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 represents a transformative class of artificial intelligence systems capable of autonomously producing diverse content based on patterns learned from large-scale data and guided by user prompts. These models can generate coherent and contextually relevant text (OpenAI et al., 2024b), synthesize photorealistic images (Rombach et al., 2022), compose original music (Copet et al., 2023), produce functional source code (Chen et al., 2021), and design 3D models and environments (Poole et al., 2022). Their generative capacity extends beyond creative tasks, with applications in scientific domains, such as predicting protein structures with atomic accuracy (Jumper et al., 2021) and assisting in the formulation of mathematical proofs (Drori et al., 2022). This versatility has made generative AI a central technology in both creative industries and scientific research.

We believe that in future, we will see the rise of **parametrically controlled LLMs** that are tuned to perform specific complex tasks, and will allow for finer-grained control of their behavior through a set of parameters instead of relying solely on natural language instructions. In this paper, we investigate a specific case to illustrate the point - the example task is that of generating *realistic end-to-end multi-turn conversation*s using large language models (LLMs) as a means to simulate dialogue episodes *in a given thematic area*. These simulated conversations would serve as structured training material that can improve downstream conversational AI applications, particularly in settings where data is scarce, human annotation is costly, or domain specificity is high. Generating whole conversations, rather than isolated responses, enables the development of systems that better capture context, discourse coherence, and speaker intent over extended interactions (Zhang et al., 2020; Roller et al., 2021).

Such simulators are increasingly important in real-world research areas ranging from healthcare and education to business advising and civic services. For instance, in training AI systems to support low-resource users—such as entrepreneurs seeking regulatory or startup guidance—few-shot or domain-specific conversations are essential, but often unavailable (Li et al., 2023). Simulated dialogues can bridge this gap by providing varied, context-rich conversations tailored to user needs and grounded in realistic scenarios (Huang et al., 2020), and allow researchers to probe system behaviors in a controlled manner—enabling stress testing for safety, bias, and human-centric factors (Bender et al., 2021) like cognitive overload. Realistic multi-turn conversation generation is not just a technical convenience—it is emerging as a core methodology for training and evaluating next-generation dialogue systems.

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.

The need for parameterized conversation control is particularly acute in domains requiring high-quality simulated dialogues, such as training data generation for conversational AI systems (Li et al., 2016a), educational dialogue design (Nye et al., 2014), therapeutic conversation modeling (Vaidyam et al., 2019), and realistic character interactions in entertainment applications (Shuster et al., 2022; Urbanek et al., 2019). Recent work by (Zheng et al., 2023) demonstrates that conversation quality assessment is multidimensional, yet current generation approaches lack explicit control over these dimensions. While current LLMs can generate plausible conversations, they face several challenges that our parameterization approach directly addresses:

1. **Challenges in conversation quality control**
   Structural Coherence: LLMs demonstrate documented difficulties maintaining consistency across extended dialogues. Research by (Gao et al., 2018) confirms a deterioration in response quality as conversation history increases, while (Xu et al., 2022) identifies specific challenges in entity tracking and resolution of coreference over multiple turns. More recent studies by (Dziri et al., 2022) quantify inconsistencies in model-generated dialogues, showing that even state-of-the-art models exhibit significant contradiction rates. Our framework addresses these issues through explicit parameters for narrative coherence, memory utilization, and contradiction detection, building on techniques from computational narratology (Mani, 2012).

   Knowledge Progression: Studies by (Kim et al., 2020) show that effective knowledge transfer in educational dialogues requires careful calibration of complexity progression. Our parameters for explanation progression, conceptual density, and learning framework provide fine-grained control over knowledge transfer dynamics, drawing on established pedagogical frameworks (Bloom, 1956; Anderson et al., 2001) and cognitive load theory (Sweller et al., 2011).

   Character Consistency: Current approaches struggle to maintain consistent character voices and knowledge states throughout extended conversations. (Li et al., 2016b) and (Zhang et al., 2018) demonstrate that explicit persona modeling improves response consistency, but challenges persist in maintaining these personas across turns. Our parameterization includes explicit controls for character consistency, knowledge asymmetry, and backstory depth to address these challenges, incorporating insights from computational models of personality (Mairesse et al., 2007) and literary character development (Bamman et al., 2013, 2014).

2. **Challenges in conversation generation methodology**

   Control Granularity: Existing approaches typically offer coarse-grained control through natural language instructions, which can be ambiguous and inconsistently interpreted by models (Mishra et al., 2022). Recent work by (Min et al., 2022) shows significant variance in how models interpret the same natural language instructions. Our parameterization aim to provide control over conversation properties, similar to approaches in other generative domains such as text-to-image generation (Nichol et al., 2022) and music synthesis (Agostinelli et al., 2023).

   Theoretical Grounding: Current conversation generation approaches often lack connection to established theoretical frameworks in linguistics and dialogue management. (Larsson and Traum, 2000) and (Traum and Larsson, 2003) provide formal models of tracking the state of dialogue that have not been fully used in the generation of neural conversations. Our parameter set establishes formal connections to speech act theory (Searle, 1969; Austin, 1975), information theory (Shannon, 1948), computational narratology (Mani, 2012), and dialogue management models (Young et al., 2013; Williams et al., 2016), creating a bridge between neural approaches and classical dialogue system theory.

   Evaluation Framework: (Deriu et al., 2020)

identifies significant gaps in conversation evaluation methodologies, a finding echoed by (Mehri and Eskenazi, 2020), who demonstrate poor correlation between automated metrics and human judgments of conversation quality. (See et al., 2019) further shows that human quality assessments depend on multiple dimensions that current automatic metrics do not capture comprehensively. Our parameterization approach enables systematic variation of conversation properties, facilitating controlled experiments to assess quality dimensions and potentially leading to more nuanced evaluation methodologies.

The key contributions of this paper are:

1. A comprehensive taxonomy of 35 conversation parameters with 9 dominating factors organized into six dimensions that capture the essential aspects of high-quality conversations, extending prior work on dialogue quality factors (See et al., 2019; Mehri and Eskenazi, 2020)

2. Analysis of parameter necessity and sufficiency, identifying a core set of essential parameters while eliminating redundancies, informed by dimensionality reduction approaches to conversation modeling (Larochelle et al., 2009; Lowe et al., 2018)

3. Formal theoretical connections between our parameters and established models in computational linguistics (Jurafsky and Martin, 2000), dialogue management (Young et al., 2013), and information theory (Xu et al., 2020), creating a bridge between neural approaches and classical dialogue system theory

4. Preliminary experimental validation demonstrating how modern LLMs can effectively implement a subset of these parameters through prompt conditioning, building on recent advances in controlled text generation (Khalifa et al., 2021; Yang and Klein, 2021)

5. A proposed research agenda for implementing the full parameter set through architectural modifications (Hu et al., 2017; Keskar et al., 2019), developing efficient parameter encoding methods (Li and Liang, 2021; Lester et al., 2021), and creating benchmark datasets (Welleck et al., 2019; Dziri et al., 2020)

Our methodology combines computational approaches with insights from linguistics, psychology, and education. We evaluate our framework through a series of controlled experiments comparing conversations generated with systematically varied parameter settings. Results demonstrate statistically significant differences in generated conversation properties when parameter values are manipulated, confirming the effectiveness of our approach for a subset of parameters. For parameters that current LLMs struggle to implement reliably, we provide a detailed analysis of limitations and propose architectural modifications to address these challenges.

Our parameterization framework represents a significant step toward more controllable, higher-quality conversation generation with LLMs. By providing a standardized approach to conversation quality control, we aim to influence the theoretical understanding and practical capabilities of conversational AI systems.

2 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 one sample per prompt per 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 E. Human protocols and agreement statistics are summarized in the main text and fully detailed in Appendix A.

For each task, we create a set of simulators and control the parameters to generate task-specific conversations 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. The evaluation is done by manually eliminating similar topics from the generated results. 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 using the mean squared error: $\frac{1}{n}\sum \text{set value} - \text{inferred value}$. Categorical parameters are evaluated using multi-class classification accuracy, where correct classifications receive a score of 1 and incorrect classifications receive a score of 0.

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.

This evaluation measures the semantic distance between conversation segments and the initial topic focus using sentence embedding techniques. We employ BERT-based sentence embeddings to compute cosine similarity scores: $1 - \cos(\text{embedding}(\text{utterance}_i) - \text{embedding}(\text{utterance}_0))$ between the entrepreneur's utterances at each turn and the main business topic established in the conversation opening.

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.

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. Full task definitions, rater rubrics, sampling details, and statistical procedures appear in Appendix A.

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.

3 Experiments

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 4. Advanced models such as *Gemini-2.5-pro* and *DeepSeek-R1* exhibit superior topic diversification capabilities, generating 141 and 143 distinct topics, respectively, with corresponding entropy values of 5.266 and 5.275. 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, highlighting the need for structured parameter control. 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 conversa-

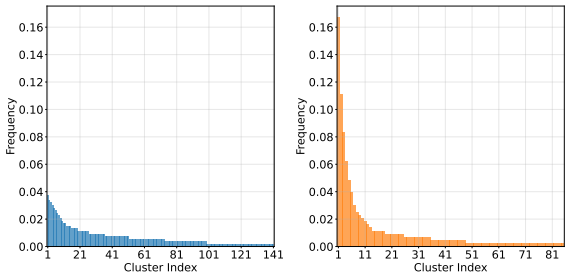


Figure 1: Topic frequency distributions of the *gpt-4o-mini* (orange) and *gemini-2.5-pro* (blue). Clusters are sorted in descending frequency. More advanced model (gemini-2.5-pro) produced a more diverse topic compared to the less advanced model (gpt-4o-mini)

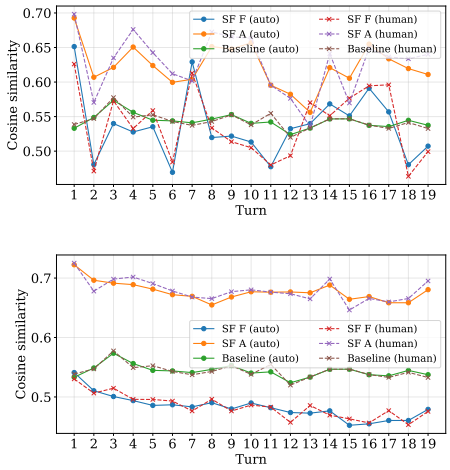


Figure 2: The cosine similarity between the entrepreneur's utterance to the main topic in different smoothness factors for two models: (a) *gpt-4o-mini*, (b) *claude-3.7-sonnet*. *claude-3.7-sonnet* is showing a high separation between the highest and lowest smoothness factor, showing better understanding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

tional flow.

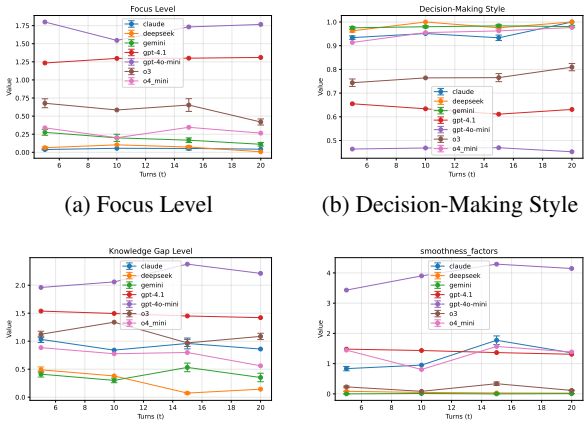
We also examine sentence diversity by calculating semantic diversity through the cosine similarity of embeddings generated by all-MiniLM-L6-v2. (Table 1). The embedding diversity rankings partially diverge from topic-level diversity measures, suggesting that models may employ different strategies for 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 accuracy, where advanced models achieve 0.8-1.0 accuracy rates while lighter models like *gpt-4o-mini* struggle to maintain consistent classification performance, often hovering around 0.4-0.6 accuracy. The smoothness factor analysis (Figure 3(d)) demonstrates that parameter control effectiveness varies significantly 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 relation-



(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. D

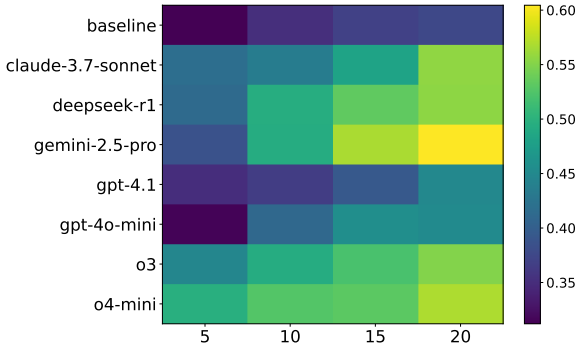


Figure 4: Concept-revisit rate by turns for each model with knowledge gap level of user set to 1 (most knowledgeable). All models exhibit a higher revisit rate with turn progression.

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 3: 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. D for full human results and agreement analyses

ship 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 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 3, 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.

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

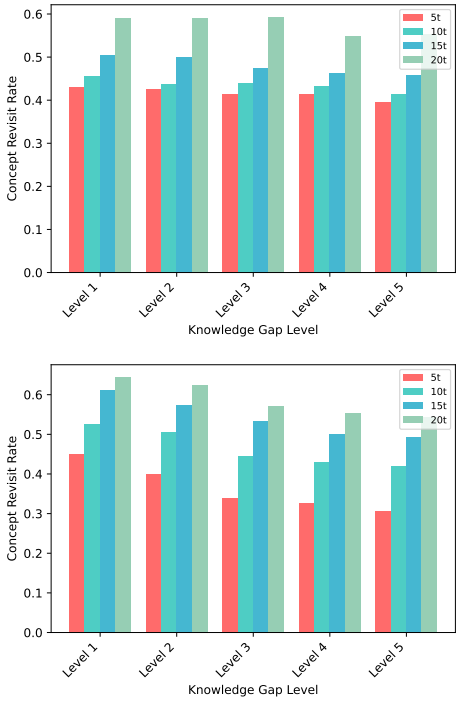


Figure 5: Concept-revisit rate by knowledge-gap level for two models: (a) Claude, (b) Gemini. Knowledge Gap Level 1 is the smallest knowledge gap, and Knowledge Gap Level 5 is the highest. *Gemini-2.5-pro* shows a more significant difference when modifying Knowledge Gap Level.

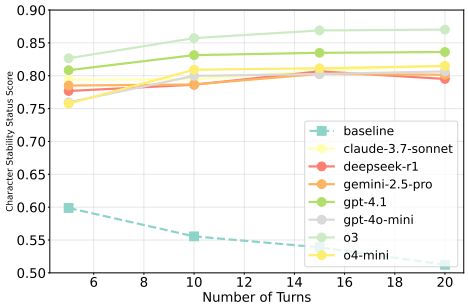


Figure 6: Character Parameter stability over turns, the baseline has a decreasing stability over turns, while all other models with character properties show an increase in stability score. The agreement between the human and LLM judge at baseline is κ = **0.9316**, and κ = **0.9241** for *claude-3.7-sonnet*

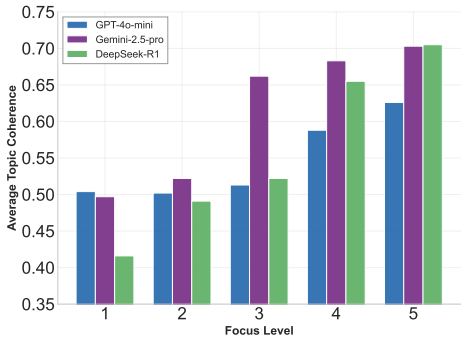


Figure 7: Average topic coherence between turns, models have parameter sensitivity issues on intermediate values, but all models can differentiate the lowest and highest value.

show relatively flat performance curves across the 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.

4 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.

5 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.

6 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.

7 LLM usage

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

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, and 1 others. 2023. Musiclm: Generating music from text.

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.

Lorin W. Anderson, David R. Krathwohl, Peter W. Airasian, Kathleen A. Cruikshank, Richard E. Mayer, Paul R. Pintrich, James Raths, and Merlin C. Wittrock. 2001. A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives. *Educational Horizons*, 83(3):154–159.

John L. Austin. 1975. *How to Do Things with Words*. Oxford University Press.

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.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. pages 352–361.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A bayesian mixed effects model of literary character. pages 370–379.

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.

Benjamin S. Bloom. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. Longmans, Green.

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.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36:47704–47720.

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.

Iddo Drori, Akshay Kharkar, William R. Sickinger, Eden Kochman, Suwen P. Ng, Kavi Hadash, Yilu Ge, Joshua B. Tenenbaum, Ce Liu, and Hendrik Strobelt. 2022. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. 119(32):e2123433119.

Nouha Dziri, Ehsan Kamalloo, Kory W. Mathewson, and Osmar Zaiane. 2020. Evaluating coherence in dialogue systems using entailment.

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.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. pages 1371–1374.

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.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. 70:1587–1596.

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

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, and 1 others. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

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.

Yoojung Kim, Tolga Soyata, and Razieh Fatemeh Behnagh. 2020. Designing adaptive conversational agent tutors to enhance computer science learning through dialogue-based scaffolding. *Computers in Human Behavior*, 113:106496.

Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin. 2009. Exploring strategies for training deep neural networks. *Journal of machine learning research*, 10(1).

Staffan Larsson and David R. Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6(3-4):323–340.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models.

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

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation.

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*.

Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2018. Towards an automatic turing test: Learning to evaluate dialogue responses.

François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. 30:457–500.

Inderjeet Mani. 2012. *Computational Modeling of Narrative*. Morgan & Claypool Publishers.

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

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? pages 7102–7113.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to GPTk's language. *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models.

Benjamin D. Nye, Arthur C. Graesser, and Xiangen Hu. 2014. AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24(4):427–469.

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.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion.

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.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

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

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, and 1 others. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage.

John Sweller, Jeroen J. G. Van Merrienboer, and Fred G. W. C. Paas. 2011. Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4):295–312.

David R Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In *Current and new directions in discourse and dialogue*, pages 325–353. Springer.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. pages 673–683.

Ajoy N. Vaidyam, Hannah Wisniewski, John D. Halamka, Matcheri S. Kashavan, and John B. Torous. 2019. Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference.

Jason D. Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, and Deepak Ramachandran. 2016. The dialog state tracking challenge series: A review. 8(1):1–33.

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

Ying Xu, Zhiting Cao, and Gonzalo G. de Polavieja. 2020. A theory of usable information under computational constraints. *Entropy*, 22(9):1014.

Kevin Yang and Dan Klein. 2021. Fudge: Controlled text generation with future discriminators.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. POMDP-based statistical spoken dialog systems: A review. 101(5):1160–1179.

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.

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 $N{=}5$ 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.

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 Prompts

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

B.1 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>.

B.2 Parameterized Prompt

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

B.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.

B.4 Fundamentals

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"])

B.5 Participants

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

B.6 Learning Approach

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")

B.7 Conversation Dynamics

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

B.8 Linguistic Patterns

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)

B.9 Content Attributes

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"])

B.10 Implementation Guidelines

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

B.11 Parameter Validation

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

B.12 Output Format

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,
        "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 More Results

C.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 4). 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.

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 4: Topic diversity and topic entropy of baseline models.

C.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)

C.3 Human-Evaluated Results for Table 3

We asked human raters to score the per-condition average errors corresponding to Table 3 (*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 5: Human-evaluated average performance errors corresponding to Table 3.

Correspondence with automatic metrics. We measure correspondence between Table 3 and Table 5 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 3.

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)

D Supplemental Human Evaluation

D.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 6: Human A/B coherence win-rates.

D.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 7: Human–automatic correspondence by parameter.

D.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 8: Human inference of KGL from transcripts.

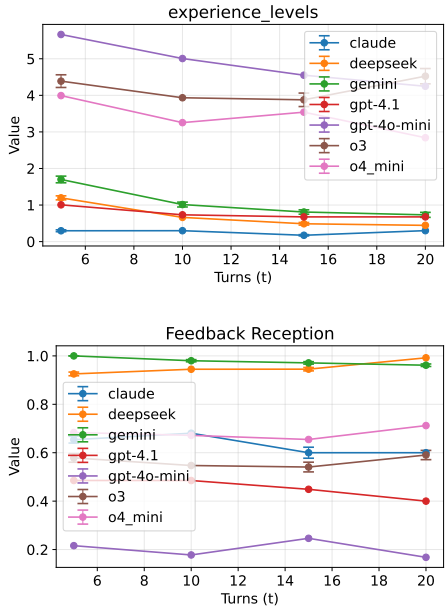


Figure 8: Additional figures on parameter adherence

E 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 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.