTRa-base, Modern-G-BERT-134M, and Modern-G-BERT-1B. Additionally, context window size is varied (1, 4, 8, 12 utterances) and transition-matrix regularization is compared against label smoothing ($\epsilon \in \{0.0, 0.1, 0.2\}$) as an alternative regularization strategy (see Appendix A.2.2). This systematic evaluation across 300 configurations (7 encoders × 2 architectures × 5 TM weights × 4 context lengths, plus label smoothing variants), evaluated using 5-fold cross-validation, shows that architectural improvements hold consistently across language models ranging from 110M to 1B parameters. Table 1 summarizes the model architectures and their properties.

Transition Matrix Baseline: The transition matrix baseline provides a competitive non-learning baseline. This model predicts the most likely next category given only the previous utterance's category, using the empirical transition matrix computed from training data. It requires no neural training or context encoding and serves as a practical reference point for the value of learned contextual representations.

Model	Attention	History
Transition Matrix	—	—
GPT-5-mini (LLM)	—	✓
gpt-oss-120b (LLM)	—	✓
Simple RNN	Pooling	—
Tanaka (2019)	Hier. Attn.	✓
BERT	Pooling	—
BERT+History	Multi-head	✓

Table 1: Model architectures and properties.

RNN Baselines: Two RNN-based baselines are implemented: (1) Simple RNN, a basic recurrent architecture over utterance embeddings, and (2) the Tanaka et al. (2019) architecture, which uses hierarchical attention mechanisms to model multi-turn context for NDAP. All RNN baselines are trained with the same transition-matrix regularization as BERT models. Results are shown in Section 4.4.1.

LLM Baseline: To contextualize fine-tuned BERT performance against state-of-the-art language models, both proprietary (GPT-5-mini) and open-source (gpt-oss-120b, a 120B-parameter Mixture of Experts model) LLMs are evaluated in a zero-shot setting. Each model receives conversation history (last 12 turns) and all 60 category descriptions, returning top-3 predictions with confidence scores. This baseline tests whether large-scale pretraining alone captures dialogue-flow patterns without explicit structural constraints. The full prompt template is provided in Appendix A.2.7.

4.3 Evaluation Metrics

Models are evaluated on predictive correctness and dialogue-flow alignment via three metric categories:

Predictive Correctness Macro-F1, weighted F1, and Top-3 accuracy serve as the primary measures. Top-3 accuracy acknowledges that multiple next acts can be plausible. Note that these metrics assume a single correct answer, which is a simplification: in NDAP, multiple dialogue acts are often genuinely valid continuations, so moderate absolute scores are expected.

[1] Code and data available at https://anonymous.4open.science/r/tm-reg-8E74

Encoder	Best$_{\lambda=0}$	Best$_{\lambda>0}$	λ	Δ	%Gain
GBERT-large	.097	**.102**	0.5	+.005	+5.1%
GBERT-base	.092	.098	0.5	+.005	+6.0%
ModernGBERT-1B	.089	.096	0.2	+.007	+8.5%
EuroBERT-610M	.087	.096	0.5	+.009	+10.5%
EuroBERT-210M	.080	.095	0.5	+.014	+18.0%
ModernGBERT-134M	.076	.086	0.5	+.009	+12.2%
GELECTRa-base	.060	.070	1.0	+.009	+16.4%
Mean	.083	.092	–	+.008	+11.0%

Table 2: TM regularization effect by encoder (macro-F1, 60 categories, 5-fold CV). Compares best result at λ_{tm}=0 vs best at λ_{tm}>0. Sorted by best macro-F1.

λ_{tm}	Macro-F1	W-F1	Top-3	Cum70	JS
0.0	.309 ±.011	.495 ±.008	**.789** ±.011	.913 ±.019	.299 ±.017
0.2	.314 ±.013	.496 ±.010	.783 ±.014	.915 ±.016	.258 ±.013
0.5	**.319** ±.014	.499 ±.012	.781 ±.013	.916 ±.014	.225 ±.011
1.0	.315 ±.015	.499 ±.012	.777 ±.013	**.918** ±.013	.205 ±.008
1.5	.317 ±.013	**.501** ±.010	.773 ±.011	.916 ±.009	**.201** ±.007

Table 3: Cross-dataset validation on HOPE (15 English counselling dialogue act classes, 5-fold CV, mean ± std).

Dialogue-Flow Alignment To assess how well model predictions adhere to the conversational structure observed in the training data, a group of metrics based on a first-order empirical transition matrix $\mathbf{T}$ is employed: Cumulative Accuracy at 70% (Cum70) and Jensen-Shannon (JS) Divergence. Cum70 measures whether the predicted category falls within the set of most likely transitions that together account for 70% of the empirical probability mass—capturing whether predictions align with pragmatically plausible continuations. JS divergence measures the overall distributional alignment between predicted and empirical transition distributions.

Rationale for First-Order Metrics While dialogue context is inherently rich and multi-turn, the dialogue-flow metrics deliberately rely on a first-order (last-act-to-next-act) transition matrix. This is a practical necessity. Constructing higher-order transition matrices (e.g., second-order, based on the last two acts) would be infeasible given the 60 leaf categories, as it would lead to a combinatorial explosion of states ($60^2 = 3600$ potential source pairs) and result in an extremely sparse and unreliable matrix with this dataset size. The first-order matrix thus serves as a robust and computable proxy for measuring the model's grasp of foundational, short-term dialogue coherence.

Transition Matrix Computation The transition matrix is computed from training data within each CV fold, measuring whether the model internalized valid dialogue flows.

4.4 Results

4.4.1 TM Regularization Effect by Encoder

Table 2 shows the effect of transition-matrix regularization on the 60-category classification task, comparing each encoder's best result at λ_{tm}=0 versus its best result at λ_{tm}>0.

All encoders improve with transition-loss regularization (Table 2). Relative gains range from +5.1% (GBERT-large) to +18.0% (EuroBERT-210M), with a mean improvement of +11.0%. Notably, smaller encoders show larger relative gains, suggesting TM regularization partially compensates for limited model capacity. Optimal λ_{tm} values cluster around 0.5, with only GELECTRa-base benefiting from higher regularization (λ_{tm}=1.0). Architecture effects vary by encoder (Appendix B Table 16). Figure 4 visualizes values averaged across architectures and context sizes for each λ_{tm} value; this averaging reveals larger relative improvements, ranging from 9% (GBERT-large) to 42% (GELECTRa-base), consistent with the abstract.

4.4.2 Cross-Dataset Validation

To validate generalization beyond German counselling, the transition-matrix regularizer is evaluated on the HOPE dataset (Malhotra et al., 2022), an English counselling corpus with 15 dialogue act categories. Table 3 summarizes results using XLM-RoBERTa-base and BERT-base encoders with both BERT and History architectures. Additional evaluation on Switchboard (SWDA), a non-counselling benchmark with highly skewed transition distributions, is provided in Section A.2.4.

Results indicate generalization across languages (German to English), counselling modalities (online text-based to spoken), and category systems (60-class OnCoCo taxonomy to 15-class HOPE scheme). Macro-F1 improves from 0.309 to 0.319 (+3.2% relative) at λ_{tm}=0.5, and JS divergence drops from 0.299 to 0.201 (33% reduction).

4.4.3 Label Smoothing Comparison

Table 4 compares transition-matrix regularization against label smoothing (Szegedy et al., 2015) across all seven encoders. TM regularization con-

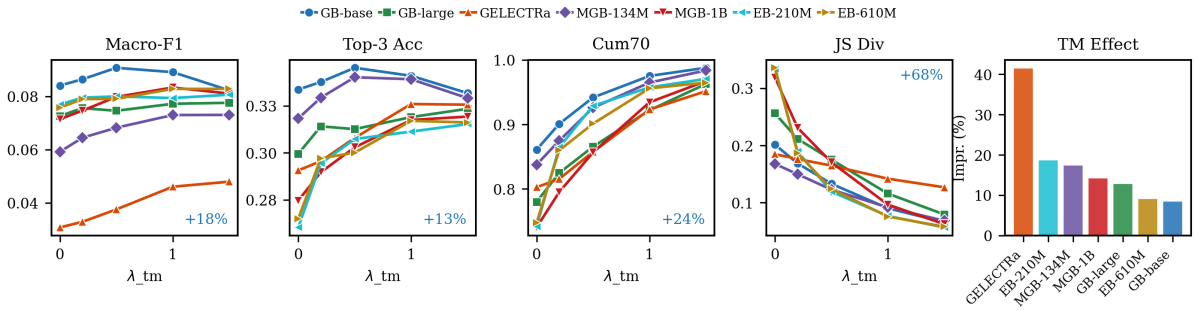

Figure 4: Effect of transition loss weight (λ_{tm}) on 60-category classification. Left four panels show how Macro-F1, Top-3 Accuracy, Cum70, and JS Divergence change with increasing λ_{tm} across encoders. Right panel compares relative macro-F1 gains from TM regularization by encoder.

	Label Smoothing			TM Reg.	
Encoder	ϵ=0.0	ϵ=0.1	ϵ=0.2	Best	Δ
GBERT-large	.079	.074	.069	**.102**	+.023
GBERT-base	.073	.081	.071	**.098**	+.017
ModernGBERT-1B	.059	.074	.065	**.096**	+.022
EuroBERT-610M	.085	.078	.077	**.096**	+.010
EuroBERT-210M	.075	.066	.067	**.095**	+.019
ModernGBERT-134M	.049	.050	.047	**.086**	+.035
GELECTRa-base	.033	.033	.031	**.070**	+.036
Mean	.065	.065	.061	**.092**	+.023

Table 4: Label smoothing vs. TM regularization (macro-F1, 60 categories, 5-fold CV). Δ = TM best − best LS.

Model	λ_{tm}	**Cfg**	**Macro-F1**	**W-F1**	**Top-3**	**Cum70**	**JS**
LLM Baselines (zero-shot, 5-fold CV)							
GPT-5-mini	–	–	.091 ±.011	.132 ±.014	.254 ±.014	.550 ±.035	–
gpt-oss-120b	–	–	.072 ±.006	.108 ±.007	.174 ±.010	.400 ±.021	–
RNN Baselines (5-fold CV)							
Simple RNN	0.5	–	.003 ±.000	.014 ±.002	.208 ±.012	.751 ±.037	.198 ±.006
Tanaka (2019)	1.5	–	.008 ±.002	.031 ±.008	.235 ±.017	.802 ±.035	.209 ±.015
Transition Matrix Baseline (5-fold CV)							
TM Only	–	–	.056 ±.003	.115 ±.011	.315	1.00	–
Fine-tuned Encoders (best config, 5-fold CV)							
GB-large	0.5	H4	**.102 ±.007**	**.171 ±.012**	.342 ±.023	.894 ±.013	.175 ±.004
GB-base	0.5	B8	.098 ±.010	.158 ±.015	**.354 ±.024**	.933 ±.008	.143 ±.006
MGB-1B	0.2	H8	.097 ±.011	.159 ±.021	.314 ±.026	.821 ±.022	.235 ±.015
EB-610M	0.5	H4	.097 ±.012	.160 ±.019	.334 ±.024	.939 ±.007	.109 ±.007
EB-210M	0.5	H12	.095 ±.015	.158 ±.020	.331 ±.027	.937 ±.011	.125 ±.004
MGB-134M	0.5	H4	.086 ±.007	.147 ±.009	.358 ±.021	**.942 ±.005**	.122 ±.001
GELECTRa	1.0	H4	.070 ±.007	.128 ±.011	.335 ±.015	.933 ±.019	**.121 ±.003**

Table 5: Results by model (60 categories, 5-fold CV). Cfg = best architecture (B=BERT, H=History) + context length.

sistently outperforms label smoothing on all encoders, with a mean improvement of +0.023 macro-F1 over the best label smoothing configuration. The relative gains are largest for weaker models (GELECTRa-base: +0.036, ModernGBERT-134M: +0.035), confirming that transition-based priors provide the strongest benefits when baseline performance is low.

4.4.4 Model Comparison and Effect of Regularization Strength

Figure 4 visualizes the effect of transition loss weight (λ_{tm}) across architectures and metrics. Dialogue-flow metrics (Cum70, JS divergence) improve monotonically with increasing λ_{tm}, while predictive metrics (Macro-F1, Top-3) peak around λ_{tm}=1.0–1.5. The right panel shows that TM regularization provides larger effect sizes than architecture changes (BERT→History).

Table 5 compares performance across seven German BERT variants (110M–1B parameters), two LLMs, and baselines. A transition matrix baseline that predicts using only empirical $P(\text{next}|\text{prev})$ from training data achieves macro-F1 of 0.056, outperforming RNN models (0.003–0.008) but falling well short of fine-tuned encoders (0.070–0.102). Among fine-tuned encoders, GBERT-large achieves the highest macro-F1 (0.102), while smaller models like ModernGBERT-134M achieve competitive performance with TM regularization. The Cfg column indicates the best architecture (B=BERT, H=History) and context length for each encoder. History-augmented models dominate, with optimal context lengths of 4–12 utterances. LLMs (GPT-5-mini, gpt-oss-120b) achieve lower dialogue-flow alignment (Cum70: 0.40–0.55) despite competitive macro-F1, indicating that pretraining alone does not capture transition patterns.

Significance Testing. Paired bootstrap tests with Benjamini-Hochberg correction confirm that TM regularization yields robust but configuration-dependent improvements, with mid-sized encoders showing the most consistent gains (Table 14). History-based architectures provide benefits only for specific encoders and do not consistently outperform standard BERT (see Appendix B for encoder-specific results).

5 Discussion

The results highlight several insights:

1. Transition regularization provides consistent benefits: The transition-matrix regularizer improves all metrics across configurations (Table 4). This dual improvement in predictive accuracy and dialogue-flow alignment reflects the multi-modal nature of NDAP: cross-entropy treats all non-gold predictions as equally wrong, while TM regularization provides *distributional* supervision encoding plausible next acts. Moreover, the proposed regularizer is genuinely model- and architecture-agnostic: it yields consistent gains across all seven encoder variants (110M–1B parameters) and both attention-pooling and history-augmented architectures, and can be integrated as a drop-in objective without any architectural changes or sequence-level decoding.

2. Structural priors matter more than model scale: Across a 10× size range (110M–1B parameters), TM regularization provides larger gains than encoder choice (Table 5). ModernGBERT-1B achieves best macro-F1 at λ_{tm}=0.2 but exhibits lower dialogue-flow alignment (Cum70=.822, JS=.236) than smaller models with higher λ_{tm}—larger models may require less aggressive regularization but still benefit from structural priors.

3. Differential benefits across encoders and architectures: Encoder-level analysis (Appendix B) reveals a clear separation. TM regularization acts as a general-purpose inductive bias that disproportionately benefits weaker encoders (GELECTRa-base: +0.036, ModernGBERT-134M: +0.035 macro-F1), whereas explicit history modeling provides selective gains that do not generalize across encoder families (GELECTRa: +1.93 points; GBERT-base: −0.16 points). Across all encoders, domain-specific transition priors consistently outperform uniform label smoothing (Table 4).

4. LLM comparison validates transition regularization: GPT-5-mini achieves lower macro-F1 than the best fine-tuned model Table (5) and substantially lower dialogue-flow alignment (Cum70: 0.550 vs 0.894). The open-source gpt-oss-120b performs worse (macro-F1: 0.072) due to unreliable schema adherence in structured outputs, resulting in a 17% parse error rate. The low Cum70 for both LLMs confirms that even state-of-the-art LLMs do not naturally acquire dialogue-flow transition patterns from pretraining alone, validating the need for explicit structural priors.

5. Client simulation subset shows strong benefits: On a 28-category client-side subset (Appendix A.2.1), TM regularization yields +18% weighted F1 improvement (0.225→0.265) confirming that the approach is effective for the client simulation use case.

5.1 Effect of Regularization Strength

Response pattern: All metrics improve with transition-loss weight, with dialogue-flow metrics (cumulative coverage, JS divergence) continuing to improve at higher λ_{tm} values while predictive metrics peak around λ_{tm}=1.0–1.5. The increasing trend suggests the regularizer provides a stable optimization signal without creating competing objectives.

Implicit regularization effect: The transition-matrix regularizer also acts as an implicit regularizer against overfitting. Without it (λ_{tm}=0.0), 83% of runs show overfitting magnitude >0.5. Positive λ_{tm} reduces this: at λ_{tm}=1.5, overfitting magnitude falls to 0.20 with no runs exceeding 0.5.

Practical recommendation: For practitioners, a $\lambda_{\text{tm}} \in [0.5, 1.0]$ as a default is recommended. Across all configurations, λ_{tm}=0.5 most frequently achieves optimal macro-F1, while λ_{tm}=1.0 performs within 5% of optimal in over 70% of configurations, making it a robust choice when extensive tuning is not feasible. This pattern also holds for HOPE.

6 Conclusion

This paper proposed a transition-matrix KL regularizer for NDAP that incorporates empirical dialogue-flow structure into the training objective. Across experiments on a fine-grained German counselling taxonomy and cross-dataset transfer to other dialogue corpora, the regularizer consistently improves both predictive performance and alignment with observed dialogue-flow dynamics. Additionally, we presented history-augmented architectures that leverage broader dialogue context through multi-head attention.

A natural next step is integrating NDAP with LLM control, using predicted categories to condition prompts or guide sampling strategies for client simulation. Additionally, fine-tuning decoder-only architectures on NDAP could leverage their autoregressive nature to better model dialogue sequences while potentially benefiting from transition-matrix regularization.

Limitations

The primary dataset consists of 76 conversations with approximately 5,500 utterances—a relatively small corpus for training neural models. While cross-validation and cross-dataset experiments provide evidence of robustness, the limited size restricts conclusions about generalization to larger-scale deployments or substantially different counselling contexts.

The conversations are role-play sessions conducted by social science students, not recordings of real clinical counselling. While participants followed structured scenarios designed to reflect authentic counselling dynamics, student role-plays may lack the pragmatic complexity, emotional depth, and unpredictability of genuine therapeutic interactions. Models trained on this data may not fully capture the nuances present in real-world counselling.

Annotations were produced by trained students and subsequently reviewed by a domain expert in a sequential process. This workflow precludes formal inter-rater agreement metrics, which are standard for validating annotation reliability. Although expert review ensures quality control, the absence of quantified agreement limits assessment of annotation consistency and may affect reproducibility of the label assignments.

The transition matrix loss assumes transitions are stable across the corpus; in applications with significant domain shift or concept drift, alternative regularization approaches may be needed. Our evaluation uses gold dialogue act labels for the conversation history; although this is common practice in NDAP, these markings would have to be predicted in fully autonomous operations, which could potentially lead to chain errors. End-to-end evaluation combining NDAP with LLM-based response generation remains unexplored.

Ethical considerations

The counselling dialogue data used in this study were collected from structured role-play sessions conducted by social science students in a university course setting. Participants acted in predefined client and counsellor roles; no real clients, patients, or personal therapeutic information were involved. All participants provided informed consent for research use of the data, and all conversations were anonymized prior to analysis. The dialogues were annotated by paid social science students with prior training in counselling concepts. Annotation included both dialogue act labeling and anonymization of the conversations. All annotations were subsequently reviewed by a domain expert to ensure consistency and quality.

We acknowledge that dialogue act prediction technology for counselling contexts could potentially be misused. However, the intended application, controllable client simulation for counselor training, serves an educational purpose that may improve the quality of counseling services. The data set released contains only simulated conversations and poses minimal privacy risk.

Acknowledgments

Generative AI tools (ChatGPT and Claude) were used throughout the article for coding assistance, to identify relevant literature, LaTeX figure and table preparation, and to improve clarity and style of writing. They were not used to generate scientific claims or experimental results. All content was reviewed, verified, and finalized by the authors.

References

Jens Albrecht, Robert Lehmann, Aleksandra Poltermann, Eric Rudolph, Philipp Steigerwald, and Mara Stieler. 2025. OnCoCo 1.0: A Public Dataset for Fine-Grained Message Classification in Online Counseling Conversations. *arXiv preprint*. ArXiv:2512.09804 [cs].

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Timothy W. Bickmore, Daniel Schulman, and Candace Sidner. 2013. Automated interventions for multiple health behaviors using conversational agents. *Patient Education and Counseling*, 92(2):142–148.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sirui Chen, Yuan Wang, Zijing Wen, Zhiyu Li, Changshuo Zhang, Xiao Zhang, Quan Lin, Cheng Zhu, and Jun Xu. 2023. Controllable Multi-Objective Re-ranking with Policy Hypernetworks. In *Proceedings*

of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 3855–3864. ArXiv:2306.05118 [cs].

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via CRF-attentive structured network. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, Sigir '18, pages 225–234, Ann Arbor, MI, USA. Association for Computing Machinery. Number of pages: 10 tex.address: New York, NY, USA.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding Attention in Sequence-to-Sequence Models for Dialogue Act Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7594–7601.

Orianna Demasi, Yu Li, and Zhou Yu. 2020. A Multi-Persona Chatbot for Hotline Counselor Training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3623–3636, Online. Association for Computational Linguistics.

Shaoxiong Feng, Xuancheng Ren, Hongshen Chen, Bin Sun, Kan Li, and Xu Sun. 2020. Regularizing Dialogue Generation by Imitating Implicit Scenarios. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6592–6604, Online. Association for Computational Linguistics.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11(67):2001–2049.

Ananya Ganesh, Martha Palmer, and Katharina Kann. 2021. What Would a Teacher Do? Predicting Future Talk Moves. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4739–4751, Online. Association for Computational Linguistics.

Jeroen Geertzen. 2009. Dialogue act prediction using stochastic context-free grammar induction. In *Proceedings of the EACL 2009 workshop on computational linguistic aspects of grammatical inference*, pages 7–15.

Ruifeng He, Rui Zhang, Xiang Li, and Yu Su. 2022. GALAXY: a generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the 60th annual meeting of the association for computational linguistics*, pages 3106–3118.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 332–342.

Xisen Jin, Wenqiang Lei, Zhaochun Ren, Hongshen Chen, Shangsong Liang, Yihong Zhao, and Dawei Yin. 2018. Explicit State Tracking with Semi-Supervision for Neural Dialogue Generation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1403–1412. ArXiv:1808.10596 [cs].

Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, and Laura E. Barnes. 2017. HDLTex: Hierarchical Deep Learning for Text Classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 364–371. ArXiv:1709.08267 [cs].

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, Wsdm '22, pages 735–745, Virtual Event, AZ, USA. Association for Computing Machinery. Number of pages: 11 tex.address: New York, NY, USA.

Theresa B. Moyers, Lauren N. Rowell, Jennifer K. Manuel, Denise Ernst, and Jon M. Houck. 2016. The Motivational Interviewing Treatment Integrity Code (MITI 4): Rationale, Preliminary Reliability and Validity. *Journal of Substance Abuse Treatment*, 65:36–42.

Masaaki Nagata and Tsuyoshi Morimoto. 1994. First steps towards statistical modeling of dialogue to predict the speech act type of the next utterance. *Speech Communication*, 15(3):193–203.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in neural information processing systems*, volume 35, pages 27730–27744. Curran Associates, Inc. Citation Key: Ouyang_rlhf_2022.

Lohith Ravuru, Seonghan Ryu, Hyungtak Choi, Haehun Yang, and Hyeonmok Ko. 2022. Multi-Domain Dialogue State Tracking By Neural-Retrieval Augmentation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 169–175, Online only. Association for Computational Linguistics.

N. Reithinger, R. Engel, M. Kipp, and M. Klesen. 1996. Predicting dialogue acts for a speech-to-speech translation system. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 2, pages 654–657 vol.2.

Eric Rudolph, Philipp Steigerwald, and Jens Albrecht. 2025. Comparing human roleplayers and LLM-simulated clients in online counselling training: An analysis of counselling patterns. In *Proceedings of the 18th international conference on educational data mining*, pages 381–387, Palermo, Italy. International Educational Data Mining Society.

Guokan Shang, Antoine Tixier, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2020. Speaker-change Aware CRF for Dialogue Act Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 450–464, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Elizabeth Shriberg, Rebecca Bates, Paul Taylor, Andreas Stolcke, Daniel Jurafsky, Klaus Ries, Noah Coccaro, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3–4):439–487.

Kotaro Shukuri, Ryoma Ishigaki, Jundai Suzuki, Tsubasa Naganuma, Takuma Fujimoto, Daisuke Kawakubo, Masaki Shuzo, and Eisaku Maeda. 2023. Meta-control of Dialogue Systems Using Large Language Models. *arXiv preprint*. ArXiv:2312.13715 [cs].

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374. Place: Cambridge, MA Publisher: MIT Press.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. *arXiv preprint*. ArXiv:1512.00567 [cs].

Koji Tanaka, Junya Takayama, and Yuki Arase. 2019. Dialogue-act prediction of future responses based on conversation history. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop*, pages 197–202.

Stefan Ultes, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Dongho Kim, Iñigo Casanueva, Paweł Budzianowski, Nikola Mrkšić, Tsung-Hsien Wen, Milica Gašić, and Steve Young. 2017. PyDial: A Multi-domain Statistical Dialogue System Toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78, Vancouver, Canada. Association for Computational Linguistics.

Nicolas Wagner and Stefan Ultes. 2024. On the Controllability of Large Language Models for Dialogue Interaction. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–221, Kyoto, Japan. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *International conference on learning representations (ICLR) 2022*.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022a. Anno-MI: A Dataset of Expert-Annotated Counselling Dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181, Singapore, Singapore. IEEE.

Zixiu Wu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2022b. Towards Automated Counselling Decision-Making: Remarks on Therapist Action Forecasting on the AnnoMI Dataset. In *Interspeech 2022*, pages 1906–1910. ISCA.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Matthias Zimmermann. 2009. Joint segmentation and classification of dialog acts using conditional random fields. In *Interspeech 2009*, pages 864–867. ISCA.

A Model Architecture Details

A.1 Hyperparameter Settings

A.2 Ablation Study

Ablation studies are conducted to validate the design choices: (1) evaluating on a client simulation subset, (2) comparing transition-matrix regularization against label smoothing, (3) evaluating the effect of synthetic data augmentation, (4) exploring alternative KL formulations, and (5) investigating hierarchical label embeddings.

A.2.1 Client Simulation Subset

For client simulation applications, predicting client-side dialogue acts is the primary goal. TM regularization is evaluated on a subset containing only client-side categories, filtering for counselor→client and client→client transitions.

Dataset. The client simulation subset contains 2,176 instances across 28 client-side categories. A train/test split of 1,604/572 utterances with conversation-level partitioning is used. The following ablation studies (Sections A.2.2–A.2.3) are conducted on this subset to provide additional insights into design choices.

Hyperparameter	Value
Training Configuration	
Epochs	10
Batch size	64
Learning rate scheduler	Cosine
Gradient clipping	1.0
Mixed precision (AMP)	Enabled
Early Stopping	
Patience	3 epochs
Min delta	0.001
Monitor metric	Validation macro-F1
Model Architecture	
Context utterances	{1, 4, 8, 12}
History length	10 categories
Attention heads	8
Transformer dropout	0.2
Regularization Grid	
λ_{tm}	{0.0, 0.2, 0.5, 1.0, 1.5}
Label smoothing ϵ	{0.0, 0.1, 0.2}
Cross-Validation	
Number of folds	5
Split strategy	Conversation-level
Random seed	42

Table 6: Hyperparameter settings. Context utterances and regularization weights were varied in grid search; other values were fixed across all experiments.

Effect of Transition Regularization. Table 7 shows the effect of varying λ_{tm} on the client simulation subset.

λ_{tm}	W-F1	Top-3	Trans.	Cum70
0.0	.225	.451	.599	.801
0.2	.255	.495	.743	.894
0.5	**.265**	**.505**	.799	.925
1.0	.251	.497	.838	.955
1.5	.253	.494	**.875**	**.961**

Table 7: Effect of λ_{tm} on client simulation subset (28 categories). Trans.=transition validity. Cum70=cumulative accuracy at 70%.

TM regularization yields +18% weighted F1 improvement (0.225→0.265 at λ_{tm}=0.5), stronger than the +17% improvement on the full 60-category task. Transition validity improves by +46% (0.599→0.875 at λ_{tm}=1.5).

Encoder Comparison. Table 8 compares encoder performance on the client simulation subset.

Best weighted F1 (0.292) is achieved with GBERT-large at λ_{tm}=0.2. Notably, the absolute F1 scores on the 28-category subset are substantially higher than on the 60-category task (0.292 vs 0.164), reflecting the reduced complexity of the classification problem. However, the relative improvement from TM regularization remains consistent, confirming that dialogue-flow priors are valuable across task granularities.

Encoder	W-F1	Top-3	Params
GBERT-large	**.292**	.506	336M
ModernGBERT-1B	.287	.507	1B
GBERT-base	.282	**.518**	110M
EuroBERT-610M	.281	.492	610M
EuroBERT-210M	.274	.513	210M
ModernGBERT-134M	.271	.538	134M
GELECTRa-base	.252	.517	110M

Table 8: Encoder comparison on client simulation subset (28 categories, best λ_{tm} per encoder). Model size (110M–1B) shows no consistent correlation with performance.

A.2.2 Transition-Matrix vs. Label Smoothing

Table 9 compares our transition-matrix regularization against label smoothing (Szegedy et al., 2015) with $\epsilon \in \{0.0, 0.1, 0.2\}$ on the client simulation subset. While label smoothing provides modest gains for some architectures, transition-matrix regularization consistently outperforms across most encoder variants.

	Label Smoothing			TM Reg.	
Encoder	ϵ=0.0	ϵ=0.1	ϵ=0.2	Best	Δ
GBERT-large	.244	.268	.255	**.292**	+.024
GBERT-base	.253	.263	.260	**.282**	+.019
ModernGBERT-134M	.256	.251	**.277**	.271	–.006
GELECTRa-base	.206	.197	.208	**.253**	+.044
EuroBERT-210M	.243	.223	.252	**.275**	+.023
EuroBERT-610M	.229	.225	.241	**.281**	+.041
Mean	.239	.238	.249	**.276**	+.024

Table 9: Label smoothing vs. TM regularization (weighted F1). TM outperforms on 5 of 6 encoders.

The mean improvement of TM regularization over the best label smoothing configuration (ϵ=0.2) is +0.024 F1 points. Notably, the gains are largest for architectures with lower baseline performance (GElectra-base: +4.4%, EuroBERT-610M: +4.1%), suggesting that domain-specific structural priors are particularly valuable when model capacity is limited. Table 4 in the main paper shows that this pattern holds even more strongly on the full 60-category task.

A.2.3 Synthetic Data Augmentation

Given the substantial class imbalance in the client simulation subset (Gini coefficient 0.51), LLM-based synthetic data augmentation was explored.

A two-phase strategy was employed: (1) prompt-based augmentation using GPT-5-mini to balance minority classes up to 100 examples each, and (2) persona-based generation using expert-designed client profiles to introduce lexical and stylistic diversity. The augmented training set contains 8,487 instances (81% synthetic), reducing the Gini coefficient from 0.51 to 0.06.

Table 10 compares BERT-based models trained on real data only versus real+synthetic data on the client simulation subset.

Encoder	Real	+Synth.	Δ
GBERT-large	**.292**	.281	–.011
GBERT-base	**.282**	.264	–.018
ModernGBERT-134M	.271	**.283**	+.011
GELECTRa-base	.253	**.257**	+.004
EuroBERT-210M	.275	**.283**	+.008
EuroBERT-610M	**.281**	.276	–.006
Mean	**.276**	.274	–.002

Table 10: Real vs. real+synthetic training (weighted F1). Synthetic augmentation provides no consistent benefit for BERT-based models.

Results are mixed: synthetic data slightly improves smaller models (ModernGBERT-134M, EuroBERT-210M) but degrades larger models (GBERT-large, GBERT-base). The overall mean shows no significant difference ($p > 0.17$, independent t-test), suggesting that pretrained transformers are sufficiently data-efficient for this task.

In contrast, RNN baselines benefited substantially from synthetic augmentation: Simple RNN achieved F1=0.234 vs 0.097 on real data only. This suggests that pretrained transformers are sufficiently data-efficient for fine-grained dialogue act prediction, while RNNs require substantially more examples to converge. For this task, domain-specific structural priors (like transition matrices) appear more valuable than quantity-based augmentation when using pretrained encoders.

A.2.4 Cross-Dataset Validation on SWDA and Alternative KL Formulations

To test generalization beyond counselling and investigate the effect of skewed transition distributions, we evaluated TM regularization on the Switchboard Dialogue Act Corpus (SWDA; Shriberg et al., 1998; Stolcke et al., 2000), a benchmark of English telephone conversations filtered to 37 classes (excluding rare categories with fewer than 50 occurrences). Table 11 shows results using XLM-RoBERTa.

λ_{tm}	F1	Top-3	Cum70	JS
0.0	**.169**	.765	.901	.150
0.5	.162	.767	.929	.073
1.5	.144	**.771**	**.986**	**.023**

Table 11: SWDA 37-class results. Dialogue-flow metrics (Cum70, JS) improve substantially but macro-F1 decreases with TM regularization.

Unlike counselling datasets (OnCoCo, HOPE), TM regularization on SWDA improves dialogue-flow alignment (JS: 0.150→0.023, Cum70: 0.901→0.986) but *decreases* macro-F1. This divergence reflects SWDA's highly skewed transition distribution: analysis of the empirical transition matrix reveals that 69% of source classes have a single dominant successor—the "Statement-non-opinion" (sd) category, which accounts for 34% of all transitions. The regularizer thus pushes predictions toward this dominant class, improving flow alignment but suppressing minority class predictions.

To address this skewness, alternative formulations of the transition-matrix loss were explored: (1) reverse KL divergence, which is more permissive for confident predictions, and (2) entropy-weighted KL, which down-weights samples from high-entropy source categories where multiple successors are plausible. Neither variant improved over standard forward KL (Table 12), suggesting that the macro-F1 degradation on SWDA stems from the dataset's inherent transition structure rather than the KL formulation.

KL Variant	Accuracy	Macro-F1
Forward KL (baseline)	**0.502**	**0.150**
Reverse KL	0.500	0.149
Entropy-weighted	0.496	0.146
Entropy + Reverse	0.495	0.146

Table 12: Alternative KL formulations on SWDA 37-class (λ_{tm}=0.5). Forward KL performs best.

This finding suggests that TM regularization is most effective when transition patterns are more balanced, as in counselling domains where dialogue follows structured phases (problem exploration → intervention → resolution) rather than converging on a single dominant act type.

A.2.5 Context Window Size

The number of preceding utterances used as context is varied (1, 4, 8, 12). Table 13 shows macro-F1

across context lengths and TM weights. Performance peaks at 4 utterances with $\lambda \geq 0.5$ (macro-F1=0.080). Longer context windows provide diminishing returns, suggesting that recent dialogue history is most informative for NDAP.

	TM Weight (λ_{tm})				
Ctx	0.0	0.2	0.5	1.0	1.5
1	.065 ±.012	.069 ±.012	.072 ±.010	.071 ±.011	.073 ±.009
4	.070 ±.019	.073 ±.020	**.080 ±.013**	**.080 ±.014**	**.080 ±.011**
8	.066 ±.018	.076 ±.015	.074 ±.020	.079 ±.013	.079 ±.010
12	.065 ±.013	.070 ±.019	.075 ±.018	.077 ±.014	.077 ±.012

Table 13: Macro-F1 by context window size and TM weight (mean ± std, 5-fold CV). Best at 4 utterances with $\lambda \geq 0.5$.

A.2.6 Hierarchical Label Embeddings

The OnCoCo taxonomy organizes 60 leaf categories into a five-level hierarchy. Experiments explored whether explicitly encoding this structure could improve predictions by embedding each hierarchy level (K_1–K_5) independently and integrating them with the conversation context via cross-attention. This approach is inspired by hierarchical text classification methods (Kowsari et al., 2017). However, experiments showed only marginal improvements over the base architecture (+0.5% weighted F1), insufficient to justify the additional complexity. The hypothesis is that the pretrained encoder already captures sufficient semantic structure, and that the transition-matrix regularizer—which implicitly encodes label relationships through observed co-occurrence patterns—provides a more effective inductive bias than explicit hierarchy embeddings.

A.2.7 LLM Baseline Prompt Template

The following prompt template is used for GPT-5-mini zero-shot evaluation. Category descriptions and conversation history are inserted at the indicated placeholders.

```
You are an expert in dialogue act classification
for German online counseling.

## Task
Predict the dialogue act category of the NEXT
utterance in this counseling conversation.

## Categories (60 total)
[CATEGORY_CODE]: [DESCRIPTION]
... (all 60 categories with descriptions)

## Conversation History (last 12 turns)
[SPEAKER] ([CATEGORY_CODE] | [DESCRIPTION]): [TEXT]
... (conversation turns with speaker, category, text)

## Output Format (JSON)
Return your top 3 predictions:
{
  "predictions": [
    {"category": "CODE", "confidence": 0.6},
    {"category": "CODE", "confidence": 0.25},
    {"category": "CODE", "confidence": 0.15}
  ]
}
```

B Significance Tests by Encoder

We report paired bootstrap significance tests (10,000 iterations) with Benjamini-Hochberg FDR correction for multiple comparisons. Table 14 summarizes results across regularization strengths, demonstrating consistent positive effects at all λ_{tm} values tested. Tables 15 and 16 show encoder-specific results.

λ_{tm}	Tests	Pos.	%Pos.	Sig.	%Sig.	Mean Δ	Med. Δ
0.2	42	36	86%	9	21%	+0.0061	+0.0050
0.5	44	37	84%	25	57%	+0.0069	+0.0063
1.0	45	40	89%	22	49%	+0.0089	+0.0091
1.5	42	33	79%	19	45%	+0.0075	+0.0070

Table 14: Summary of TM regularization significance tests by λ_{tm} value. Tests: paired comparisons (encoder × architecture × context). Pos.: positive effect over λ_{tm}=0. Sig.: significant after FDR correction (α=0.05).

Encoder	Δ Macro-F1	Sig. (FDR<0.05)
ModernGBERT-134M	+0.93%	5/8 (62%)
EuroBERT-610M	+0.80%	4/5 (80%)
ModernGBERT-1B	+0.72%	3/7 (43%)
GBERT-base	+0.70%	5/6 (83%)
GELECTRa-base	+0.68%	4/8 (50%)
GBERT-large	+0.47%	3/7 (43%)
EuroBERT-210M	+0.38%	1/3 (33%)
Overall	**+0.69%**	**25/44 (57%)**

Table 15: TM effect (λ_{tm}=0 vs 0.5) by encoder. Mid-sized encoders show highest significance rates.

Encoder	Δ Macro-F1	Sig. (FDR<0.05)
GELECTRa-base	+1.93%	4/4 (100%)
ModernGBERT-134M	+0.93%	2/4 (50%)
EuroBERT-210M	+0.75%	2/3 (67%)
EuroBERT-610M	+0.64%	2/4 (50%)
ModernGBERT-1B	+0.50%	1/4 (25%)
GBERT-large	−0.08%	1/3 (33%)
GBERT-base	−0.16%	1/4 (25%)
Overall	**+0.67%**	**13/26 (50%)**

Table 16: Architecture effect (BERT→History, λ_{tm}=0.5) by encoder. Benefits weaker encoders only.