Can LLMs Generate High-Quality Task-Specific Conversations?

Anonymous ACL submission

Abstract

This paper introduces a parameterization framework for controlling conversation quality in large language models. We explore nine key parameters across six dimensions that enable precise specification of dialogue properties. Through experiments with state-of-the-art LLMs, we demonstrate that parameter-based control produces statistically significant differences in generated conversation properties. Our approach addresses challenges in conversation generation, including topic coherence, knowledge progression, character consistency, and control granularity. The framework provides a standardized method for conversation quality control with applications in education, therapy, customer service, and entertainment. Future work will focus on implementing additional parameters through architectural modifications and developing benchmark datasets for evaluation.

1 Introduction

Generative AI systems can now produce coherent text, images, and code at scale (OpenAI et al., 2024b; Rombach et al., 2022; Chen et al., 2021). However, for multi-turn dialogue, model behavior remains difficult to specify and verify. Even when outputs appear plausible, properties such as topic coherence, knowledge progression, and persona stability can vary substantially across turns. This gap motivates a controllability interface that is both expressive enough to describe desired dialogue properties and measurable enough to evaluate whether those properties were realized.

In this paper, we study whether current LLMs admit a parameterized control interface for multi-turn conversation generation. Our objective is generation-facing: can we expose a set of dialogue attributes as explicit parameters, and does varying these parameters induce recoverable, systematic changes in the generated transcripts? Concretely, we treat end-to-end conversation generation as a controlled simulation setting, where the transcript itself must manifest the configured properties, rather than relying on post-hoc prompting heuristics. (Zhang et al., 2020; Roller et al., 2021)

In domains such as business advising, civic services, and educational support, researchers often require long, context-rich dialogues that reflect specific user backgrounds, knowledge levels, and interaction styles. Yet existing simulator designs typically express these properties as natural-language instructions, leaving the intended attributes underspecified and difficult to measure. As a result, it is often unclear whether observed gains come from true controllability, model quality, or prompt artifacts. (Li et al., 2023; Huang et al., 2020; Bender et al., 2021)

We address this by introducing a parameter schema paired with a recoverability-based evaluation protocol. The schema makes conversation properties explicit and systematically variable; the protocol tests whether those properties are actually realized by asking independent judges to recover the parameters from transcripts. This framing directly operationalizes parameter adherence as the central validity criterion for a controllable generation interface, and separates controllability from incidental improvements in generic text quality. (Zheng et al., 2023)

The key contributions of this paper are:

1. A taxonomy of 35 conversation parameters with 9 dominating factors organized into six dimensions that capture essential aspects of multi-turn conversation control.

2. Analysis of parameter impact and redundancy under our recoverability-based evaluation protocol, identifying a core subset of high-impact parameters and documenting where additional parameters provide diminishing or overlapping gains.

3. A recoverability-based evaluation protocol that operationalizes parameter adherence as a transcript-level criterion for controllable generation.

4. A benchmark study showing that modern LLMs can implement a subset of these parameters through prompt conditioning, with measurable and statistically significant behavioral differences when parameters are varied.

5. An external validity design that tests whether recovered parameters transfer across dialogue regimes beyond simulator-only corpora.

In this paper, we introduce a parameterization framework for LLM-based conversation generation. Unlike unstructured prompting approaches, this parameterization enables precise specification of conversation properties that can be systematically varied, measured, and optimized. This approach builds upon prior work in controlled text generation (Keskar et al., 2019; Dathathri et al., 2019; Khalifa et al., 2021) but extends these techniques specifically for multi-turn dialogue contexts with novel parameter dimensions.

2 Related Work

A large body of work studies controllable generation by conditioning on attributes or control variables, including early approaches to conditional language modeling and plug-in control mechanisms (Keskar et al., 2019; Dathathri et al., 2019; Khalifa et al., 2021). More recent work on steerability and preference-conditioned generation shows that LLM behavior can be moved along interpretable axes and optimized for diverse user preferences, but is often evaluated at the utterance level or with short contexts (Liang et al., 2024; Wang et al., 2024). In contrast, our target is *transcript-level control*: parameters such as smoothness, focus, knowledge gap, and decision style must remain stable (or evolve in a prescribed way) across many turns, and success is defined by whether the configured attributes are recoverable from the resulting multi-turn conversation.

LLM-driven simulators are increasingly used to generate task-specific dialogues when real conversations are scarce, especially in low-resource settings. Recent work demonstrates practical frameworks for synthetic dialogue generation, but also highlights a central risk for prompt-based control: apparent improvements may reflect template artifacts rather than portable conversational properties (Suresh et al., 2025; Castillo-Bolado et al., 2024). This concern motivates our external validity design, where we evaluate whether parameters recovered from real corpora can be used to condition continuation generation in directionally consistent ways across different dialogue regimes.

Conversation quality is known to be multidimensional, and prior work documents persistent gaps between automatic metrics and human judgments, especially as context length grows (Deriu et al., 2020; Mehri and Eskenazi, 2020; See et al., 2019; Zheng et al., 2023). A growing set of benchmarks now explicitly targets multi-turn capabilities, interactive agent behavior, and long-horizon instruction following, providing evidence that strong single-turn performance does not guarantee robust multi-turn behavior (Kwan et al., 2024; He et al., 2024; Sirdeshmukh et al., 2025; Jia et al., 2025; Castillo-Bolado et al., 2024). Our evaluation objective differs from generic capability scoring: we measure whether a *specified control setting* is realized in the transcript, operationalized via parameter recoverability and directionally consistent metric shifts under controlled perturbations.

Maintaining stable persona traits and role-specific behavior over extended interactions remains a known failure mode in dialogue generation, including issues with entity tracking, coreference, and contradiction over multiple turns (Li et al., 2016; Zhang et al., 2018; Xu et al., 2022; Dziri et al., 2022). Recent benchmarks focusing on dynamic role-play and persona-conditioned interaction further emphasize that long-horizon consistency is difficult even when local responses appear plausible (Yuan et al., 2025; Tan et al., 2025). Our parameter schema explicitly includes user identity and decision-making style controls, and our adherence metrics quantify where models succeed and where they fail under long contexts and cross-domain settings.

Finally, recent work proposes robustness-oriented multi-turn benchmarks and holistic evaluation in spoken dialogue regimes, emphasizing distribution shift, error accumulation, and interaction-level brittleness (Xiang et al., 2025; Yan et al., 2025). These findings align with our motivation for external validity testing: if a control interface is meaningful, it should remain interpretable and recoverable beyond a single synthetic generation setup.

3 Methods

For this exploratory study, we selected nine parameters that has the highest impact from the 35 factors of conversation quality, which are spread across the six dimensions.

Turn: The number of turns of the conversation.

Industry Context: The initial field of this conversation.

Knowledge Gap Level: The prior knowledge the entrepreneur has of the conversation's field. This is a method used in (Baskar et al., 2025) to measure the model's knowledge alignment with the entrepreneur. We define the gap as a 1-5 integer value, where 1 refers to an expert with a deep understanding of the domain, and 5 refers to a complete novice with minimal business knowledge about their ideas.

Smoothness Factor: A grade A-F indicating conversation flow, with A referring to a perfectly flowing conversation with logical transitions, and F referring to a highly disjointed conversation with random topic jumping.

Focus Level: A grade 1-5 indicating how focused the entrepreneur is on this conversation. 1 refers to free-flowing, wide-ranging conversation covering many aspects, and 5 refers to laser-focused on specific details of implementation.

Identity: The initial setting of the entrepreneur's background, which is used by (Aher et al., 2023) to simulate gender and racial diversity.

Technical Language Level: A 0-1 float number indicating the level of technical language the entrepreneur is using in the conversation. Similar methods were used in (Scarlatos et al., 2025) to trace knowledge levels in system-user conversation.

Formality Level: A 0-1 float number indicating the formal phrase usage in the conversation.

Decision-Making Style: The style of response the entrepreneur treats the system's response. It can be one of analytical, Intuitive, consultative, or impulsive.

The exact definition of other parameters used in the prompt, the precise definition of the value of each parameter, and examples can be found in the Appendix.

Prompt Engineering The data set is created by constructing parameterized prompts that combine three key components: a base conversation generation prompt specifying the business advisory scenario, detailed parameter definitions for each dimension, and the specific parameter values for each conversation instance. For each experimental condition, we systematically vary the parameter values while maintaining consistent entrepreneur background profiles and industry contexts. The final prompt is fed to the target LLM to generate complete multi-turn conversations. The full prompt structure with an example implementation is in the Appendix.

Model Selection We evaluate four state-of-the-art LLMs: *Gemini-2.5-pro* (Comanici et al., 2025), *Claude-3.7-sonnet* (Bai et al., 2022), *o3*, *o4-mini* (OpenAI, 2025), with other smaller or open-source LLMs: *Deepseek-r1* (DeepSeek-AI et al., 2025), *gpt-4o-mini* (OpenAI et al., 2024a), *Llama3.1:70b* (Grattafiori et al., 2024). Unless otherwise noted: temperature $= 0.7$, top-p $= 0.95$, max-tokens $= 8192$, no custom stop tokens. We generate by default 20 samples per random seed.

Baseline We use prompt-based simulation using Claude Model *claude-3.7-sonnet* (Bai et al., 2022) as our baseline since it has the best performance among all other vanilla LLMs. (see the Appendix for baseline model comparison). Baseline results are produced using only the target turn, a random initial character setting with a brief background and previous experience with no special prompts or parameters, and rely solely on the LLM's ability to generate outputs.

Evaluation Methods We assess our simulator along five axes using the same generated corpora: (i) *Topic Diversity* (topic counts and entropy), (ii) *Parameter Adherence* (recovery of set parameters from transcripts), (iii) *Topic Drift* (embedding-based distance to initial topic), (iv) *Character Properties Stability* (persona consistency and contradiction/memory errors), and (v) *Entity Revisit Rate* (and usefulness). Operational definitions and prompts for each axis appear in Appendix F. Human protocols and agreement statistics are summarized in the main text and fully detailed in Appendix A. To mitigate circularity in LLM-as-judge evaluation, we calibrate the judge against human labels on a held-out subset and report human-judge correspondence, and replicate core recoverability results with a consensus on multiple judges from a different provider when the generator under evaluation matches the default judge family.

For each task, we create a set of simulators and control the parameters to generate task-specific con-

versations from the same topic.

Topic Diversity We use a random seed to create 800 entrepreneurs' background data. The generated parameters are then injected into the prompt and fed into each LLM. We estimate topic diversity by clustering conversation topic summaries using sentence embeddings with cosine distance and a min cluster size set per dataset to maintain comparable granularity. For transparency, we additionally report sensitivity across a small range of clustering thresholds in Appendix F. We also compared the diversity of the topics by entropy: $H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$, where $p(x_i)$ is the probability of topic x_i.

Parameter Adherence We generate 200 entrepreneurs' background data and randomized conversation parameters. These are fed into each LLM across four different conversation lengths: 5, 10, 15, and 20 turns, resulting in a total of 800 conversations. The evaluation employs a hybrid human-LLM assessment framework in which both human annotators and *Claude-sonnet-3.7* serve as judges.

The evaluation protocol provides judges with only the conversation transcript, requiring them to infer the original parameters based on predefined parameter definitions. For numerical parameters (all on the 1–5 Likert scale), adherence is measured by mean squared error: $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$, where y_i is the set value and $\hat{y}_i$ is the inferred value.

Topic Drift We generate 200 20-turn entrepreneur conversations, each with smoothness factor set to A (highest topic adherence) and F (lowest topic adherence), along with 200 baseline conversations without smoothness factor control, resulting in 600 total conversations for topic drift analysis. The smoothness factor parameter controls the degree to which conversations maintain thematic coherence versus allowing natural topic exploration and deviation from the original business concept.

We compute topic drift as cosine distance between the embedding of the entrepreneur utterance at turn t and the embedding of the opening topic cue: $\text{Drift}(t) = 1 - \cos(\mathbf{e}_t, \mathbf{e}_0)$, where $\mathbf{e}_t = \text{Embed}(\text{utterance}_t)$ and $\mathbf{e}_0 = \text{Embed}(\text{topic cue})$. Unless otherwise noted, embeddings use all-MiniLM-L6-v2 with mean pooling and ℓ_2 normalization.

Character Properties Stability We generate 500 20-turn conversations with both the entrepreneur's formality and technical levels randomized between 0 and 1, then 500 more with the formality parameter omitted and another 500 with the technical parameter omitted.

Character stability is evaluated by:

Formality Level: Formality is determined by a composite of vocabulary sophistication, sentence structure, and pronoun usage.

Technical Language Level: The technical level is determined by the density of the domain terminology, the complexity of the concepts, and the usage of jargon. The final stability score is calculated by the $1 - 0.5(\text{Formality Error} + \text{Technical Level Error})$

Entity Revisit Rate We generate 100 entrepreneurs' background information with Knowledge Gap Level parameters ranging from 1-5, where this parameter measures the knowledge disparity between the user's existing background and their proposed business concept. Each entrepreneur profile is used to generate conversations in four different lengths (5, 10, 15, and 20 turns).

The evaluation is done by first extracting NER and core concepts using BERT. We then track when previously mentioned entities reappear in subsequent turns in the conversation. The concept of a recall rate is calculated as $\frac{1}{T-1}\sum_{t=2}^{T}|\text{Entities}_t \cap \bigcup_{i=1}^{t-1}\text{Entities}_i|$, where Entities_t represents the set of entities mentioned at turn t, and T is the total duration of the conversation.

External Datasets for Validity A concern in prompt-based controllable generation is that apparent control may reflect prompt artifacts in a synthetic setting rather than a portable interface that corresponds to properties observed in real dialogues. To test the validity of our parameter schema beyond simulator-only corpora, we evaluate our framework on three publicly accessible, multi-turn dialogue datasets that represent distinct dialogue regimes. We use the Fin-Ally (Das et al., 2025), MultiWOZ 2.2 (Zang et al., 2020), and STARv2 (Zhao et al., 2023) as our datasets. We use these datasets to answer two validity questions. First, the parameter recoverability on real data. Given only a human-written transcript and parameter setting, can independent judges produce the similar dialogue that maintains the same conversation and user parameters? This tests whether our parameters correspond to observable properties of real dialogues. Second, the cross-domain portability. When we condition generation on a real dialogue prefix from each dataset and vary a single

Model	Embedding diversity
claude	0.2912
deepseek-r1	0.4161
o3	0.3360
o4-mini	0.2830
gpt-4.1	0.3436
gpt-4o-mini	0.2085
gemini	0.3747
llama3.1:70b	0.0576
baseline	0.1075

Table 1: Embedding diversity (sentence embedding).

Model	Topic diversity	Topic entropy
claude	111	4.469
deepseek-r1	143	5.275
o3	136	4.464
o4-mini	154	5.311
gpt-4.1	140	4.578
gpt-4o-mini	84	3.859
gemini	141	5.266
llama3	5	0.888
baseline	35	2.985

Table 2: Topic diversity and topic entropy.

Parameter	Fin-Ally	STARv2	MultiWOZ
Formality (ICC)	0.8119	0.7632	0.7407
Technicality (ICC)	0.7876	0.7495	0.7211
Focus (ICC)	–	0.7084	0.6834
Smoothness (ICC)	–	0.6764	0.6572
Decision Style (κ)	0.8320	–	–
Knowledge Gap (ICC)	0.7795	–	–

Table 3: Human agreement on external corpora (two blinded raters). Fin-Ally is short in the released CSV, so we omit long-horizon labels.

Dataset	Setting	Foc. MSE↓	Smth. MSE↓	Drift↓
STARv2	Unctrl.	0.9433	0.8954	0.4141
STARv2	Param.	**0.7587**	**0.7015**	**0.3276**
MultiWOZ	Unctrl.	0.9117	0.8696	0.3909
MultiWOZ	Param.	**0.7340**	**0.6818**	**0.3026**

Table 4: External validity via prefix-to-continuation generation. "Param." conditions generation on parameters recovered from the real prefix. Drift uses the same embedding-based topic drift metric as Sec. 3.

parameter while holding all other settings fixed, do we observe systematic and directionally consistent changes in transcript-level measurements across domains? These tests are designed to separate true controllability from dataset-specific prompting effects, and to ground our simulator evaluation in external corpora with different conversational structures.

Human Evaluation Protocols We reuse the exact conversation from our automatic evaluations. Human studies cover parameter adherence, smoothness factor, knowledge-gap sensitivity, decision-making style classification, concept-revisit rate. Each item is rated by 2 blinded raters, with stratified sampling by model, length, and parameter level. We report MSE/τ for numeric parameters and macro-F1/κ for categorical labels. For external datasets, annotators label a subset of the dataset with summaries and key parameters to test transfer and recoverability. Full task definitions, rater rubrics, sampling details, and statistical procedures appear in Appendix A.

4 Experiments

Parameter adherence varies across models with improving accuracy over extended conversations. A central concern with prompt-based controllable generation is that apparent control may be an artifact of the synthetic prompt distribution rather than a portable interface that corresponds to properties present in real dialogue. We therefore complement our simulator-only adherence curves with two external evaluations that mirror our main protocol while isolating different failure modes. Whether the parameters are recoverable from human-written transcripts under our definitions, and whether conditioning on those recovered parameters produces directional and measurable changes when continuing real multi-turn dialogues.

Table 3 reports inter-rater agreement on a labeled subset of three public corpora using the same parameter rubrics as Sec. 3. Agreement is consistently high for style-linked dimensions with strong surface realizations, including Formality and Technicality. Across both business-oriented dialogue (Fin-Ally) and task-oriented dialogue (STARv2/MultiWOZ). Discourse-level properties that require longer-range judgments, such as Focus and Smoothness, achieve moderate but stable agreement on STARv2 and MultiWOZ. This can reflect the expected annotator ambiguity when a transcript mixes local coherence with occasional topic exploration. Fin-Ally is short in the released dataset, so we do not report long-horizon labels there; instead, we focus on parameters that can be inferred reliably from local behavior, where Decision-Making Style reaches strong agreement and Knowledge Gap remains recoverable.

Table 4 then evaluates whether this recoverability can be translated into portable control in a realistic deployment setting. We use a prefix-to-continuation protocol that given a human-written

dialogue prefix from STARv2 or MultiWOZ, we generate a continuation either without control (Unctrl.) or conditioned on the parameters recovered from that prefix (Param.). If parameterization is valid beyond simulator corpora, it should reduce adherence error and stabilize topic tracking even when the context is out-of-distribution relative to our synthetic prompts. This is exactly what we observe. The result shows the parameterized approach effectively reduce the error rate in both datasets. The consistency of these gains across two distinct dialogue regimes supports the interpretation that our interface captures transferable dialogue properties rather than prompt-specific artifacts.

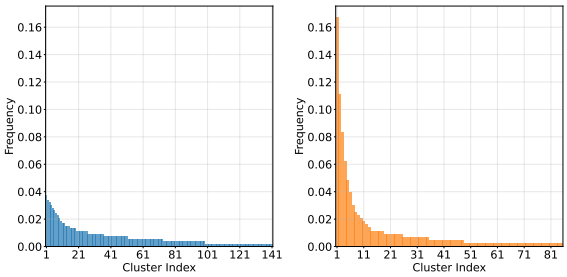


Figure 1: Topic frequency distributions of the *gpt-4o-mini* (orange) and *gemini-2.5-pro* (blue). Clusters are sorted in descending frequency. More advanced model (gemini-2.5-pro) produced a more diverse topic compared to the less advanced model (gpt-4o-mini)

Simulators have bias on topic selection, and may not generate a diverse pool of topics. The simulators can be classified into two broad camps according to their approach to exploring the subject matter, as shown in Table 7. Advanced models such as *Gemini-2.5-pro* and *DeepSeek-R1* exhibit superior topic diversification capabilities. These models demonstrate a more uniform attention distribution across thematic domains, closely approximating human-like conversational breadth. In contrast, less capable models like *GPT-4o-mini* produce more constrained topic distributions, while lightweight models such as *Llama3.1:70b* show severe limitations with only 5 distinct topics.

The baseline approach without parameterization yields poor diversity metrics. Mid-tier systems occupy an intermediate position, with respectable topic coverage but exhibiting concentration patterns around familiar conceptual clusters. This shows model architectures can explore diverse thematic spaces while maintaining coherent conversational flow.

We also examine sentence diversity by calculating semantic diversity through the cosine similarity of embeddings generated by all-MiniLM-L6-v2. (Table 1). The embedding diversity rankings partially diverge from topic-level diversity measures, suggesting that models may employ different strategies for achieving variation, and they may use similar words or add additional definitions (e.g., AI-driven business vs. non-AI-driven) to express different topics.

Beyond quantitative diversity measures, we observe systematic biases in topic selection patterns. For example, when generating food-related business scenarios, models frequently default to vegan or health-conscious options regardless of user specifications. This tendency toward "safe" or socially desirable recommendations indicates inherent training biases that may limit the authenticity of generated conversations.

Adding Smoothness factor improves topic correlation. Adding a smoothness factor to simulate the conversation flow not only creates a diversified conversation but also improves the model's adherence to the main topic. (Figure 2). Both the small and the more advanced models can improve adherence to the main topic after setting a high smoothness factor, and advanced models can successfully create a more significant difference between high and low smoothness factors. Without the smoothness factor, the model can only provide a conversation that has low correlation to the given topic. Human-labeled results show high agreement between human annotators and LLM judges ($\kappa = \mathbf{0.91}$).

Parameter adherence varies across models with improving accuracy over extended conversations. Analysis of parameter adherence across conversation turns reveals substantial differences in model capabilities, with most parameters showing improved accuracy as conversations progress. As shown in Figures 3(a-c), advanced models such as Claude and Gemini demonstrate superior parameter implementation, with MSE errors for the focus level, the knowledge gap level, and the experience level decreasing from initial values to more accurate parameter representation over 20 conversations compared to the ground truth. This improvement pattern suggests that models require several turns to fully establish and maintain specified parameter values. The evaluation of the decision-making style (Figure 3 (b)) shows binary classification ac-

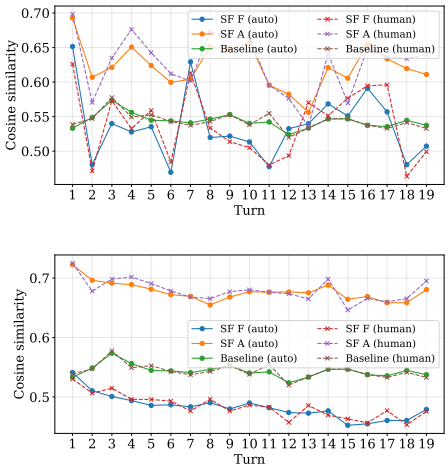
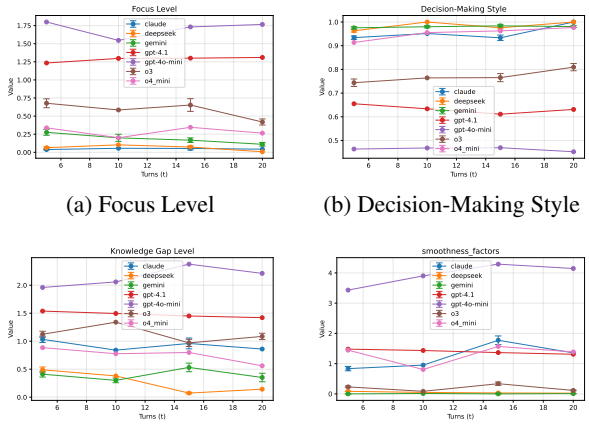


Figure 2: The cosine similarity between the entrepreneur's utterance to the main topic in different smoothness factors for two models: (a) *gpt-4o-mini*, (b) *claude-3.7-sonnet*. *claude-3.7-sonnet* is showing a high separation between the highest and lowest smoothness factor, showing better understanding and adherence to parameters.



(a) Focus Level (b) Decision-Making Style

(c) Knowledge Gap Level (d) Smoothness Factors

Figure 3: Model metric curves vs. conversation turns. The comparison between human evaluation can be found in App. E

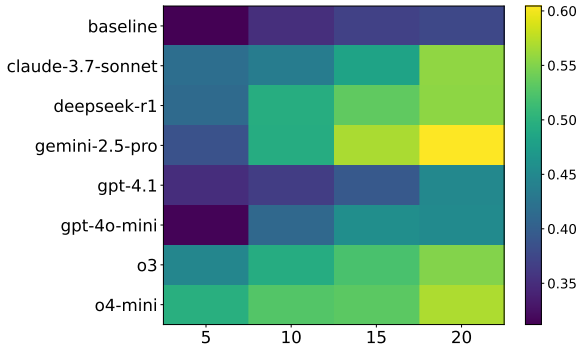


Figure 4: Concept-revisit rate by turns for each model with knowledge gap level of user set to 1 (most knowledgeable). All models exhibit a higher revisit rate with turn progression.

curacy, where advanced models achieve 0.8-1.0 accuracy rates while lighter models like *gpt-4o-mini* struggle to maintain consistent classification performance. The smoothness factor analysis (Figure 3(d)) demonstrates that parameter control effectiveness varies by model architecture, with Claude maintaining clear parameter differentiation while smaller models show less distinct parameter implementation regardless of specified values.

Knowledge gap parameters influence concept revisit patterns in advanced models. The relationship between Knowledge Gap Level and concept revisit behavior reveals substantial differences in advanced models' adaptation capabilities, as shown in Figure 8. *Gemini-2.5-pro* exhibits a clear inverse correlation between knowledge gap and revisit rate, with highly knowledgeable users (Level 1) showing revisit rates of approximately 0.5-0.6, while novice users (Level 5) demonstrate lower revisit rates around 0.1-0.2 across all conversation lengths. This pattern aligns with pedagogical theory, where experts benefit from reinforcement of complex concepts, while beginners require more linear information introduction. Conversely, Claude shows a lower differentiation between knowledge gap levels, but a higher differentiation over turns. This shows that some models cannot correctly simulate a conversation with low revisit rates.

With a high knowledge gap level, all models show a higher revisit rate compared to the baseline. (Figure 6). Advanced models, including *o3*, *gpt-4.1*, and *Claude-3.7-sonnet*, maintain high character consistency scores that improve over extended conversations, while mid-tier models show respectable but more variable performance. The baseline approach demonstrates significantly lower consistency. This suggests that sophisticated parameter implementation requires substantial model capacity to fully understand and adhere to the parameters, but all models can obtain a significant level of performance increase.

Character parameters are stable across all models. The character parameter study shows that all models can reach high parameter stability over turns, although more advanced models have better performance (Figure 6). All models exhibit improved stability trajectories over conversation

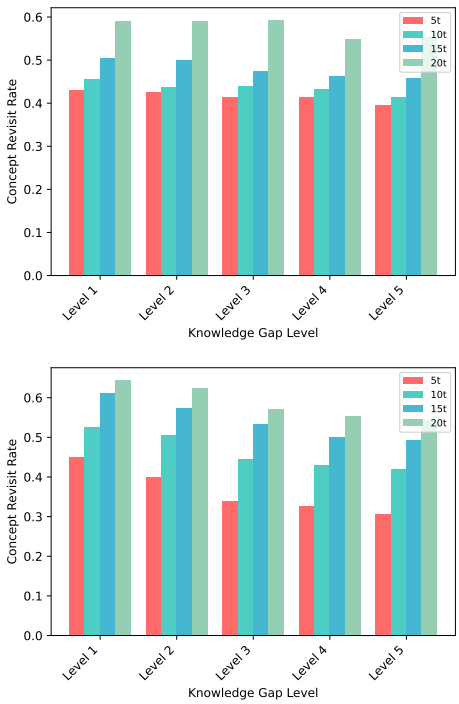


Figure 5: Concept-revisit rate by knowledge-gap level for two models: (a) Claude, (b) Gemini. Knowledge Gap Level 1 is the smallest knowledge gap, and Knowledge Gap Level 5 is the highest. *Gemini-2.5-pro* shows a more significant difference when modifying Knowledge Gap Level.

Model	Turns	Formality	Technical	Full
claude-3.7	5	0.280	0.252	0.206
claude-3.7	10	0.305	0.265	0.205
claude-3.7	15	0.298	0.258	0.192
claude-3.7	20	0.292	0.252	0.184
o3	5	0.255	0.212	0.173
o3	10	0.222	0.175	0.143
o3	15	0.215	0.162	0.131
o3	20	0.212	0.155	0.130

Table 5: Average performance errors for Formality Only, Technical Only, and Full Parameters across varying conversation turns. Human evaluation shows a correlation of Pearson's r = 0.938. See App. E for full human results and agreement analyses

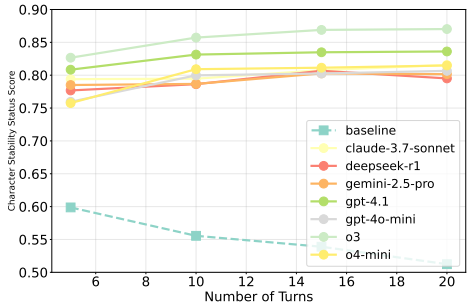


Figure 6: Character Parameter stability over turns, the baseline has a decreasing stability over turns, while all other models with character properties show an increase in stability score. The agreement between the human and LLM judge at baseline is κ = **0.9316**, and κ = **0.9241** for *claude-3.7-sonnet*

length, with consistency scores rising from initial values. This could be because the model does not have enough context initially, but the performance stabilizes after 15 turns.

We also performed an ablation analysis presented in Table 5, where we test the error of the model when only the formality parameters of the model or technical parameters are given. The result shows that the combined parameter implementation yields benefits exceeding the sum of individual components in both models. This suggests that adding more specified parameters to the model may further increase the model's capability of simulating complex conversations.

5 Conclusion

We create a comprehensive parameterization framework for controlling LLM-based conversation generation, demonstrating both the potential and limitations of current approaches to fine-grained dialogue control. Our experiments with the simulator show that advanced models can effectively differentiate between parameter values and maintain improving consistency over long conversations.

6 Limitations

Several issues are unaddressed in this exploratory study. We only provide the necessary parameters for conversation generation, not an exhaustive set of parameters that covers all aspects. More parameters could be added to the prompt since we have already proven that interconnected parameters can improve conversation quality.

A fine-tuned LLM with human-labeled conversation parameters as a dataset may increase the simulator's sensitivity to intermediate values. We are using the default temperature settings. More analysis could be made on different parameter settings and fine-tuned open-source LLMs.

Parameterized settings cannot increase the model's factual accuracy. Adding a factual accuracy parameter can prompt the LLM to provide incorrect information, but they are also not sensitive enough to intermediate parameters and does not decrease the hallucination rate compared to the vanilla model. A RAG-based approach is still needed to decrease the simulator's hallucination.

7 Ethical Considerations

We adhere to the ACL Code of Ethics and will submit the Responsible NLP checklist.

Our contribution is a *simulation* framework for controllable, multi-turn conversations. Risks include persuasive misuse, misrepresentation of synthetic dialogues as human, and style emulation of specific groups. We label released samples as *synthetic*, recommend visible and machine-readable provenance, prohibit deceptive use, and discourage targeting real individuals.

All API models are used under their providers' standard research terms; Llama-3.1-70B is used under its open-source license. No user PII is sent; prompts contain synthetic content only. We vary parameters that affect coherence/focus to study controllability, not to generate harmful content. Degradation is limited to stylistic/structural properties in benign domains, with disallowed-content checks. We stress that parameterization does *not* ensure truthfulness; downstream use should add retrieval/verification.

The *Identity* parameter can reinforce stereotypes if entangled with behavior. We decouple identity from behavioral controls, randomize combinations, and encourage reporting group-wise summaries and disparity checks. The framework must not be used to infer protected attributes.

Corpora are model-generated; no PII is collected or released. Any artifacts (prompts, code, metrics) will be filtered for PII/copyright and released under a research license forbidding deceptive use.

Annotators worked only with synthetic text, provided informed consent, could withdraw, and were fairly compensated; no demographics were collected.

We provide prompts/parameter grids, seeds, and evaluation code for replication and report computing to limit environmental impact.

Simulators aid pre-deployment testing but do not replace human studies; final evaluations require real users under appropriate ethical review and consent.

8 LLM usage

We use LLM mainly for paraphrasing sentences and checking grammar mistakes.

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. pages 337–371.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback.

Sarvesh Baskar, Tanmay Tulsidas Verelakar, Srinivasan Parthasarathy, and Manas Gaur. 2025. From guessing to asking: An approach to resolving the persona knowledge gap in llms during multi-turn conversations.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? pages 610–623.

David Castillo-Bolado, Joseph Davidson, Finlay Gray, and Marek Rosa. 2024. Beyond prompts: Dynamic conversational benchmarking of large language models. *Preprint*, arXiv:2409.20222.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, and 1 others. 2021. Evaluating large language models trained on code.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Sarmistha Das, Priya Mathur, Ishani Sharma, Sriparna Saha, Kitsuchart Pasupa, and Alka Maurya. 2025. Fin-ally: Pioneering the development of an advanced, commonsense-embedded conversational ai for money matters. *Preprint*, arXiv:2509.24342.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo Ponti, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? pages 5271–5285.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, and 1 others. 2024. The llama 3 herd of models.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. 2024. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *Preprint*, arXiv:2410.15553.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. 38(3):1–32.

Qi Jia, Ye Shen, Xiujie Song, Kaiwei Zhang, Shibo Wang, Dun Pei, Xiangyang Zhu, and Guangtao Zhai. 2025. One battle after another: Probing llms' limits on multi-turn instruction following with a benchmark evolving framework. *Preprint*, arXiv:2511.03508.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. 2021. A distributional approach to controlled text generation.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *Preprint*, arXiv:2401.16745.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. pages 994–1003.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable text generation for large language models: A survey. *Preprint*, arXiv:2408.12599.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, and 1 others. 2024a. Gpt-4o system card.

OpenAI. 2025. Introducing OpenAI o3 and o4-mini. Accessed: 2025-07-29.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and 1 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. pages 300–325.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-resolution image synthesis with latent diffusion models. pages 10684–10695.

Alexander Scarlatos, Ryan S Baker, and Andrew Lan. 2025. Exploring knowledge tracing in tutor-student dialogues using llms. pages 249–259.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. 2025. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. *Preprint*, arXiv:2501.17399.

Sathya Krishnan Suresh, Wu Mengjun, Tushar Pranav, and Eng Siong Chng. 2025. Diasynth: Synthetic dialogue generation framework for low resource dialogue applications. *Preprint*, arXiv:2409.19020.

Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh Murthy, Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, Silvio Savarese, Huan Wang, Caiming Xiong, and Shelby Heinecke. 2025. Personabench: Evaluating ai models on understanding personal information through accessing (synthetic) private user data. *Preprint*, arXiv:2502.20616.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8642–8655, Bangkok, Thailand. Association for Computational Linguistics.

Hao Xiang, Tianyi Tang, Yang Su, Bowen Yu, An Yang, Fei Huang, Yichang Zhang, Yaojie Lu, Hongyu Lin, Xianpei Han, Jingren Zhou, Junyang Lin, and Le Sun. 2025. Rmtbench: Benchmarking llms through multi-turn user-centric role-playing. *Preprint*, arXiv:2507.20352.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. pages 5180–5197.

Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. 2025. Uro-bench: Towards comprehensive evaluation for end-to-end spoken dialogue models. *Preprint*, arXiv:2502.17810.

Dingbo Yuan, Yipeng Chen, Guodong Liu, Chenchen Li, Chengfu Tang, Dongxu Zhang, Zhenkui Wang, Xudong Wang, and Song Liu. 2025. Dmt-rolebench: A dynamic multi-turn dialogue based benchmark for role-playing evaluation of large language model and agent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25760–25768.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. *Preprint*, arXiv:2007.12720.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? pages 2204–2213.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation.

Jeffrey Zhao, Yuan Cao, Raghav Gupta, Harrison Lee, Abhinav Rastogi, Mingqiu Wang, Hagen Soltau, Izhak Shafran, and Yonghui Wu. 2023. Anytod: A programmable task-oriented dialog system. *Preprint*, arXiv:2212.09939.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Human Evaluation Protocols

A.1 Participants

We recruited five experienced university student volunteers with 3 different expertise areas located in the US (each >6 months of prior annotation work) for the human study. Annotators completed a brief screening task. No personally identifying information was collected. The study was reviewed as non–human-subjects research under institutional guidance because annotators evaluated synthetic text only and provided no private information; nevertheless, we followed standard consent and data-minimization practices.

Limitation. With $N=10$ raters this pilot primarily establishes feasibility and inter-rater reliability bounds; it does not provide population-level estimates. We therefore report uncertainty and treat the human study as confirmatory for construct validity and judge calibration, not as a definitive benchmark.

A.2 Blinding

Annotators were strictly blind to (i) the generating model identity/version, (ii) all parameter settings, and (iii) the study hypotheses. Each task UI displayed only the initial topic cue (one-line description) and the conversation transcript in a structured JSON-like view. Conversation length was visible by necessity but randomized across items.

A.3 Tasks and Rater Rubrics

All tasks used written rubrics with anchors and examples.

- **Parameter Adherence.**
  - *Focus Level (1–5):* We use the exact same definition as in the prompt, with examples for each score level.
  - *Smoothness (A–F→1–6):* Anchors include turn-to-turn linkage checks.
  - *Knowledge Gap Level (1–5):* inferred user expertise; 1 (expert) to 5 (novice). Anchors emphasize initiative, prerequisite vocabulary, and abstraction level.
  - *Formality / Technicality (0.0–1.0 sliders):* defined via lexicon sophistication, syntax, register; and domain-term density/jargon use. Anchored examples provided.
  - *Decision-Making Style (categorical):* {analytical, intuitive, consultative, risk-averse, impulsive}. Definitions with exemplars; single-best-label instruction.
- **Character Consistency Audit.** Per-turn tagging of violations: *contradiction* (factual/persona), *memory error*, *style drift*. Raters mark presence/absence and leave a short note at first occurrence.

- **KGL Sensitivity.** Blind inference of KGL with a scaffolding checklist (evidence of prerequisite knowledge, abstraction level, need for definitions, rate of requests for clarification).

- **Entity Revisit Usefulness.** For a sampled subset of turns, raters label entity revisits as {reinforce, advance, clarify, superficial}, with examples.

- **Controlled Degradation.** For items generated under low-quality settings (e.g., Smoothness=F), raters answer a binary check: "Does the transcript manifest the specified degradation?" with short justification.

For the external datasets, the annotators are given the original dialogue, and in addition to annotating the original dataset based on the 9 main parameters, they are also prompted to provide a summarization of the dialogue as the original prompt to the model for future generation tasks.

A.4 Sampling and Power

We stratified by *model* (all models under test), *turn length* (5/10/15/20), and *parameter level* (extremes + midpoints where applicable). Within each stratum, we uniformly sampled conversation instances from the automatic pool. Given $N{=}2$ raters, the study is underpowered for small effects; our *a priori* goal was precision of reliability (95% CI width $\leq$0.20 for κ in high-agreement regimes) and coarse effect-size estimation. We therefore report bootstrap CIs and avoid null-hypothesis claims when CIs are wide.

A.5 Rater Recruitment, Training, and Procedure

Annotators completed: (i) a 30-minute tutorial with screenshots, (ii) a guided practice set of 12 items with gold explanations, (iii) a qualification quiz (pass threshold: $\geq$80% agreement with gold rationales). Items were presented in randomized order with per-rater unique shuffles. Model and parameter metadata were removed from the payload and filenames. Each session included regular, unlabeled gold checks.

A.6 Quality Control and Exclusions

We applied: (i) gold-check filters (remove rater-session segments with $<$70% gold agreement), (ii) attention checks (detect straight-lining on sliders), (iii) time-on-task filters (remove items below 10th percentile viewing time unless justification note present), and (iv) duplication prevention.

A.7 Metrics and Statistical Analysis

- **Agreement.** For numeric scales we report Intraclass Correlation Coefficient (ICC(2,k)) and Krippendorff's α; for categoricals (Decision Style) we report Cohen's κ. All with nonparametric 95% CIs via 1,000 bootstrap replicates clustered by item.

- **Mixed-Effects Models.** We fit linear/logistic mixed-effects models with fixed effects for parameter level, model, and turn length, and random intercepts for prompt and rater. Model selection used AIC with maximal random-effects structure feasible for convergence.

- **Multiple Testing.** We control the false discovery rate across families of related hypotheses via Benjamini–Hochberg ($q{=}0.10$) and report both raw and adjusted p-values when shown.

A.8 LLM-Judge Calibration ("Judge Card")

We calibrate an LLM judge on a rater-labeled subset (70/30 train/holdout split, stratified by task). For each metric we fit a bias-correction model (ordinary least squares for numeric, multinomial logistic for categorical) mapping LLM scores to human scores. We report: (i) human–LLM correlation (Pearson/Spearman) on the holdout set, (ii) calibration plots, (iii) post-calibration RMSE / macro-F1, and (iv) recommended use constraints (e.g., avoid mid-level Smoothness adjudication without human review). To prevent family overlap bias, we avoid using a judge from the same provider as any compared generator in sensitivity analyses and replicate with an alternative judge to assess robustness.

A.9 Data and Artifact Availability

We release (under a research license): anonymized prompts for annotation, rater rubrics/instructions, gold examples, and the code used for sampling, UI rendering, agreement computation, bootstrap, and mixed-effects modeling. Generated transcripts are released where license permits; otherwise we provide hashes and per-turn feature summaries.

B Selecting the High-Impact Parameters

We first define a neutral configuration c_{neutral} over all 35 parameters. For each parameter p, we specify

two extreme but valid values $(p_{\text{low}}, p_{\text{high}})$ consistent with our parameter definitions, while keeping all remaining parameters fixed at their neutral values.

We then sample M entrepreneur background profiles (we use M=200 in our planned experiment, matching the scale used for parameter adherence in Sec. 3). For each background b and each parameter p, we generate a matched pair of conversations:

$$c_{p,b}^{\text{low}}: \text{ generated with } c_{\text{neutral}} \text{ but with } p \leftarrow p_{\text{low}},$$
$$c_{p,b}^{\text{high}}: \text{ generated with } c_{\text{neutral}} \text{ but with } p \leftarrow p_{\text{high}}.$$

All runs use the same model, the same random-seed policy, and the same fixed conversation length to remove confounds. This yields, for each parameter p, a dataset

$$D(p) = \{(b, c_{p,b}^{\text{low}}, c_{p,b}^{\text{high}}) \mid b \in \{1, \ldots, M\}\},$$

in which any systematic differences between $c_{p,b}^{\text{low}}$ and $c_{p,b}^{\text{high}}$ can be attributed to the single parameter p.

Behavioral effect size. Given the single-parameter datasets $D(p)$, we quantify how much each parameter changes conversation behavior using the same metrics as in Sec. 3. For a given parameter p and metric A, we compute the signed difference for each matched pair:

$$\Delta_{p,b}^{A} = A(c_{p,b}^{\text{high}}) - A(c_{p,b}^{\text{low}}).$$

We summarize its magnitude by the mean effect size:

$$E^{A}(p) = E_b\left[\Delta_{p,b}^{A}\right].$$

To compare parameters across metrics with different scales, we z-normalize effect sizes per metric:

$$\tilde{E}^{A}(p) = \frac{E^{A}(p) - \mu_A}{\sigma_A},$$

where μ_A and σ_A are the mean and standard deviation of $\{E^{A}(p)\}_{p=1}^{35}$ for metric A.

Controllability via parameter recovery. We also measure how controllable each parameter is, using the same parameter-recovery idea as in our main study. For each p, we pool all conversations in $D(p)$ and ask *Claude-3.7-sonnet* to infer whether a given transcript was generated with p_{low} or p_{high} (without seeing the parameter values). This yields a classification accuracy $\text{Acc}(p)$ for distinguishing low vs. high settings. We convert this into a $[0, 1]$ controllability score:

$$C(p) = \max\big(0,\, 2 \cdot (\text{Acc}(p) - 0.5)\big),$$

so that random guessing (Acc=0.5) maps to $C(p)$=0 and perfect identification maps to $C(p)$=1.

Parameter	Mean $\lvert\tilde{E}(p)\rvert$	$C(p)$	$S(p)$
Knowledge Gap Level	1.393	0.9245	**1.288**
Smoothness Factor	1.329	0.9264	**1.231**
Focus Level	1.090	0.9082	**0.9971**
Technical Language Level	0.9557	0.9265	**0.8855**
Formality Level	0.8747	0.9457	**0.8272**
Decision-Making Style	0.8024	0.9097	**0.7319**
Identity (User)	0.8192	0.8459	**0.6929**
Industry Context	0.7171	0.8159	**0.5851**
Complexity Progression	0.5462	0.7530	0.4113
Practical–Theoretical Balance	0.5040	0.7271	0.3664
Framework	0.4786	0.7063	0.3380
Feedback Reception	0.4418	0.6752	0.2983
Emotional Journey	0.4266	0.6438	0.2746
Factual Accuracy	0.4036	0.5940	0.2397

Table 6: Single-parameter sensitivity ranking. We remove *Turns* since it is always included as a base setting in all experiments.

Overall sensitivity index. For each parameter, we define an overall sensitivity index that combines how much it changes the conversation with how reliably that change is recoverable from the transcript:

$$S(p) = C(p) \cdot \frac{1}{|\mathcal{A}(p)|} \sum_{A \in \mathcal{A}(p)} \tilde{E}^{A}(p),$$

where $\mathcal{A}(p)$ is the subset of metrics theoretically relevant to parameter p.

Intermediate values. For intermediate parameter values, we agree with the reviewers' reading and explicitly report this limitation in Sec. **??**: current models reliably separate extreme settings, but show weak discrimination among mid-range settings. Our goal is to empirically document this limitation, not to claim that the framework currently yields perfect fine-grained control.

C Prompts

In this section, we present the prompt used for conversation generation.

C.1 Raw Prompt

Raw Prompt

Create a K-turn conversation between an AI adviser and an entrepreneur trying to work on <A business field>. In the conversation, the AI adviser is an informed business coach in a Small Business Development Corporation, and the entrepreneur is a < entrepreneur's demographic background > with a focus on <entrepreneur's idea>.

C.2 Parameterized Prompt

Below is the complete prompt to the LLM for parameterized conversation generation:

Parameterized Prompt

Create a K-turn conversation between an AI adviser and an entrepreneur trying to work on <A business field>. In the conversation, the AI adviser is an informed business coach in a Small Business Development Corporation, and the entrepreneur is a < entrepreneur's demographic background > with a focus on <entrepreneur's idea>.
Core structural parameters that define the conversation's basic framework:

- **Purpose:** The primary intent of the conversation
  - advisory: Problem-solving and guidance-focused dialogue
  - educational: Knowledge transfer and learning-oriented
  - exploratory: Discovery and brainstorming-centered
  - evaluative: Assessment and critique-focused
- **Turns:** Total number of conversation turns (exchanges between participants)
- **Turn Balance:** Distribution of conversation contributions between participants (expressed as ratio, e.g., "55:45" means user speaks 55% of turns, advisor 45%)
- **Arc:** Overall narrative structure of the conversation
  - problem-solution: Identifies issues and develops solutions
  - exploration-conclusion: Broad investigation leading to specific outcomes
  - question-answer: Sequential inquiry and response pattern
  - build-refine: Iterative development and improvement process
- **Initiator:** Which participant starts the conversation
  - user: Entrepreneur begins with question or problem
  - assistant: Advisor opens with inquiry or observation
- **Topic Scope:** Array of subject areas that may be covered during the conversation (e.g., ["food business", "marketing", "operations"])

Parameters defining the characteristics and relationship between conversation participants:

- **Knowledge Gap Level (KGL)**
  - **1:** Expert with deep understanding of business domain
  - **2:** Advanced practitioner with solid foundational knowledge and some specialized expertise
  - **3:** Moderate familiarity with business concepts
  - **4:** Basic understanding with significant knowledge gaps requiring guidance
  - **5:** Complete novice with minimal business knowledge about their ideas
- **Assistant Parameters:**
  - **Identity:** Role and background description (e.g., "experienced business advisor with small business expertise")
  - **Consistency Level:** How consistently the assistant maintains their role and expertise (0.0 = highly variable, 1.0 = perfectly consistent)
- **User Parameters:**
  - **Identity:** Role and background description (e.g., "early-stage food business entrepreneur")
  - **Focus Level (FL)**
    - **1:** Free-flowing, wide-ranging conversation covering many aspects
    - **2:** Mostly broad discussion with occasional deep dives into specific areas
    - **3:** Balanced focus with some exploration of tangential topics
    - **4:** Primarily focused on core issues with minimal tangential exploration
    - **5:** Laser-focused on specific details of implementation
  - **Prior Knowledge Level:** User's existing expertise in the domain (1 = complete novice, 2 = limited

knowledge, 3 = moderate level understanding, 4 = extensive previous experience, 5 = expert level)

  - **Decision-Making Style (DMS)**
    - **Analytical:** Focuses on data, metrics, and logical analysis
    - **Intuitive:** Relies on gut feeling and personal judgment
    - **Consultative:** Seeks multiple perspectives before deciding
    - **Risk-averse:** Prioritizes minimizing potential downsides
    - **Impulsive:** Makes quick decisions without extensive deliberation
  - **Feedback Reception (FR)**
    - **Receptive:** Eagerly accepts and builds upon advice
    - **Balanced:** Considers advice thoughtfully with moderate acceptance
    - **Skeptical:** Questions most suggestions, needs convincing
    - **Resistant:** Pushes back against most advice, difficult to persuade

Parameters controlling how knowledge is delivered and educational objectives are achieved:

- **Framework:** Educational methodology employed
  - socratic: Question-driven discovery learning
  - didactic: Direct instruction and explanation
  - collaborative: Joint problem-solving approach
  - experiential: Learning through practical examples and scenarios
- **Practical-Theoretical Balance:** Ratio of practical application to theoretical concepts (0.0 = purely theoretical, 1.0 = purely practical)
- **Complexity Progression:** Array showing how conceptual difficulty increases throughout the conversation (e.g., [0.3, 0.5, 0.7, 0.8] indicates gradual complexity increase)
- **Industry Context:** Specific sector or domain focus (e.g., "food-business", "technology", "healthcare")

Parameters governing interpersonal interactions and emotional progression:

- **Formality:** Level of professional versus casual communication (0.0 = highly casual, 1.0 = highly formal)
- **Emotional Journey:** Array of emotional states and their intensities throughout the conversation
  - Each entry contains an emotion and intensity level (0.0 = minimal, 1.0 = maximum)
  - Example: ["uncertainty": 0.8, "curiosity": 0.7, "confusion": 0.5, "understanding": 0.6, "confidence": 0.7]
- **Relationship Development:** How much the participant relationship evolves during the conversation (0.0 = static relationship, 1.0 = significant relationship building)
- **Disagreement Handling:** Approach to managing conflicting viewpoints
  - diplomatic: Respectful acknowledgment and gentle correction
  - direct: Clear, straightforward disagreement
  - avoidant: Minimizing or redirecting conflict
  - collaborative: Working together to resolve differences

Parameters controlling language use and communication style:

- **Technical Language Level:** Degree of specialized terminology and jargon (0.0 = plain language only, 1.0 = highly technical)
- **Question Types:** Distribution of different inquiry styles
  - **Closed:** Yes/no or specific factual questions
  - **Open:** Broad, exploratory questions requiring detailed responses
  - **Rhetorical:** Questions posed for emphasis rather than response
  - **Clarifying:** Questions seeking to understand or confirm information
  - Values should sum to 1.0 (e.g., "closed": 0.2, "open": 0.5, "rhetorical": 0.1, "clarifying": 0.2)
- **Response Style:** Communication characteristics
  - **Conciseness:** Brevity versus elaboration (0.0 = very verbose, 1.0 = extremely concise)
  - **Directness:** Straightforward versus indirect communication (0.0 = highly indirect, 1.0 = completely direct)
  - **Formality:** Professional versus casual language (0.0 = very casual, 1.0 = highly formal)

Parameters ensuring quality and comprehensiveness of conversation content:

- **Factual Accuracy:** Degree of correctness in information provided (0.0 = potentially inaccurate, 1.0 = verified accuracy)
- **Example Specificity:** Level of detail in illustrations and case studies (0.0 = general examples, 1.0 = highly specific, detailed examples)
- **Stakeholder Perspectives:** Array of viewpoints to be considered during the conversation (e.g., ["customer", "supplier", "regulator", "competitor"])

When generating conversations using these parameters:

1. Begin by establishing participant identities and knowledge levels
2. Follow the specified conversation arc while maintaining turn balance
3. Progress complexity according to the defined progression array
4. Incorporate emotional journey elements at appropriate conversation points
5. Ensure content addresses multiple stakeholder perspectives
6. Maintain consistency with linguistic pattern specifications
7. Adapt technical language level to participant knowledge asymmetry

Before conversation generation, validate that:

- All numerical parameters fall within specified ranges (0.0-1.0)
- Question type distributions sum to 1.0
- Turn balance ratios are mathematically consistent
- Complexity progression shows logical advancement
- Stakeholder perspectives are relevant to industry context

Generated conversations should follow this structure:

```
{
  "metadata": {
    "participantRoles": {...},
    "conversationArc": "...",
    "totalTurns": n
  },
  "conversation": [
    {
      "turn": 1,
      "speaker": "user|assistant",
      "content": "...",
      "emotionalState": "...",
      "complexityLevel": 0.x
    },
    ...
  ],
  "analysis": {
    "parameterAdherence": {...},
    "learningObjectivesMet": [...],
   "stakeholderPerspectivesCovered": [...]
  }
}
```

Here is an example input about a user's background:

```
{
  "conversationParameters": {
    "fundamentals": {
      "purpose": "advisory",
      "turns": 12,
      "turnBalance": "55:45",
      "arc": "problem-solution",
      "initiator": "user",
      "topicScope":
           ["food business",
           "marketing", "operations"]
    },
    "participants": {
      "knowledgeGapLevel": 3,
      "assistant": {
        "identity":
           "experienced business advisor",
        "consistencyLevel": 0.85
      },
      "user": {
        "identity":
            "early-stage food"
            "business entrepreneur",
        "focusLevel": 3,
        "priorKnowledgeLevel": 0.4,
      "decisionMakingStyle": "analytical",
        "feedbackReception": "receptive"
      }
    },
    "learningApproach": {
      "framework": "socratic",
      "practicalTheoreticalBalance": 0.7,
      "complexityProgression":
          [0.3, 0.5, 0.7, 0.8],
      "industryContext": "food-business"
    },
    "conversationDynamics": {
      "formality": 0.7,
      "emotionalJourney": [
        {"uncertainty": 0.8},
        {"curiosity": 0.7},
        {"understanding": 0.6},
        {"confidence": 0.7}
      ],
      "relationshipDevelopment": 0.5,
      "disagreementHandling": "diplomatic"
    },
    "linguisticPatterns": {
      "technicalLanguageLevel": 0.6,
      "questionTypes": {
        "closed": 0.2,
```

Model	Topic diversity	Topic entropy
claude	25	2.366
deepseek-r1	18	2.195
o3	27	2.493
o4-mini	33	2.880
gpt-4.1	31	2.762
gpt-4o-mini	12	1.012
gemini-2.5-pro	28	2.511
llama3.1:70b	5	0.810
claude-3.7-sonnet	35	2.985

Table 7: Topic diversity and topic entropy of baseline models.


```
        "open": 0.5,
        "rhetorical": 0.1,
        "clarifying": 0.2
      },
      "responseStyle": {
        "conciseness": 0.5,
        "directness": 0.6,
        "formality": 0.7
      }
    },
    "contentAttributes": {
      "factualAccuracy": 0.9,
      "exampleSpecificity": 0.6,
      "stakeholderPerspectives":
      ["customer", "supplier",
      "regulator", "competitor"]
    }
  }
}
```


C.3 Conversation Parameters Structure

The conversation generator operates using a hierarchical parameter system organized into six main categories: Fundamentals, Participants, Learning Approach, Conversation Dynamics, Linguistic Patterns, and Content Attributes.

D More Results

D.1 Baseline Performance Comparison

We compare the performance of different models in terms of topic diversity and topic entropy when given the baseline prompt. (Table 7). The result shows *claude-3.7-sonnet* has the best topic diversity, and smaller models like *llama3.1:70b* have the same poor performance compared to the parameterized version.

D.2 More Parameter Adherence Results

Experience Level We categorize the experience level using the prior knowledge level in the original prompt and calculate the MSE between the actual and predicted value. All models show a decrease in MSE with higher turns. (Figure 8)

D.3 Human-Evaluated Results for Table 5

We asked human raters to score the per-condition average errors corresponding to Table 5 (*Formality Only*, *Technical Only*, *Full*) across turn lengths. Errors were judged using the same rubric (Sec. A) and aggregated to match the table layout. Below we report the human-evaluated averages and alignment with the automatic metrics.

Model	Turns	Formality	Technical	Full
claude-3.7	5	0.283	0.255	0.208
claude-3.7	10	0.302	0.268	0.207
claude-3.7	15	0.300	0.260	0.193
claude-3.7	20	0.294	0.254	0.186
o3	5	0.257	0.214	0.174
o3	10	0.223	0.176	0.144
o3	15	0.216	0.163	0.132
o3	20	0.213	0.156	0.131

Table 8: Human-evaluated average performance errors corresponding to Table 5.

Correspondence with automatic metrics. We measure correspondence between Table 5 and Table 8 across all cells.

- Pearson $r = \mathbf{0.938}$ (95% CI: 0.922–0.952)
- Spearman $\rho = 0.934$ (95% CI: 0.916–0.948)
- ICC(2,k) $= 0.931$ (95% CI: 0.907–0.948)

These values indicate high alignment (> 0.93) between human ratings and the automatic metrics used in Table 5.

Feedback Reception The measurement of feedback reception is categorized into four types described in the prompt, and the result is calculated based on the rate of correct classification. The response indicates that some advanced models achieve a very high level of accuracy by combining a mixture of LLM and human decision-making, demonstrating that these models can accurately simulate the user's sentiment based on a description. Other advanced models and small models show less optimal results in this role-playing setting. (Figure 8)

While simulators can generate good responses, they may fail to create bad ones While models demonstrate clear differentiation between extreme parameter values in focus levels (Level 1 vs Level 5), they exhibit poor sensitivity to intermediate parameter settings. In Figure 7, all three models show relatively flat performance curves across the

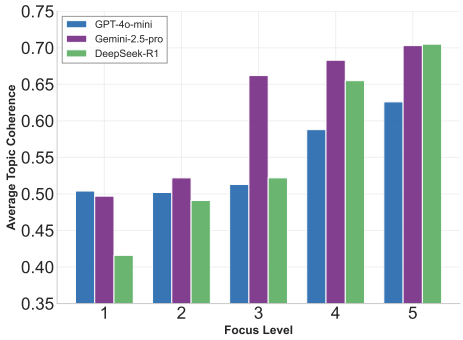


Figure 7: Average topic coherence between turns, models have parameter sensitivity issues on intermediate values, but all models can differentiate the lowest and highest value.

middle range (Levels 2-4), with topic coherence scores clustering around 0.45-0.55 regardless of the specified focus level. This suggests that models can successfully implement "very focused" versus "very unfocused" conversation styles but struggle to generate nuanced variations in between.

Similar behavior is observed in Figures 2 and 6. In Figure 2, both models show only marginally lower cosine similarity scores compared to the baseline, failing to achieve the expected degradation specified by smoothness factor F (*Highly disjointed with random topic jumping*). In Figure 6, *claude-3.7-sonnet* demonstrates minimal differentiation between knowledge gap levels 1 and 5, while *gemini-2.5-pro* exhibits comparable limitations, conflating performance across levels 3-5 despite maintaining clear separation between the extreme values (levels 1 and 5).

The insensitivity of the parameter may be due to a lack of fine-tuning. With only the definitions for each level provided to the LLM, models can only rely on pre-trained representations to map abstract parameter descriptions to concrete output behaviors. Given sufficient examples of intermediate quality levels between "highly focused" and "completely unfocused" conversations, the model could possibly provide a more distinguishable result. Further, post-training alignment procedures through RLHF further reinforce the model's tendency to produce helpful, coherent responses, creating systematic resistance to generating lower-quality content regardless of parameter specifications, which lowers the model's ability to generate poor-quality conversations.

E Supplemental Human Evaluation

E.1 A/B Topic Coherence vs. Baseline (Win-Rate)

Model	Win-rate vs. Baseline
Gemini-2.5-pro	0.8911
Claude-3.7-sonnet	0.8458
Deepseek-r1	0.8465
o3	0.8347
o4-mini	0.8410
gpt-4o-mini	0.7026

Table 9: Human A/B coherence win-rates.

E.2 Parameter Adherence (Human vs. Automatic)

Parameter	Pearson r	Human $\leftrightarrow$ Auto agreement
Focus	0.914	ICC=0.901
Smoothness	0.944	ICC=0.932
KGL	0.931	ICC=0.919
Formality	0.953	ICC=0.941
Technicality	0.946	ICC=0.934
Decision Style	–	$\kappa = 0.918$, macro-F1=0.882

Table 10: Human–automatic correspondence by parameter.

E.3 KGL Sensitivity (Human Recovery Accuracy)

Model	Exact match	± 1 tolerance
Gemini-2.5-pro	0.6314	0.9108
Claude-3.7-sonnet	0.6025	0.8969
Deepseek-r1	0.6122	0.8703
o3	0.5867	0.8668
o4-mini	0.6160	0.8826
gpt-4o-mini	0.4701	0.7995
Baseline	0.3846	0.7321

Table 11: Human inference of KGL from transcripts.

F Evaluation Tasks

Here, we first introduce our evaluation tasks and explain the methods in Section 3.

Topic Diversity The conversation needs a topic to start. After setting the topic area before the simulation, LLM will pick a subtopic based on the configured parameters to best suit the entrepreneur's background. In this task, we compare the distributions of topics mentioned by the simulator.

Parameter Adherence To evaluate whether the conversation generated follows the given parameters, we evaluate the difference between the settled

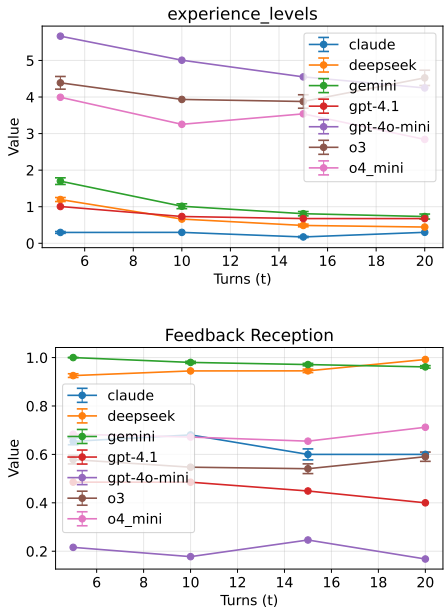


Figure 8: Additional figures on parameter adherence

parameters vs. the inferred parameters given only the generated conversation.

Topic Drift Natural dialogue often involves gradual topic transitions that can lead to substantial drift from the original subject matter, making thematic coherence throughout extended conversations a challenge. We measure the semantic distance between conversation segments to quantify how far the dialogue deviates from its initial topic focus. We calculate sentence embedding to compute cosine similarity scores between the opening conversational topic and subsequent dialogues, tracking the drift over turns.

Character Properties Stability Consistent character portrayal across conversation turns is essential for believable dialogues, yet current LLMs often exhibit personality inconsistencies that undermine conversation quality. This evaluation measures character stability by analyzing linguistic markers, decision-making patterns, and domain expertise demonstrations throughout generated conversations. We measure deviations between the character's behavior in conversation versus their given background or parameters.

Entity Revisit Rate Effective conversations demonstrate sophisticated information management by strategically reintroducing previously mentioned entities, concepts, and topics, creating coherent narrative threads rather than generating unrelated information. We quantify how frequently and effectively the conversation simulator references earlier elements by tracking named entities and key concepts from earlier turns, then analyzing whether their subsequent appearances serve meaningful conversational purposes.