BEYOND MODE COLLAPSE: GENERATING DIVERSE SYNTHETIC EXPERT CONVERSATIONS VIA GENERATIVE FLOW NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

High quality synthetic data is central to post-training LLMs for adaptive-AI applications that represent the diverse expert strategies and decisions in conversations. Prompting LLMs directly or conditioning them on end-use scenarios yields low-diversity data that collapses onto dominant modes. We propose a method to generate diverse high-quality synthetic data using Generative Flow Networks (GFlowNets). We show that training GFlowNets to generate latent conversation structure using a Gaussian-mixture density over key interaction features (e.g., confusion-episode dynamics, scaffolding–directive balance) enables sampling expert strategies in proportion to their prevalence in the training data. Across two structurally distinct domains, tutoring and emotional-support dialogues, our GFlow-based synthetic data generation approach offers a better balance of fidelity, mode-coverage and authenticity than reinforcement-learning and end-to-end LLM baselines, without copying training data. Evaluated on three downstream outcome prediction tasks, classifiers trained on synthetic GFlowNet-generated conversations provide a stronger training signal than competitive synthesis baselines.

1 INTRODUCTION

Building adaptive AI applications to support tasks involving experts (e.g., answering student questions, providing emotional support) requires understanding the decisions that experts make as conversations unfold. These decisions might include taking actions for probing knowledge, explaining, redirecting, and resolving (Chi et al., 2014). As LLMs are increasingly deployed in these domains, including autonomous agentic workflows (Kwan et al., 2024; Sun et al., 2025), reproducing expert decisions faithfully for post-training and evaluation becomes essential. Yet real expert conversations are typically difficult and expensive to obtain; they may also be privacy-sensitive and ultimately cannot be collected at the scale these systems require (de Kok et al., 2023; Nikolenko et al., 2019).

Synthetic conversations offer a scalable alternative for post-training LLMs (Kurakin et al., 2023). While current LLM-based methods can reproduce plausible, fluent surface language of expert dialogues, they struggle to faithfully reproduce the deeper structure underlying diverse, adaptive decisions that drive a long goal-directed conversation (Laban et al., 2026). Prompting with example scenarios or seed data can add surface diversity, such as varied topics (Soudani et al., 2026), but the underlying decision patterns can still collapse onto the dominant modes of interaction in the data (O'Mahony et al., 2024; Hamilton, 2024). Attempted to steer data generation with a handcrafted proxy quality reward can optimize synthesis toward the dominant highest-scoring mode, suppressing the very diversity that downstream post-training (e.g., RLHF) relies on.

Recent advances in measuring structural and cognitive properties of synthetic conversations (e.g., expert decisions, challenges) (Ion & Collins-Thompson, 2026) and factored conversation generation (Macina et al., 2023) have opened new opportunities to control the latent structure of a conversation directly, and to measure whether the resulting synthetic conversations exhibit the intended expert decisions and cognitive dynamics (Ion & Collins-Thompson, 2026). In particular, Generative Flow Networks (GFlowNets) (Bengio et al., 2021) provide a controllable way to achieve mode coverage by sampling discrete sequences in proportion to their rewards. In this work, we leverage the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

capabilities of GFlowNets to train and sample diverse conversation sequences in latent space using a mode-specific Gaussian Mixture Model (GMM). We pose the following research questions.

RQ1: Do the capabilities of GFlow networks flexibly address the needs for diverse synthetic data generation?

RQ2: How does data generated using GFlowNets compare against baselines on fidelity, coverage and authenticity?

RQ3: Does GFlowNet-generated data improve downstream task performance in structured expert conversations?

We evaluate our GFlow-based synthetic data generation against competitive Reinforcement Learning (RL) and LLM-only generation baselines in two expert assistance domains: (1) math-tutoring and (2) emotional support dialogues. We use multidimensional measures of synthetic data quality that include distributional fidelity, mode coverage, and authenticity with respect to real conversations. We show that GFlow-based synthetic data generations yield higher fidelity expert conversations, while maintaining coverage of diverse expert strategies, without plagiarizing real training conversations. We also evaluate the utility of using synthetic data for downstream outcome prediction tasks. We find that the conversation diversity achieved from using GFlowNet-generated training data provides a stronger training signal than baseline synthetic data approaches.

2 Background

Mode collapse in pretrained LLMs is the phenomenon of generating data with limited linguistic and content diversity. Post-training on diverse synthetic data is one way to push a model beyond this distribution (O'Mahony et al., 2024), but only if the synthetic data is itself diverse. Extensive work on multi-turn conversations shows that LLMs degrade as conversations lengthen (Kwan et al., 2024). However, such evaluations have largely focused on capabilities and turn-level quality: benchmarks probe recollection, refinement, and follow-up and find, while synthetic-text toolkits standardize fluency, utility, privacy, and distributional metrics such as FID and MAUVE (Ramesh et al., 2025).

Surface-level interventions such as temperature scaling and scenario prompting can increase variation but tradeoff other quality aspects and limited by LLM's and practitioner's abilities (Nguyen et al., 2025; Chang et al., 2024). Real expert conversations are available only in small quantities and cannot be collected at the scale post-training requires. Factoring conversations in a latent move space and then realizing the moves into dialogues provides a useful alternative to control conversation generation (Macina et al., 2023; Soudani et al., 2026). Generative Flow Networks (Bengio et al., 2021) were introduced to learn a policy that samples discrete, compositional objects with probability proportional to a terminal reward. They provide a promising alternative to reward-maximizing approaches to represent and generate conversations with diverse expert trajectories.

The utility of synthetic data for post-training concentrates in the rare, difficult, and safety-relevant regions, like escalating disclosures, adversarial or multi-intent requests, and extended unresolved confusion that likelihood-driven generation samples least visit (Tan & Lee, 2025). Recovering this coverage by hand-enumerating persona or scenario taxonomies (Zamfirescu-Pereira et al., 2023) bounds diversity by the designer's anticipation and does not scale to the combinatorial, sequentially dependent space that scenario factors induce, leaving the tail effectively unreachable (Novak et al., 2026). This motivates sampling compositional structure with non-negligible probability on low-density regions, rather than enumeration or unmodified base-model sampling.

3 Method

We generate synthetic conversations through a two-stage pipeline: (i) a Generative Flow Network (GFlowNet) plans a sequence of speaker roles and abstract discourse moves underlying a conversation (discourse-move skeleton), and (ii) a large language model realizes each move into a natural-language utterance. This separation lets the structural model learn from a compact, annotation-derived reward while leaving surface-form generation to a capable LLM.

We define the synthetic data generation framework (GFlowNet state, reward, and training objective over an abstract move set) (§ 3.1–§ 3.2), then instantiate and evaluate the framework on two domains of expert assistance - math tutoring, and emotional support dialogues (§ 4.1).

3.1 Preliminaries: Generative Flow Networks

A Generative Flow Network (GFlowNet) (Bengio et al., 2021) learns a stochastic policy for constructing a compositional object through a sequence of actions, trained so that the probability of sampling a complete object x is proportional to a non-negative reward $R(x)$, rather than concentrated on the arg-max object as under reward-maximizing reinforcement learning. Construction is formalized as a walk on a directed acyclic graph whose nodes are partially built objects: a forward policy $P_F(\cdot \mid s; \theta)$ proposes the next action from state s, and generation terminates at a fully constructed object $x \in \mathcal{X}$. Among the family of objectives that enforce P_F-induced terminal probability proportional to $R(x)$ (flow matching, detailed balance, sub-trajectory balance) we use *trajectory balance* (Malkin et al., 2022), which introduces a single learned scalar Z_θ approximating the partition function $\sum_x R(x)$ and matches, along each complete trajectory, the forward flow to the reward:

$$\mathcal{L}_{\text{TB}}(x; \theta) = \Big(\log Z_\theta + \sum_t \log P_F(a_t \mid s_{t-1}; \theta) - \log P_B(a_t \mid s_t) - \log R(x)\Big)^2, \tag{1}$$

with P_B a (possibly fixed) backward policy.

3.2 Discourse-Move Trajectories as a Generative Flow Network

Trajectories. We represent the conversation τ between an expert and a user as a sequence of speaker, move-label and utterance triplets (z_t, m_t, u_t) defined as $\tau = ((z_0, m_0, u_0), (z_1, m_1, u_1), ..., (z_t, m_t, u_t))$ where the speaker role $z_t \in \{\text{E}, \text{C}\}$ is either an expert (E) or their counterpart (C). The moves m_t belong to a domain-specific move vocabulary $\mathcal{M}_{\text{E}}$ and $\mathcal{M}_{\text{C}}$. We write $\mathcal{M} = \mathcal{M}_{\text{E}} \cup \mathcal{M}_{\text{C}}$ for the full move set. The utterance u_t is the observed natural language utterance of the speaker.

Factored structural representation. We factor conversation generation into (i) generation of an interaction structure and (ii) natural-language realization of that structure. We define the structural action at turn t as $a_t = (z_t, m_t)$ and represent the structural trajectory of a conversation as

$$\tau_{STRUCT} = (a_1, a_2, \ldots, a_T) = ((z_1, m_1), (z_2, m_2), \ldots, (z_T, m_T)).$$

The structural state at turn t is the trajectory prefix $s_t = (a_1, \ldots, a_t)$, with $s_0 = \varnothing$. Both our GFlowNet and reinforcement-learning conditions operate over this structural state space rather than generating utterances directly.

Policy. For the policy, the speaker z_t is drawn from a fixed, data-derived speaker-transition distribution $P_{\text{role}}(z_t \mid s_{t-1})$ estimated from the corpus. Only the move policy is learned, so the forward policy factors as

$$P_F(a_t \mid s_{t-1}; \theta) = P_{\text{role}}(z_t \mid s_{t-1})\ P_F(m_t \mid s_{t-1}, z_t; \theta), \tag{2}$$

with the move factor $P_F(m_t \mid s_{t-1}, z_t; \theta)$ carrying learnable parameters θ. We use P_F for this forward policy throughout, including in the trajectory-balance objective and reward (Eq. 3).

We parameterize $P_F(m_t \mid s_{t-1}, z_t; \theta)$ with a GRU over the structural history: a recurrent state $\mathbf{h}_t$ summarizes the prefix s_t, updated with the embedded (speaker, move) pair at each step (Appendix D). Conditioning on the recurrent summary keeps the policy compact while capturing longer-range structure than a one-step Markov policy.

Reward. We factor the trajectory reward into a terminal, distributional term and local, transition-level term:

$$\log R(\tau_{\text{STRUCT}}) = \beta \log R_{\text{term}}(\tau_{\text{STRUCT}}) \ + \ \alpha \sum_{t=1}^{T} \log\big(r_t \cdot w_t\big). \tag{3}$$

The terminal term $R_{\text{term}}(\tau_{\text{STRUCT}})$ scores the plausibility of the *whole* structural trajectory under a corpus-level feature map $\phi : \mathcal{T}_{\text{STRUCT}} \to \mathbb{R}^d$ (length statistics, per-move rates, structural proportions, positional statistics, outcome indicators); by default we use a Gaussian plausibility score $R_{\text{term}}(\tau_{\text{STRUCT}}) = \exp\big(-\|(\phi(\tau_{\text{STRUCT}}) - \mu)/\sigma\|_2\big)$ against the expert feature mean/std (μ, σ). The local term $r_t = \max\big(\hat{P}_{\text{data}}(m_t \mid m_{t-1}, z_{t-1}, z_t),\, \varepsilon\big)$ is an empirical bigram transition probability estimated from the corpus (floored at ε), rewarding locally coherent move sequences.

Training. We minimize the trajectory balance loss of Eq. 1 with the reward of Eq. 3. Because a discourse-move skeleton is built by a single, monotonically growing sequence of moves, each state has a unique parent, so the backward policy P_B is degenerate and its term is a constant absorbed into $\log Z_\theta$ We train in two regimes: an *offline* regime that teacher-forces P_F on real training trajectories with $\log R(x)$ precomputed once per trajectory, and an *online* regime that samples fresh trajectories from the current policy, scores them under R, and teacher-forces the same trajectories to obtain a differentiable $\sum_t \log P_F$—closing the gap between the training distribution and the policy's own sampling distribution. We train all models with Adam, gradient-norm clipping, and early stopping on a held-out split.

Sampling and Natural-Language Realization We generate conversations in two stages. **(1) Skeleton sampling.** We first sample a move skeleton: starting from a target length drawn from the corpus turn-count distribution, we roll out P_F turn by turn. At each step, we fix the next role from the speaker-transition model, mask illegal moves, and sample the next move until a closing move or the target length is reached.

(2) Realization. We then pass the sampled skeleton sequence of (role, move) pairs to a large language model . The LLM model returns one utterance per (role, move) pair as a structured object matching the fixed schema, resulting in the full natural language conversation.

4 Evaluations

We now describe the datasets, features, and metrics for evaluating our proposed GFlow-based conversation synthesis approach. Our empirical results comprise results based on both task-independent, multidimensional dataset quality measures and task-specific training effectiveness.

4.1 Datasets and Dataset Features

We use two expert datasets: (1) Math Mentor Database (MMDB) - a dataset of goal-directed math tutoring conversations, where a tutor is addressing a student's misconception with regards to specific math problems (Ion & Collins-Thompson, 2026), and (2) Emotional Support Conversations (ESConv) - a dataset of emotional support conversations where a support seeker interacts with a support provider regarding a specific emotional issue (Liu et al., 2021).

We represent each real or synthetic conversation x as a fixed-length vector $\phi(x) \in \mathbb{R}^d$ of scalar statistics computed from its sequence of (role, move, utterance) turns. We borrow the feature taxonomy from (Ion & Collins-Thompson, 2026), who define features along five dimensions: 1) *surface* features describe stylistic form independent of content (message-length statistics, turn counts), 2) *content* features are per-move occurrence rates over the full move/strategy vocabulary (e.g. scaffolding or direct-guidance rates for tutoring, question or suggestion rates for ESConv), 3) *structure* features aggregate those rates into coarser category shares (e.g. tutor-academic vs. socio-emotional turns, or empathic vs. directive strategies), 4) *dynamics* features capture how content changes over the conversation (e.g. confusion-episode duration, the shift in scaffolding density from the first to the last third, or the position of the first suggestion), and 5) *outcome* features are binary or scalar end-state indicators (e.g. whether the session resolved, whether a breakthrough occurred). Table 1 summarizes the two source datasets; the full per-dataset feature list is in Appendix 5.

4.2 Metrics

Our evaluation measures can be grouped along three axes: 1) *Skeleton-level* metrics (JS_{marg}, JS_{trans}) test whether generated discourse-move trajectories reproduce real move dynamics; 2)

	MathMentorDB	ESConv
Domain	1:1 math tutoring	emotional-support dialogue
Roles	tutor, student	supporter, seeker
Move vocab.	24	9
Train convs.	2,500	1,040
Held-out convs.	1,000	260
Avg. turns/conv.	23.8	29.5

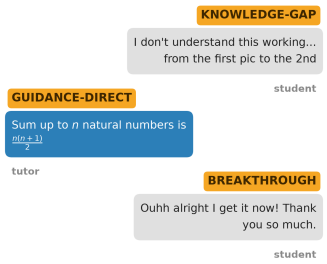


Table 1: **Left:** summary statistics for the two datasets. **Right:** a real excerpt from MathMentorDB's training split showing the tutor and student exchange, and the annotated move labels.

Realized-conversation metrics (pMSE, and the fidelity/coverage decomposition $\mathrm{IP}_\alpha/\mathrm{IR}_\beta$) test whether final transcripts match real conversations in a shared feature space; and 3) *Authenticity* diagnostics test whether that match was novel or achieved by reproducing training records. We summarize the metrics below and refer the readers to Appendix G for specific details on their computation.

JS-Divergence on skeleton moves. We report the Jensen–Shannon divergence over the move distributions $\mathrm{JS}_{\mathrm{trans}}$, averaging $\mathrm{JS}\big(\hat{P}_{\mathrm{gen}}(\cdot \mid c) \,\|\, \hat{P}_{\mathrm{real}}(\cdot \mid c)\big)$ over contexts $c = (z_{t-1}, z_t, m_{t-1})$ to capture how well generated state sequences match real ones (lower is better).

pMSE and mode coverage on realized conversations. To measure conversation *realism*, we report mode *coverage* and *pMSE*, both computed on the per-conversation feature vector $\phi(x)$ of § 4.1. For *mode coverage*, we fit a K-mode Gaussian mixture on the held-out real split (K=7 tutoring, K=8 ESConv) and freeze it for evaluation. Each conversation is assigned a mode $\hat{k}(x) = \arg\max_k p(k \mid \phi(x))$ and use this to report *coverage*, the number of modes with non-negligible mass among generated conversations.

For *pMSE*, we train an L_1-penalized logistic classifier to discriminate real from generated conversations and measure how far its out-of-fold propensity scores deviate from the no-signal baseline, following Ion & Collins-Thompson (2026). We report the normalized $\mathrm{pMSE}_{\mathrm{NORM}} \in [0, 1]$, where 0 means real and generated samples are indistinguishable (perfect fidelity) and larger values mean the classifier separates them (lower is better).

Sample-level fidelity, coverage, and authenticity. pMSE reports whether a discriminator separates real from generated conversations, but not *why*: the separation could be off-support generation, partial coverage of the real distribution, or memorized reuse of training data. We disentangle these with the three dimensions of *fidelity, coverage and authenticity* (Alaa et al., 2022), which are all computed using on the same feature representation $\phi(x)$. *Fidelity* (IP_α) is the fraction of generated conversations inside the ball containing the closest fraction α of real conversations; *coverage* (IR_β) measures the extent to which synthetic data spans the real data distribution; and *authenticity* measures whether synthetic conversations stay as far from their nearest real training conversation as real conversations stay from each other. In the degenerate case, synthetic data can achieve high fidelity by copying training data points; fidelity and authenticity together ensure that generated points are both within distribution and novel. We report $\mathrm{IP}_\alpha, \mathrm{IR}_\beta \in [0, 1]$ (higher is better; 1 is a perfect distributional match), as well as the *outlier* rate (synthetic conversations farther from real data than any real conversation).

4.3 Baselines

We compare against two baselines. 1) **RL+LLM** tests the effect of the GFlow trajectory-balance (TB) loss: we train an architecture- and reward-matched policy with REINFORCE (Williams, 1992) and a mean-reward baseline in place of TB. Comparing GFlow+LLM with RL+LLM under an identical reward isolates whether reward-proportional sampling from GFlowNets prevents the diversity collapse motivated in §3.1. 2) **LLM-only** tests the effect of factored generation: a strong LLM generates complete conversations end-to-end from the same framing and few-shot examples as §3.2,

across three tiers of increasing sophistication – **B1** (naive): uniformly sampled few-shot examples; **B2** (scenario-conditioned): additionally conditioned on a scenario bootstrap-resampled from real conversation metadata; **B3** (mode-proportional): few-shot examples selected in proportion to the same GMM mixture used by the `gmm_density` reward. We report the best-performing tier per domain in the main paper.

4.4 Experimental details.

For each of the three synthetic conditions [1], we generate n=1500 conversations, with GPT-5.6-TERRA and GEMINI-2.5-FLASH each; GFlow+LLM and RL+LLM conditions realize a sampled skeleton (§3.2), while the LLM-only baseline produces structure and text jointly. Sample-level metrics compare equal-size sets, scoring 200 generated against 200 held-out real conversations per draw and reporting mean $\pm$ sd over repeated draws; because each draw uses 200 per side. All distances are Mahalanobis distances in $\phi(x)$, computed using normalized whitening of the feature values on the real training split. On the generated side, move labels are the skeleton the planner committed to (GFlowNet/RL) or the model's own tags (LLM-only), so move-rate features describe the *planned* structure; whether the realized text follows that plan is not tested here. Full sampling, partition, and threshold details are in Appendix F.

5 Evaluation Results for GFlow Synthetic Data Generation

We now discuss results for our research questions on GFlowNet's suitability for diverse synthetic generation (RQ1), fidelity/coverage/authenticity against baselines (RQ2), and downstream task utility (RQ3). We refer to the three synthesis conditions as **GFlow+LLM, RL+LLM, LLM-only**.

	Math tutoring			ESConv		
Condition	$\text{JS}_{\text{trans}}\downarrow$	pMSE $\downarrow$ (all features)	pMSE $\downarrow$ (excl. surface)	$\text{JS}_{\text{trans}}\downarrow$	pMSE $\downarrow$ (all features)	pMSE $\downarrow$ (excl. surface)
Real conversations (null)	0.063	2.0	1.7	0.003	4.6	3.2
GPT-5.6-terra realization						
LLM only (**B2**, best tier)	0.367	97.3	94.3	0.067	98.4	95.6
RL + LLM (REINFORCE)[‡]	0.267	95.3	94.8	0.108	94.2	80.8
GFlow + LLM (ours)	**0.145**	**81.1**	**65.1**	**0.015**	**97.4**	**58.1**
Gemini 2.5 Flash realization						
LLM only (**B2**, best tier)	0.363	99.3	92.8	0.130	94.0	82.7
RL + LLM (REINFORCE)[‡]	0.267	97.1	96.5	0.108	90.2	81.7
GFlow + LLM (ours)	**0.145**	**80.1**	**65.1**	**0.015**	**88.6**	**51.3**
Structural upper bound[†]	0.065	44.5	2.7	0.004	72.6	6.1

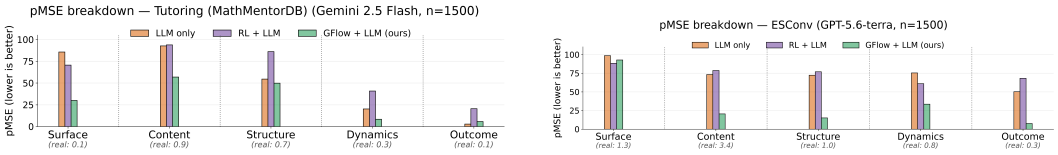
Table 2: GFlow achieved the best synthetic data fidelity across both datasets and LLM choices, as measured by state-transition fidelity (JS_{trans}) and pMSE across the three data generation conditions. A structural upper-bound is provided for reference (see § 4.4 for sample sizes, redraws, and protocol). pMSE is reported on all features of $\phi(x)$ and with the surface feature group excluded, since surface features depend on the LLM realizer rather than the planned structure. Lower is better on all metrics.

5.1 GFlow Suitability (RQ1) - GFlow Improves Synthetic Data Fidelity

Table 2 compares pMSE and JS_{trans} across the three synthesis approaches: the best-performing end-to-end LLM baseline, RL+LLM, and GFlow+LLM. An optimal baseline computed using the real conversations is included for reference. Both factored approaches (RL and GFlow) optimize a Gaussian density reward derived from conversation features. Across both evaluation datasets, the GFlow+LLM pipeline outperforms all baselines on both metrics.

[1] We identify the best performing LLM condition as **B2** in an aprior experiment with 200 samples.

Figure 1 breaks down these gains across five component dimensions (cognitive, content, structure, dynamics, and outcome; see § 4.1). Factored GFlow yields substantial improvements across all five structural dimensions, while surface-level features achieve comparable pMSE across all conditions. This pattern aligns with the design of the generative models: surface features reflect the shared LLM generation step, whereas higher-level features depend on the latent conversation structure governed by the factored model. Details with individual LLM baseline tiers are provided in Appendix E.



(a) Tutoring (b) ESConv

Figure 1: A breakdown of pMSE results by feature dimension (§4.2) shows that GFlow+LLM superior results were achieved by consistent reduction of pMSE across all non-surface feature dimensions, which capture underlying conversation strategies. Surface dimension pMSE was close across all three synthesis approaches. gemini-2.5-flash realization, n=1500, per domain.

5.2 Data Authenticity (RQ2) - GFlowNets balance fidelity and coverage while maintaining authenticity

	pMSE ↓	IP_α ↑	IR_β ↑	Auth. ↑	Outlier
Real held-out (reference)	9.4±4.5	0.94±0.03	0.96±0.02	0.70	0.050
LLM only (B2), GPT-5.6-terra	95.0±3.7	0.44±0.04	0.09±0.03	0.96	0.233
REINFORCE, GPT-5.6-terra	94.1±4.1	0.81±0.03	0.17±0.04	0.92	**0.006**
GFlowNet (TB), GPT-5.6-terra	**71.5**±3.3	**0.87**±0.04	**0.57**±0.07	0.85	0.117
LLM only (B2), Gemini 2.5 Flash	96.0±2.6	0.21±0.03	0.00±0.01	0.99	0.738
REINFORCE, Gemini 2.5 Flash	93.4±2.5	0.85±0.02	0.16±0.04	0.90	**0.007**
GFlowNet (TB), Gemini 2.5 Flash	**70.6**±4.2	**0.89**±0.04	**0.59**±0.07	0.84	0.117

Table 3: Fidelity, diversity, and authenticity on MathMentorDB by synthesis condition. Metric definitions are in § 4.2 and the sampling protocol is in § 4.4. The top row shows scores for held-out real conversations as if they had been generated, so it is the value a perfect generator reaches at this sample size. This provides an upper bound reference against which to compare the main results. ESConv results are in Appendix Table 7.

GFlow+LLM achieves the best balance of coverage and authenticity. To understand whether GFlow+LLM generates more realistic representations of real conversations, we report in Table 3 fidelity (IP_α), coverage (IR_β), and copying statistics (authenticity, outlier) for every condition on MathMentorDB computed on the conversation features $\phi(x)$. For reference, we also report the values a perfect generator attains at this sample size. GFLow+LLM generation achieves competitive fidelity and coverage (IP_α=0.87; IR_β=0.57) improving over the RL+LLM baseline by 0.06 and 0.4 respectively while maintaining moderate outliers and authenticity. RL+LLM achieves a high-fidelity but low coverage, by generating the dominant mode only. The LLM-only baseline achieve low fidelity and coverage but high authenticity due to out-of-distribution examples that are far away from any real training data points.

To better illustrate these behaviors, Figure 2 shows a visual representation of the synthetic data coverage on the real data for the three conditions, along with the separations between the synthetic and real data distributions. LLM-only generations (left) are characterized by scattered coverage of the real data (67%), with many outliers (45%). RL+LLM generations (middle) are characterized by a skewed single mode coverage in the bottom right, but only covering half of the overall data manifold (50%) and hardly any outliers (0.3% outliers). GFLow+LLM generations are characterized by more uniform and better coverage of the data manifold (89%), and spills out only modestly (14% outliers) putting it in between LLM-only and RL+LLM conditions in terms of outliers. This illustrates the tradeoff between generating diverse synthetic data and avoiding outliers.

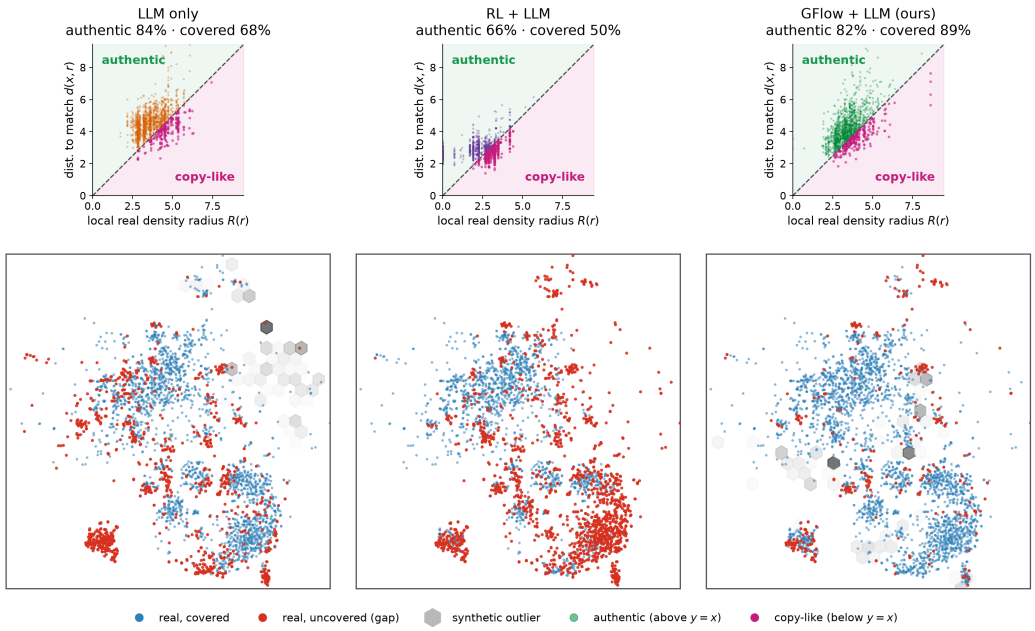


Figure 2: Coverage of the real MathMentorDB distribution for the three studied synthesis approaches, showing how our GFlow + LLM approach (right) is the only one to achieve high fidelity coverage of the entire real support, while also producing novel but realistic outliers. Visualization uses t-SNE of non-surface $\phi(x)$ (based on Gemini, n=1500 per condition). A real point is covered if there is a synthetic point within the 95th percentile (§ Appendix F). Real conversations are blue if covered by the condition and red if not; grey hexagons are synthetic outliers. Plots above the panels describe the authenticity, a local density sensitive measure of how close are synthetic samples to real ones. A synthetic point is authentic if $d(x, r) > R(r)$ (above $y = x$ line), and copy-like otherwise.

5.3 Downstream utility (RQ3) - GFlow-generated diverse conversations provide richer downstream training signals

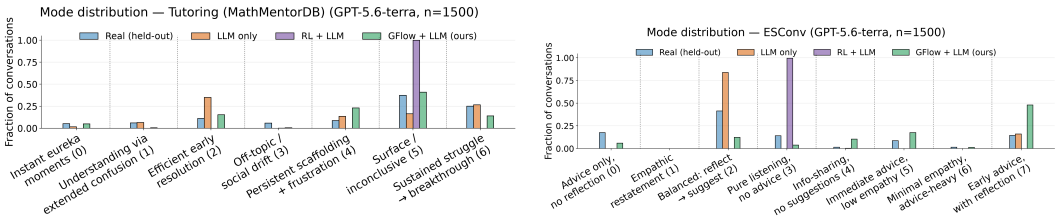
To show that diverse synthetic data can provide better signals for downstream tasks, we evaluate with a train-on-synthetic, test-on-real (TSTR) protocol (Ramesh et al., 2025) on the MMDB dataset. In this setup, we train a per-outcome logistic regression classifier on each turn's sentence embedding fused with its discourse-move label,[2] then evaluate on held-out real conversations. We select three outcomes that characterize salient important aspects of effective tutoring interaction: 1) *resolution* (did the student's confusion get resolved), 2) *breakthrough* (did the student reach a moment of insight, a marker of genuine conceptual gain rather than mere answer delivery), and 3) *off-topic drift* (did the conversation leave the instructional task, a marker of derailment). Since these outcomes can vary across modes, good performance on these tasks depends on having seen the full range of trajectories that experts can take.

		GPT-5.6-terra Realization			Gemini 2.5 Flash Realization			
Label	Real oracle	LLM	RL	GFlow	LLM	RL	GFlow	Majority floor
`resolved`	0.79±0.03	0.50±0.03	0.29±0.01	**0.65±0.03**	0.57±0.03	0.29±0.01	**0.64±0.03**	0.37
`has_breakthrough`	0.92±0.02	0.71±0.03	0.41±0.01	**0.72±0.03**	0.69±0.03	0.41±0.01	**0.81±0.03**	0.41
`has_off_topic`	0.89±0.04	0.48±0.00	0.48±0.00	**0.50±0.03**	0.48±0.00	0.48±0.00	0.48±0.00	0.48

Table 4: $F1_{macro}$ results for three conversation outcome prediction tasks, using the train-on-synthetic test-on-real (TSTR) in the tutoring domain. GFlow + LLM wins or ties every (label, realizer) pair. ± is the half-width of a 95% bootstrap CI (300 draws over real test conversations).

[2]Frozen all-MiniLM-L6-v2 (Reimers & Gurevych, 2019) embeddings, mean-pooled over three positional segments per conversation to retain coarse structure.

Diverse conversations from GFlow+LLM improve training signals. Table 4 reports the results for the outcome prediction task for the three synthesis approaches on both LLM realizations. We also include an oracle trained on the real conversations, and a majority class predictor as references. GFlow+LLM beats the baselines for both LLM realizations, except for `has_off_topic` under Gemini realization. For this outcome, all three conditions degenerate to the majority prediction suggesting no learning signals for the task. For other conditions, LLM-only ranks second, and RL+LLM is much worse for all conditions due to its mode-seeking behavior. Figure 3 shows the mode distribution across the conditions suggesting that the improved predictive power of data from GFlow+LLM arises from its better mode coverage than baselines.



(a) Tutoring (7 modes)

(b) ESConv (8 modes)

Figure 3: GMM mode distribution on the real held-out set for the conditions LLM-only (**B2**, best tier), RL+LLM, and GFlow+LLM (ours) (§ 4.2). Mode labels are our post-hoc interpretation of each cluster's dominant feature profile. RL+LLM is characterized by a single dominant mode, and GFlow+LLM reproduces more modes than LLM-only.

6 Conclusion and Limitations

Our central finding is that by using reward proportional sampling, and appropriate structure based rewards, we can generate diverse, structurally faithful synthetic conversations without relying on copying training data. Moreover, our method is able to navigate effectively between the extremes of verbatim resampling of real conversational structure, which has lowest detectability but high plagiarism rate, and end-to-end LLM generation that produces novel instances but fails to provide adequate coverage of real support. By adapting the GFlowNet framework to the problem of high-quality conversation synthesis we obtain synthetic data that is novel but that retains high fidelity and coverage of key modes.

Our work opens several interesting directions of research in high quality synthetic data generation. First, by controlling the reward structure that evaluates conversation, we can shape generation to a downstream task's specific needs. Rewards can also be informed from expert knowledge. For example, upweighting trajectories with prolonged confusion yields many conversation examples in which a student struggles before reaching understanding. Second, our generative approach is complementary to post-hoc discriminative methods that filter or rejection-sample synthetic data by quality (Seedat et al., 2023). These two approaches act on different synthesis stages and address different failures: generation determines which interactions exist in the corpus (coverage), while filtering determines which of them are kept (per-sample quality). A promising direction is to tune the two jointly: the discriminator's judgments can serve as a reward signal for the generator, while the generator learns to cover the space in ways the discriminator accepts rather than having its output pruned after the fact.

Limitations. Our study covers two domains, both equipped with an established discourse-move taxonomy. Extending the framework to other domains will require creating new discourse moves. The feature map ϕ is hand-designed, and required expert knowledge to create. Future work can investigate automatically learning features if more data is available. Our downstream evaluation trains classifiers on synthetic data as an initial step, but we leave full generative post-training for future work. We also acknowledge that a move-label abstraction is not a full representation of expertise. Exploring what drives these moves and how experts realize them are additional interesting future directions.

Reproducibility. We provide code for reproducibility in the supplementary materials.

REFERENCES

Ahmed M. Alaa, Boris van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 290–306. PMLR, 2022. URL https://proceedings.mlr.press/v162/alaa22a.html.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in neural information processing systems*, 34:27381–27394, 2021.

Allen Chang, Matthew C Fontaine, Serena Booth, Maja J Matarić, and Stefanos Nikolaidis. Quality-diversity generative sampling for learning with synthetic data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19805–19812, 2024.

Michelene TH Chi, Robert Glaser, and Marshall J Farr. *The nature of expertise*. Psychology Press, 2014.

Jip WTM de Kok, Miguel Á Armengol de la Hoz, Ymke de Jong, Véronique Brokke, Paul WG Elbers, Patrick Thoral, Alejandro Castillejo, Tomás Trenor, Jose M Castellano, Alberto E Bronchalo, et al. A guide to sharing open healthcare data under the general data protection regulation. *Scientific data*, 10(1):404, 2023.

Sil Hamilton. Detecting mode collapse in language models via narration. In Antonio Valerio Miceli-Barone, Fazl Barez, Shay B. Cohen, Elena Voita, Ulrich Germann, and Michal Lukasik (eds.), *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pp. 65–72, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.scalellm-1.5. URL https://aclanthology.org/2024.scalellm-1.5/.

Michael Ion and Kevyn Collins-Thompson. Measuring simulation fidelity via statistical detectability: A diagnostic framework for ai-generated tutoring conversations. In *Proceedings of the Thirteenth ACM Conference on Learning @ Scale*, L@S '26, pp. 114–124, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722936. doi: 10.1145/3774398.3811608. URL https://doi.org/10.1145/3774398.3811608.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*, 2023.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. MT-eval: A multi-turn capabilities evaluation benchmark for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20153–20177, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1124. URL https://aclanthology.org/2024.emnlp-main.1124/.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. In *International Conference on Learning Representations*, volume 2026, pp. 54738–54778, 2026.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3469–3483, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.269. URL https://aclanthology.org/2021.acl-long.269/.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5602–5621, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.372. URL https://aclanthology.org/2023.findings-emnlp.372/.

Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in gflownets. *Advances in Neural Information Processing Systems*, 35:5955–5967, 2022.

Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs. In *International Conference on Learning Representations*, volume 2025, pp. 70333–70366, 2025.

Sergey I Nikolenko et al. Synthetic data for deep learning. 2019.

Andrej Novak, Milana Grbić, Matej Ivaniček, and Dragan Matić. Prompting shapes the statistical tails of llm-generated biomedical data. *Statistical Analysis and Data Mining: An ASA Data Science Journal*, 19(3):e70093, 2026.

Laura O'Mahony, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. Attributing mode collapse in the fine-tuning of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, volume 2, pp. 2, 2024.

Krithika Ramesh, Daniel Smolyak, Zihao Zhao, Nupoor Gandhi, Ritu Agarwal, Margrét V. Bjarnadóttir, and Anjalie Field. SynthTextEval: Synthetic text data generation and evaluation for high-stakes domains. In Ivan Habernal, Peter Schulam, and Jörg Tiedemann (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 487–499, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-334-0. doi: 10.18653/v1/2025.emnlp-demos.35. URL https://aclanthology.org/2025.emnlp-demos.35/.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Nabeel Seedat, Nicolas Huynh, Boris Van Breugel, and Mihaela Van Der Schaar. Curated llm: Synergy of llms and data curation for tabular augmentation in low-data regimes. *arXiv preprint arXiv:2312.12112*, 2023.

Heydar Soudani, Roxana Petcu, Evangelos Kanoulas, and Faegheh Hasibi. A survey on recent advances in conversational data generation. *ACM Comput. Surv.*, 58(10), April 2026. ISSN 0360-0300. doi: 10.1145/3795686. URL https://doi.org/10.1145/3795686.

Weiwei Sun, Xuhui Zhou, Weihua Du, Xingyao Wang, Sean Welleck, Graham Neubig, Maarten Sap, and Yiming Yang. Training proactive and personalized llm agents. *arXiv preprint arXiv:2511.02208*, 2025.

Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. Unmasking implicit bias: Evaluating persona-prompted LLM responses in power-disparate social scenarios. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1075–1108, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.50. URL https://aclanthology.org/2025.naacl-long.50/.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581388. URL https://doi.org/10.1145/3544548.3581388.

A LLM-USE

We used LLMs for generating synthetic data because that is the topic of the work. We used LLM-as-a-judge for analyzing the conversation modes and labeling the ESConv features similar to how features were labeled in MathMentorDB (MMDB) dataset. To achieve this, we used Claude to identify expert features in ESConv dataset that are analogous to MathMentorDB dataset such as `question_rate`, `self_disclosure_rate`. After an initial taxomy generation, the features were validated by one of the authors for their correctness and relevance to the emotional support domain in accordance with liteature in the space. Some AI-assistance was part of implementation as we used AI-auto-complete enabled code editors for implementation. We also used LLMs to refine the arguments in the writing. We did not used LLMs to come up with ideas for the work.

B ETHICAL CONSIDERATIONS

We use two existing research datasets, MathMentorDB (Ion & Collins-Thompson, 2026) and ES-Conv (Liu et al., 2021), under their respective terms of use, and collect no new human data. One motivation for synthetic data is to reduce reliance on sensitive real conversations, but generation from real data does not by itself guarantee privacy. For this reason we report copying diagnostics (§5.2), which show near-copy rates at or below those of held-out real data. These diagnostics do not constitute a formal privacy guarantee, and settings that require one should combine our approach with techniques such as differentially private training.

Both domains involve potentially vulnerable users, including students and people seeking emotional support. Synthetic conversations that are realistic by our metrics are not validated as safe or clinically appropriate. Systems trained on them should undergo domain-expert review and appropriate evaluation before deployment, particularly in mental-health settings, where synthetic data is not a substitute for professional oversight.

Because our generator reproduces the structural distribution of its training corpus, it also reproduces that corpus's biases, including which interaction patterns and user populations are represented. Reward reshaping can shift this distribution deliberately, which is useful for covering rare cases but could also be used to generate misleading or manipulative interactions. We will release code and data for research purposes and encourage users to document the reward and data choices behind any synthetic corpus they produce.

C FULL DATASET FEATURES

Table C describes the full feature set used for both datasets.

D GFLOW AND RL POLICY PARAMETERIZATION

We parameterize the move policy $P_F(m_t \mid s_{t-1}, z_t; \theta)$ of Eq. 2 with a GRU over the structural history. A recurrent state $\mathbf{h}_{t-1}$ summarizes the (role, move) pairs seen through turn $t-1$, updated at each step with the embedded pair $[\mathbf{e}(z_{t-1}); \mathbf{e}(m_{t-1})]$. The role z_t itself is not predicted by this network – it is drawn beforehand from the fixed P_{role} of Eq. 2. Given z_t and $\mathbf{h}_{t-1}$, the move is sampled from a masked softmax over z_t's legal move set,

$$P_F(m_t \mid s_{t-1}, z_t; \theta) = \text{MaskedSoftmax}_{\mathcal{M}_{z_t}}\big(W_m[\mathbf{h}_{t-1}; \mathbf{e}(z_t)] + b_m\big). \tag{4}$$

When a Markov policy suffices, $\mathbf{h}_{t-1}$ is replaced by the embedded (role, move) pair from turn $t-1$ alone and a projection of each active state tracker, concatenated through a shallow MLP.

Dimension	MathMentorDB (tutoring), 47 feat.	ESConv (emotional support), 30 feat.
Surface (6 / 6)	avg_student_len, std_student_len, avg_tutor_len, std_tutor_len, n_messages, turn_ratio	avg_seeker_len, std_seeker_len, avg_supporter_len, std_supporter_len, n_messages, turn_ratio
Content (21 / 8)	direct_rate, scaffolding_rate, correction_rate, confirm_pos_rate, confirm_neg_rate, tutor_question_rate, knowledge_check_rate, attempt_rate, knowledge_gap_rate, understanding_rate, breakthrough_rate, explain_problem_rate, explain_reasoning_rate, student_question_rate, student_confirm_rate, frustration_rate, encouragement_rate, empathy_rapport_rate, confidence_express_rate, acknowledgment_rate, knowledge_recall_rate	question_rate, restatement_rate, reflection_rate, self_disclosure_rate, affirmation_rate, suggestion_rate, information_rate, others_rate
Structure (7 / 4)	pct_tutor_academic, pct_student_academic, pct_non_academic, pct_socio_emotional, greeting_rate, platform_rate, smalltalk_rate	pct_empathic, pct_directive, pct_exploratory, pct_personal
Dynamics (10 / 9)	n_confusion_episodes, avg_confusion_duration, max_confusion_duration, instant_resolution_rate, extended_confusion_rate, scaffolding_density_first_third, scaffolding_density_last_third, scaffolding_density_shift, first_attempt_position, first_understanding_position	empathy_density_first_third, empathy_density_last_third, directive_density_first_third, directive_density_last_third, empathy_to_directive_shift, strategy_entropy, n_strategy_types, suggestion_position, first_empathy_position
Outcome (3 / 3)	resolved, has_breakthrough, has_off_topic	resolved_esc, has_suggestion, has_reflection

Table 5: The full feature vector $\phi(x) \in \mathbb{R}^{47}$ (tutoring) / $\mathbb{R}^{30}$ (ESConv), by dimension (§3.2), following the taxonomy of Ion & Collins-Thompson (2026). Per-dimension counts are (tutoring / ESConv). These are the only features used by any evaluation metric or reward in the paper (pMSE, mode coverage, DCR/copying, the GMM-density and Gaussian-z-score reward terms); the state trackers that condition the policy during generation (Appendix D) are a separate, much smaller set and are not part of $\phi(x)$.

E LLM Baseline Condition Details

This appendix expands the baseline and reference conditions used in Tables 2, and **??** (§4.3): the three LLM-baseline tiers, why their sample size differs from the GFlowNet conditions, and a short glossary of the RL+LLM, GFlow+LLM, and structural-upper-bound conditions.

LLM-baseline tiers. The end-to-end LLM baseline of §4.3 is not one configuration but three escalating tiers, each closing one specific objection to treating "prompt an LLM directly" as a weak strawman.

- **B1 (naive).** Task specification plus k=3 few-shot demonstrations drawn uniformly at random from the training corpus; no scenario conditioning.
- **B2 (scenario).** B0 plus conditional prompting: a (attribute, value) scenario descriptor—resolution outcome and target turn count for tutoring; emotional register and problem type and target turn count for ESConv—bootstrap-resampled from an actual training conversation's metadata, rather than an invented category list or a uniform prior over labels.
- **B3 (scenario+mode).** B1 plus mode-proportional few-shot selection: the k=3 demonstrations are drawn with probability proportional to the real GMM-mode mixture (§4.2), reusing the identical mixture already fit for the `gmm_density` reward, instead of a uniform draw over training conversations. This hands the LLM baseline the same distributional signal (real mode proportions) that the GFlowNet's `gmm_density` reward uses, so a residual coverage gap in Table 2/Figure 3 is evidence of reward-proportional *sampling*, not of withholding information from the baseline.

In every tier the LLM self-tags: each output turn already carries a `move` field the model chooses from the domain's legal per-role move set, so structural fidelity is measured on a label the model committed to itself, not one assigned by a separate classifier; the one role/domain combination with only a single legal value (ESConv's `Seeker-Response` sentinel) is asserted programmatically rather than left to the model. Generation uses the same k=3 few-shot cardinality and structured-output schema as our own realization pipeline (§3.2), with a JSON-escape repair retry for occasional malformed output.

F Evaluation Protocol Details

Real-side sampling and partitions. Sample-level metrics compare 200 generated conversations (drawn from the 1500 per condition) against 200 held-out real conversations, with a reference row scoring 200 held-out real conversations as if generated. For tutoring, the evaluation and reference samples are disjoint draws from the 1000 held-out conversations. ESConv has only 260 held-out conversations, so its evaluation sample is drawn from those and its reference sample from 200 training conversations withheld for this purpose, leaving 840 for reward calibration. We redraw the real side 10 times for pMSE and 50 times for IP_α, IR_β, and authenticity, redrawing the reference row as often as every other row, and report means with standard deviations.

Distances and thresholds. We estimate the whitening transform on training conversations only, never on the conversations being scored. IR_β uses the fifth-nearest real neighbour as its radius. A generated conversation is an outlier when its distance to the nearest training conversation exceeds the 95th percentile of that distance among held-out real conversations—6.52 for tutoring, 4.90 for ESConv. Outlier rates are computed over all 1500 conversations per condition; the other columns use the 200 drawn each round. The structural upper bound is scored under the same protocol at n=200, realized by GEMINI-2.5-FLASH alone.

G Metric Estimation Details

pMSE. The discriminator (§4.2) is fit with 5-fold cross-validation; each record's propensity score $\hat{p}_i$ is its out-of-fold prediction, avoiding the optimistic bias of scoring a record with a model fit on it.

G.1 Sample-Level Fidelity, Coverage, and Authenticity: Definitions

All three metrics are computed on $\phi(x)$ after a whitening transform fit on the real training split, so no single feature dominates the distances.

Mode computation

α-precision and β-recall. *α-precision* P_α is the fraction of generated conversations that fall inside the ball containing the closest fraction α of real conversations; *β-recall* R_β is its mirror on the real side, a real conversation counting as covered when its nearest generated neighbor lies within its k=5 nearest-real-neighbor radius (Eq. 9 of Alaa et al., 2022). A generator matching the real distribution yields P_α=α and R_β=β for all α, β; we summarize each curve by its integrated deviation from that diagonal,

$$\mathrm{IP}_\alpha = 1 - 2\int_0^1 |P_\alpha - \alpha|\, d\alpha, \qquad \mathrm{IR}_\beta = 1 - 2\int_0^1 |R_\beta - \beta|\, d\beta, \tag{5}$$

evaluated on a 30-step grid, so 1 is a perfect match. By Theorem 1 of Alaa et al. (2022) both curves lie on the diagonal *iff* the generated and real distributions coincide.

Authenticity and copying. High IP_α certifies that generated conversations lie on the real support but not that they are *novel* points on it: a generator replaying training conversations would score just as well. *Authenticity* tests this via the distance from each generated conversation to its nearest real *training* record (distance to closest record, DCR): a generated conversation is inauthentic if its DCR falls below that training record's own nearest-neighbor distance, i.e. it sits closer to a training point than real conversations typically do. We also report a *near-copy rate* (fraction with DCR

Tier	Combined	Surface	Content	Structure	Dynamics	Outcome
Tutoring B1 (naive)	98.7	75.0	95.9	79.0	71.0	10.9
Tutoring B2 (scenario)	98.4	72.2	90.5	61.3	47.9	5.8
Tutoring B3 (scenario+mode)	98.6	70.3	89.5	68.3	44.8	6.8
ESConv B1 (naive)	95.2	96.1	75.3	69.3	73.6	68.3
ESConv B2 (scenario)†	95.0	87.2	62.4	68.1	70.4	51.8
ESConv B3 (scenario+mode)	96.8	98.1	78.8	69.7	72.5	53.4

Table 6: Per-tier, per-feature-group pMSE for the three LLM-baseline tiers (§4.3), n=400 each, GPT-5.6-terra.

below τ_{mem}) and *outlier rate* (fraction with DCR above τ_{out}), with per-domain thresholds calibrated so a held-out real sample attains outlier rate 0.05 by construction and sets the reference for the other two: τ_{mem} is the 5th percentile of leave-one-out DCR within the training split, and τ_{out} the 95th percentile of held-out-real-to-training DCR. This follows novelty-alongside-quality reporting in candidate-generation GFlowNets (Bengio et al., 2021). For each synthetic conversation, we plot its distance to its nearest real training match, $d(x, r)$, against that match's own nearest real neighbor $R(r)$ (Alaa et al., 2022).

Near-copy and outlier thresholds. τ_{mem} is the 5th percentile of leave-one-out DCR within the training split; τ_{out} is the 95th percentile of held-out-real-to-training DCR. Both are calibrated per domain so that a held-out real sample, scored against the training split exactly as any synthetic condition is, has outlier rate 0.05 by construction – giving every other condition's near-copy and outlier rates a shared, interpretable reference point.

G.2 pMSE for LLM-baseline conditions

Table 6 summarizes the pMSE scores for the LLM conditions.

G.3 Fidelity, Diversity, Coverage for ESConv

Condition	pMSE ↓	IP_α ↑	IR_β ↑	Auth. ↑	Near-copy	Outlier
Real held-out (reference)	7.0 ±1.6	0.96 ±0.01	0.96 ±0.01	0.67	0.072	0.050
LLM only (B2), GPT-5.6-terra	95.5 ±2.3	0.43 ±0.03	0.06 ±0.02	0.94	0.000	0.123
REINFORCE, GPT-5.6-terra	96.0 ±1.6	0.27 ±0.02	0.06 ±0.02	0.96	0.001	0.282
GFlowNet (TB), GPT-5.6-terra	93.8 ±2.7	0.29 ±0.02	0.22 ±0.03	0.93	0.001	0.410
LLM only (B2), Gemini 2.5 Flash	91.6 ±2.0	0.78 ±0.04	0.42 ±0.04	0.67	0.025	0.013
REINFORCE, Gemini 2.5 Flash	92.7 ±2.0	0.52 ±0.04	0.12 ±0.02	0.75	0.025	0.103
GFlowNet (TB), Gemini 2.5 Flash	86.4 ±3.4	0.52 ±0.04	0.50 ±0.05	0.86	0.005	0.249

Table 7: Sample-level fidelity (IP_α), coverage (IR_β), and authenticity for ESConv (§G.1), n=1500 per condition, both realizers. Each draw scores 200 generated conversations against 200 held-out real conversations on $\phi(x)$, whitened on the training split; pMSE is the mean ± sd over 10 draws and IP_α, IR_β, authenticity over 50 draws at k=5, with near-copy and outlier computed over all 1500. The first row scores held-out real conversations against the training split exactly as a synthetic condition is scored, giving the value a perfect generator attains at this sample size; read every other row against it rather than against 1. pMSE here is 200 vs. 200 and so runs below the n=1500 values of Table 2, with the same ordering. The two providers used different ESConv realization prompts (§3.2), so rows compare within a provider; the GFlowNet rows are the `gmm_density` reward with the inference-time directive mask off, not the guard-on configuration.