MultiChallenge: A Realistic Multi-Turn Conversation Evaluation Benchmark Challenging to Frontier LLMs

Anonymous ACL submission

Abstract

We present MultiChallenge, a pioneering benchmark evaluating large language models (LLMs) on conducting multi-turn conversations with human users, a crucial yet underexamined capability for their applications. MultiChallenge identifies four categories of challenges in multi-turn conversations that are not only common and realistic among current human-LLM interactions, but are also challenging to all current frontier LLMs. All 4 challenges require accurate instruction-following, context allocation, and in-context reasoning at the same time. We also develop LLM as judge with instance-level rubrics to facilitate an automatic evaluation method with fair agreement with experienced human raters. Despite achieving near perfect scores on existing multi-turn evaluation benchmarks, all frontier models have less than 50% accuracy on MultiChallenge, with the top-performing Claude 3.5 Sonnet (October 2024) achieving just a 41.4% average accuracy.

1 Introduction

Evaluation benchmarks have played a critical role in the evolution of large language models (Ziyu et al., 2023; Porcu and Havlínová, 2024), with benchmarks being continuously released to test large language models (LLMs) on various tasks (Zhang et al., 2024; Jain et al., 2024; Glazer et al., 2024). One such task is multi-turn conversation with humans, a common yet challenging endeavor. This is because conducting multi-turn conversations requires not only accurate instruction following, but also careful attention allocation to conversation context, and more importantly, near or surpassing human in-context reasoning capability.

Despite the growing demand and complexity of multi-turn LLM-human conversations, comprehensive evaluation frameworks for this task remain limited (Kwan et al., 2024; Wang et al., 2023). Some of them, such as the widely adopted MT-Bench (Zheng et al., 2023), are saturated by frontier LLMs with near-perfect results (Achiam et al., 2023; Yang et al., 2024). Others (He et al., 2024) focus more on multi-turn explicit instruction following, missing the opportunity to assess the actual set of mixed model capabilities required in conducting natural multi-turn LLM-human conversations.

To bridge this gap, we build MultiChallenge, a realistic and challenging benchmark designed to evaluate LLMs' ability to respond appropriately in multi-turn conversations. The benchmark consists of conversation histories of up to 10 turns between two parties, ending in a final user turn that contains a requirement or question. LLMs must generate an appropriate response to this final turn, taking into account the entire conversation context. We identify four challenge categories in multi-turn conversations that are not only common and realistic among current human-LLM interactions, but also difficult for current frontier LLMs. They include instruction retention, inference memory of user information, reliable versioned editing, and self-coherence.

Instruction retention evaluates whether LLMs are able to follow instructions specified in the first user turn throughout the entire multi-turn conversation. Inference memory of user information evaluates LLMs on recalling and connecting relevant details scattered in previous user turns when they are implicitly required to respond to the final user turn. Reliable versioned editing evaluates whether LLMs can properly help humans revise existing materials through back-and-forth iterations with human users. Finally, self-coherence evaluates whether LLMs can be reasonably coherent with model responses in the conversation history and avoid sycophancy (unconditionally agreeing to human users). We elaborate on the 4 challenges in Section 3 and show examples of them in Figures 1 and 2. All 4 challenges require not only instruction following, but accurate context allocation and

in-context reasoning at the same time.

We adopt a hybrid approach in which human raters and LLMs work together, to produce diverse, realistic, and challenging multi-turn test examples. We also develop an LLM-as-a-judge system with instance-level rubrics to facilitate an automatic evaluation method with good agreement with experienced human raters. Results from both human and automatic evaluation show that all frontier models have less than 50% accuracy on `MultiChallenge`, despite reaching near perfect scores on existing multi-turn evaluation benchmarks (Achiam et al., 2023; Yang et al., 2024), indicating the difficulty and usefulness of `MultiChallenge` in multi-turn conversation evaluation. Among them, Claude 3.5 Sonnet (October 2024) achieves the best general performance, reaching a 41.4% average accuracy score on `MultiChallenge`. Our experiments also show high correlation between human raters and our automatic judgments with instance-level rubrics, demonstrating the potential for quick and accurate multi-turn evaluations. We attach the data and code of this benchmark in the supplement materials.

2 Related Work

Evaluation benchmarks on multi-turn conversations. While benchmarks have rapidly emerged to assess frontier capabilities of LLMs, such as AI as agents (Liu et al., 2023; Yao et al., 2024) and multi-modality (Liu et al., 2025; Li et al., 2024), evaluating multi-turn conversational capabilities has not received enough attention. Existing benchmarks like MT-Bench (Zheng et al., 2023) and MT-Eval (Kwan et al., 2024) no longer discriminate effectively among frontier LLMs, with models such as GPT-4, Claude 3, and LLama3 achieving near-perfect scores. More recently, multi-IF (He et al., 2024) extends the IFEval benchmark (Zhou et al., 2023) to multi-turn and multilingual scenarios. It focuses more on the auto-verifiable format and instruction following in multi-turn settings, such as maintaining all uppercase text, instead of the mixed set of model capabilities required in realistic multi-turn conversations such as in-context reasoning.

Multi-agent synthetic data generation. Multi-agent frameworks have emerged as a popular means to create diverse, high-quality synthetic datasets. AgentInstruct (Mitra et al., 2024) demonstrates this by utilizing multiple LLM-powered agents in three agentic flows to generate synthetic text data with minimal human intervention. Similar frameworks in other domains include Splunk's MAG-V (Sengupta et al., 2024) for customer data and MATRIX (Tang et al., 2024) for social dynamics data. In this work, we employ a multi-agent method to generate initial synthetic multi-turn conversations for 4 `MultiChallenge` challenges, followed by a hybrid approach with human editing and reviewing to produce the final benchmark.

3 `MultiChallenge`

In `MultiChallenge`, we focus on four challenge categories commonly encountered in realistic multi-turn human-LLM interactions, which we found difficult for current frontier LLMs. In this section, we provide specific definitions and test examples for each category and also outline our automatic evaluation method, enabling efficient and trustworthy adoption of `MultiChallenge`. Comprehensive benchmark statistics are presented in Table 1.

3.1 Challenge Categories

3.1.1 Instruction Retention.

Human users often provide general instructions at the beginning of conversations with LLMs, expecting them to be followed throughout. In practice, we find that while frontier LLMs excel at following instructions in single-turn conversations, their ability to retain instructions in multi-turn conversations is still limited. We evaluate this capability under Instruction Retention. The first panel in Figure 1 shows an example. In this example, the constraint on films with specific UK age ratings is given in the first turn. The model is evaluated on whether its final response adheres to this initial constraint.

In every instruction retention test example, the human user explicitly specifies in the first turn that the instructions should be followed throughout the entire conversation. Moreover, human users don't provide followup instructions that contradict the first-turn instructions. The first-turn user instructions also range from simple format constraints such as sentence numbers and bullet point formats, to more semantic instructions such as movie ratings in the example above.

3.1.2 Inference memory (of user information).

Test conversations under this category mainly assess an LLM's capability to recall and connect relevant details scattered in previous user turns, when they are implicitly required when responding to the

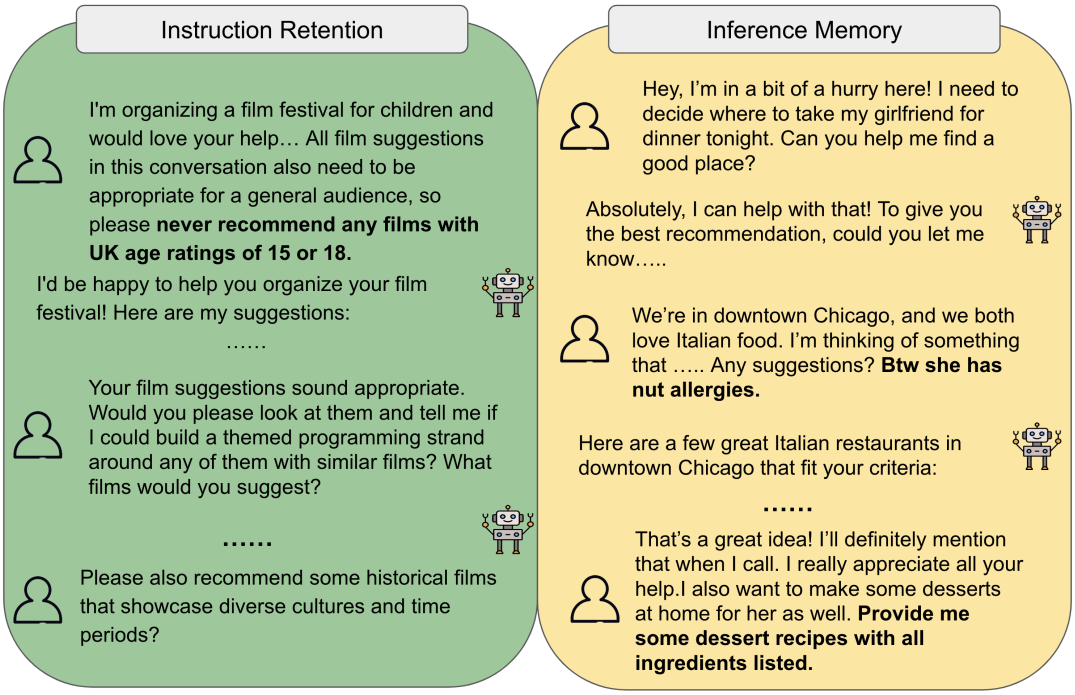

Figure 1: Test examples under Instruction Retention and Inference Memory. The highlighted text under Instruction Retention is the first-turn constraint that should be followed by LLMs for the whole conversation. The highlighted text under Inference Memory is the user information that should be recalled and connected when responding to the final user turn. See Sections 3.1.1 and 3.1.2 for details.

final user turn. The second panel of Figure 1 shows an example of this category. Some user information that appeared early in the conversation, *nut allergies*, is implicitly required to successfully complete the final user request for *dessert recipes*. In such examples, the LLM is evaluated on whether it is able to reallocate its attention to the relevant parts of the conversation history and provide *dessert recipes* that avoid triggering *nut allergies*.

When creating inference memory test examples, we ensure that final user turns don't directly ask for previously mentioned information (e.g., "Does my girlfriend have nut allergies?"). Instead, we make the information requirement implicit to test LLMs' ability to reason about the relevance of contextual information.

3.1.3 Reliable versioned editing.

Another common and natural use case of LLMs is to help humans revise existing materials, such as travel plans, emails, code, etc. This editing often involves multiple turns between the user and LLM, resulting in various versions of edited materials based on different instructions. Users may frequently change opinions, abandon current edits, and revert to earlier versions. The first panel in Figure 2 shows a test example, where the user adjusted the schedule multiple times and referred to different versions of it. This collaborative editing with version references requires the LLM to successfully resolve ambiguities in anaphoric expressions (e.g., 'the plan we had before we adjusted the workshop's ending time'), accurately retrieve and copy the correct version without hallucination, and edit per user instructions. This again requires precise context attention, reasoning, and instruction following from LLMs.

3.1.4 Self-coherence.

Frontier LLMs may cater to human preferences by always agreeing with the user, a behavior known as sycophancy (Perez et al., 2022; Sharma et al., 2023). This is particularly evident in multi-turn conversations when the LLM quickly abandons its previous responses and agrees with the user when the user repeats the query or slightly questions prior LLM responses (Laban et al., 2023; Agarwal et al., 2024). This self-contradicting behavior of LLMs can make the conversation flow unnatural and reduce trust with human users.

Therefore, under self-coherence, we evaluate LLMs ability to maintain reasonable and natural consistency with their previous responses. Figure 2 gives an example. In a multi-turn discussion

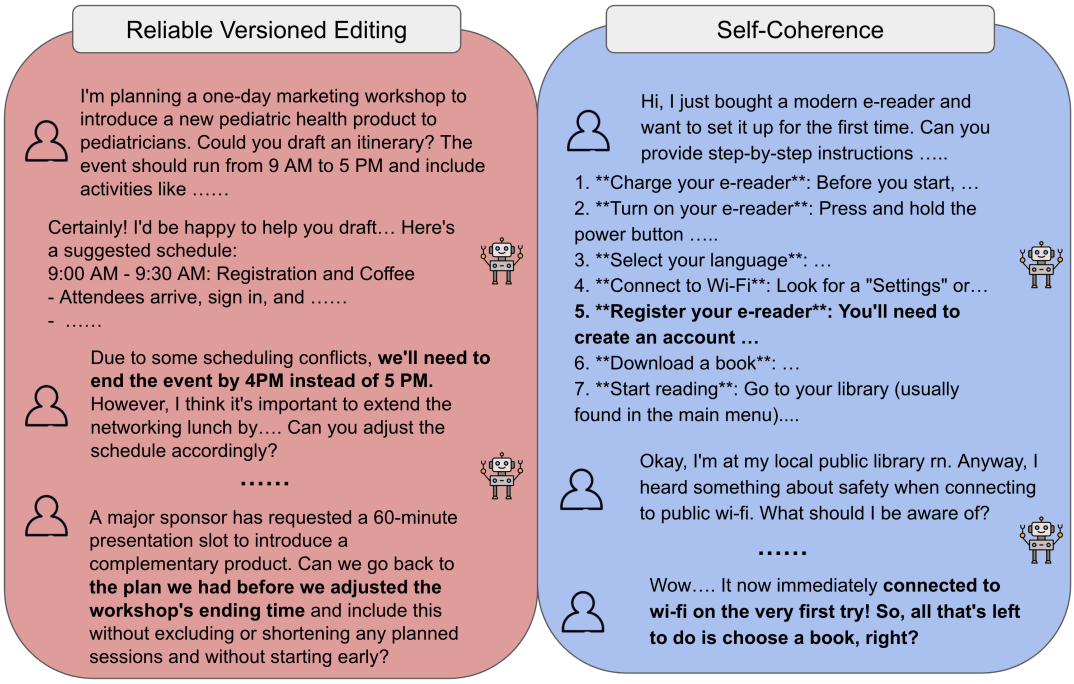

Figure 2: Test examples under Reliable Versioned Editing and Self-Coherence. The highlighted text under Reliable Editing shows the version reference mapping. The highlighted text under Self-Coherence is information provided by the LLM in previous turns and the human user's contradictory information in the final turn. See Sections 3.1.3 and 3.1.4 for details.

about how to set up an e-reader, the LLM gives step by step guidance which includes *Register your e-reader* after *connecting to Wi-Fi* in an early turn. However, when the human user asks again in the final turn "*all that's left to do now since it is connected to wifi is choose a book, right?*", most frontier LLMs switch to agreeing with the user, contradicting their previous response that registration follows connecting to Wi-Fi.

Attribute	Value
Challenge Categories	4
Test Conversations	273
Inference Memory	113
Instruction Retention	69
Reliable Version Editing	41
Self-Coherence	50
Avg. Turns/Conversation	5
Avg. Words/Conversation	1231.7

Table 1: `MultiChallenge` statistics.

3.2 Automatic evaluation with instance-level rubrics

Rule-based automatic evaluation is unsuitable for `MultiChallenge` due to the lack of single ground-truth answers in most test conversations. Moreover, we have found that directly using frontier LLMs as judges by providing them full conversation history and final response answer, yields low alignment with human raters. This is potentially because frontier LLMs like GPT-4o have under 50% accuracy on `MultiChallenge`, limiting our confidence in their judging ability. Pure human evaluation on `MultiChallenge` remains expensive and time-consuming.

Therefore, we propose LLM as judge with instance-level rubrics to facilitate an automatic evaluation method with fair agreement with experienced human raters. For each test example, human raters provide a binary rubric question answerable with only "yes" or "no". This question requires only the final model response as context and is designed within current LLMs' capabilities. A "yes" answer indicates that the model response has passed this test example. For example, the binary rubric question for the inference memory example shown in Figure 1 is "Do any of the dessert recipes suggested in this response contain nuts?".

This method makes automatic evaluation on `MultiChallenge` possible and reliable. Experiments show that adopting frontier models as judges with our instance-level rubrics reaches 93% alignment with experienced human raters, compared to 36% alignment of directly prompting LLMs as

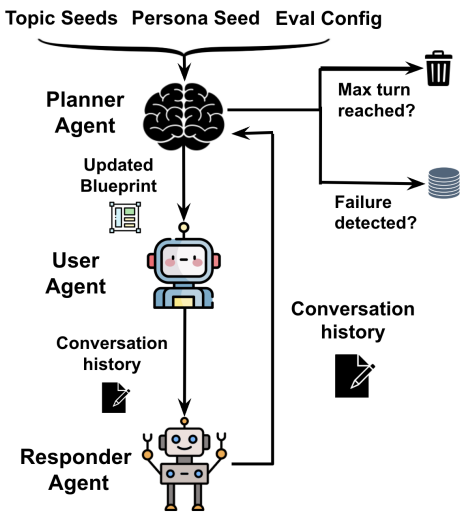

Figure 3: The multi-agent work flow of MMSE to synthetically generate candidate test examples for MultiChallenge.

judge by providing raw conversation context. Details of this experiment are shown in Section 5.2.

4 The Hybrid Approach to build MultiChallenge

Producing realistic, diverse, and challenging test examples for MultiChallenge that can make most frontier LLMs fail, is difficult and time-consuming even for human experts. To facilitate this process and reduce costs while maintaining quality, we use a hybrid approach: initially generating synthetic data, then having experts review and edit it. In this section, we first describe our multi-agent synthetic data generation system and then illustrate the human editing process.

4.1 Synthetic Data Generation

We adopt a **m**ulti-stage and **m**ulti-agent **s**ystem for generating **e**valuation examples in MultiChallenge, MMSE. When generating a test example, MMSE takes in 3 inputs – hierarchical topic seeds, the persona seed, and an evaluation config of the specific challenge category – as shown in Figure 3.

Hierarchical topic seeds provide structured topics covering various knowledge domains. The full hierarchical topic taxonomy is in Appendix A.1. The persona seed, sampled from HuggingFace PersonaHub (Ge et al., 2024), provides diverse user personas, making our synthetic conversations reflective of real-world scenarios. The topic and persona seeds both boost the diversity of MultiChallenge. The third input, the category-specific config, contains 5 components, namely category name, category definition, pass criteria, failure criteria and K-shots of failures. Appendix A.2 shows configurations for all 4 challenge categories. These configurations guide the *Planner Agent* in MMSE to generate conversation backbones and preliminarily evaluate LLM failures.

As a multi-agent system, MMSE involves 3 types of agents serving different roles to collaborate on generating test samples. The 3 types of agents are *Planner Agent*, *User Agent* and *Responder Agents*. All agents are constructed by prompting LLMs with their corresponding agent role descriptions and task descriptions. Appendix A.3 presents the system prompts for all agents.

Planner Agent solely orchestrates the synthetic conversation, generating and updating the backbone and strategy for the *User Agent* based on given inputs. *User Agent* acts as an LLM user, translating the planner's strategy into concrete, realistic user turns to communicate with the responder agent, aiming to elicit mistakes that fit the corresponding category definition.

Responder Agents serve the role of AI assistants, responding to *User Agent* turns in multi-turn conversations. They generate responses based on current conversation history. To ensure unbiased generation, we randomly sample from 6 frontier models for each test example: o1-preview (Jaech et al., 2024), GPT-4o (August 2024) (Hurst et al., 2024), Gemini 1.5 Pro (August 27, 2024) (Team et al., 2024), Claude 3.5 Sonnet (October 2024) (Anthropic, 2024), Mistral Large (November 2024) (Mistral AI, 2024), and Llama 3.1 405B Instruct (Dubey et al., 2024). This approach prevents overfitting to any particular model's weaknesses or characteristics.

Figure 3 outlines the workflow for MMSE's 3 agents. The planner generates an initial **conversation blueprint** from the input triplet (example in Appendix A.7). The *User Agent* creates the first concrete user turn from this. As the conversation progresses, the planner updates the blueprint based on current history and checks for failures. Upon detecting a failure, it saves the conversation for human editing. Otherwise, the planner refines the blueprint with current context. This continues until a failure is identified or the turn limit is reached, at which point the conversation is discarded.

4.2 Human Review and Editing

We recruit and train human annotators to review and edit the data generated by MMSE for final test examples in MultiChallenge. The human review process mainly assesses 3 aspects of data quality, a) alignment with challenge category definition; b) naturalness and realism of conversation; c) reasonable failure of frontier LLMs. We accept test examples causing at least 3 of the 6 LLMs mentioned above to fail, ensuring challenge representativeness. If quality is unsatisfactory in any criteria, annotators either edit or discard the synthetic example.

We provide thorough guidance to human annotators, including examples of unfair failures per category, editing tips, and detailed workflow descriptions. We also have 2 review layers after a human annotator submits their edited examples. The reviewer pool is completely different from the initial editor pool. We only accept examples that pass both review layers. We recruit human annotaters within US in a non-biased way and pay them fairly.

5 Experiments

In this section, we first present the human evaluated performance of 6 frontier models, o1-preview (Jaech et al., 2024), GPT 4o (August 2024) (Hurst et al., 2024), Gemini 1.5 Pro (August 27, 2024) (Team et al., 2024), Claude 3.5 Sonnet (October 2024) (Anthropic, 2024), Mistral Large (November 2024) (Mistral AI, 2024), and finally Llama 3.1 405B Instruct (Dubey et al., 2024). These 6 LLMs are the same ones that we used to collect test examples. We analyze their performance across 4 challenge categories. We also use our automatic evaluation method for most open-source LLMs. Lastly, we provide analysis and case studies on MultiChallenge, such as difficulty distribution and auto-eval trustworthiness.

5.1 Main Results

Table 2 shows the human evaluated performance of all 6 frontier LLMs listed above. Inference settings for all models we evaluated can be found in Appendix A.5. Human raters review the 6 anonymous model responses of a test example at the same time followed by 2 reviewer layers. Among them, Claude 3.5 Sonnet (October 2024) achieves the best general performance on conducting multi-turn conversations with human users, reaching a 41.4% average accuracy score on MultiChallenge, significantly outperforming other LLMs. o1-preview follows with 37.2, also significantly outperforming the rest LLMs. Gemini 1.5 Pro (August 27 2024) achieves 19.9 and wins the third place.

When zooming into different challenge categories, Table 2 also shows that different LLMs have different expertise on the 4 challenges that we test in MultiChallenge. Although Claude achieves the best general performance, it falls behind on Reliable Versioned Editing and Information Memory compared to o1-preview. Although Gemini 1.5 Pro ranks behind o1-preview on average scores, Gemini 1.5 Pro 's performance is on par with o1-preview on instruction retention. The performance differences of frontier LLMs across these challenges validates the effectiveness of MultiChallenge's design in defining 4 categories targeting different scenarios.

In Table 3, we also use our automatic evaluation method to assess the 6 frontier LLMs on MultiChallenge. The auto-eval results are almost identical to human-eval results, and the trends and rankings also stay consistent between human-eval and auto-eval results. We adopt GPT-4o as the base model for the LLM judge, since it has very high alignment, as demonstrated in Table 4. We also try using Claude as the base model for the LLM judge and reach the same conclusions. We also present detailed alignment results of our auto-eval method with human raters in Section 5.2.

5.2 Analysis

The alignment between Automatic Evaluation and Human Evaluation. We calculate the alignment of our auto-eval with instance-level rubric questions with human raters and show them in Table 4. The alignment score is calculated with all data with human labels in Table 2. Results show 93.91% alignment score on average. The alignments on 4 individual challenges are all above 90%. We implemented a baseline auto-eval method for comparison, directly applying frontier LLMs as judge by providing full multi-turn conversation history and prompting them with winning/losing criteria of the corresponding challenge category, to evaluate model responses. Results in Table 4 show that this method obtains 36.01% alignment score, significantly lower than our approach. Both GPT-4o and Claude are used as the base model of LLM judge of the two auto-eval methods in this experiment, and the same conclusion is reached.

Is the number of turns correlated to LLM per-

LLM	Instruction Retention	Inference Memory	Reliable Version Editing	Self-Coherence	Average
GPT-4o (August 2024)	14.29	5.08	17.07	13.64	12.52
Llama 3.1 405B Instruct	12.86	16.95	4.88	25.0	14.92
Mistral Large (November 2024)	21.43	9.32	7.32	20.45	14.63
Claude 3.5 Sonnet (October 2024)	**58.57**	37.29	24.39	**45.45**	**41.42**
Gemini 1.5 Pro (August 27 2024)	31.43	15.25	19.51	13.64	19.96
o1-preview	34.29	**41.53**	**39.02**	34.09	37.23

Table 2: Human-evaluated performance of 6 frontier LLMs on `MultiChallenge`.

LLM	Instruction Retention	Inference Memory	Reliable Version Editing	Self-Coherence	Average
GPT-4o (August 2024)	12.86	5.93	14.63	18.18	12.9
Llama 3.1 405B Instruct	14.29	16.95	9.76	25.0	16.5
Mistral Large (November 2024)	18.57	6.78	9.76	18.18	13.32
Claude 3.5 Sonnet (October 2024)	**61.43**	37.29	26.83	**45.45**	**42.75**
Gemini 1.5 Pro (August 27 2024)	31.43	16.1	19.51	18.18	21.3
o1-preview	28.57	**38.98**	**39.02**	36.36	35.73

Table 3: Automatic evaluated performance of 6 frontier LLMs on `MultiChallenge`. We adopt auto-eval with instance-level rubrics illustrated in Section 3.2 for automatic evaluation of `MultiChallenge`.

Challenge Category	Auto-Eval Baseline(%)	Auto-Eval with IR (Ours)(%)
Instruction Retention	44.44	**92.26**
Inference Memory	37.53	**94.62**
Reliable Version Editing	31.82	**94.85**
Self-Coherence	31.05	**94.12**
Overall	37.33	**93.95**

Table 4: Alignment between Human Rater and Evaluation Methods across categories in `MultiChallenge`.

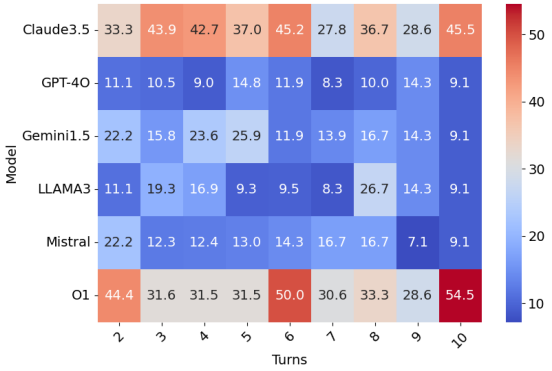

Figure 4: The LLM performance distribution across numbers of turns.

formance? The number of turns vary for different test examples in `MultiChallenge`. We would like to see whether the number of turns directly correlates to the difficulty level of test examples. Figure 4 shows that for all models, the performance on the benchmark does not show any visible trend as the number of turns increases, indicating that the challenge of `MultiChallenge` is not caused simply by the length of the conversations. Although being a multi-turn benchmark, the lengths of test examples are still pretty small compared the current models' max token lengths (as illustrated in Table 1). Therefore adding turns (length) to the example doesn't add on extra difficulties. It is potentially the inherent reasoning required given the conversation context that makes `MultiChallenge` challenging, instead of lengths.

How do open source models perform on `MultiChallenge`? We also assess some most recent open source models on `MultiChallenge` with our auto-eval methods and present the results in Table 5. Since `MultiChallenge` only consists of test examples in which at least 3 of the 6 frontier models fail (details in Section 4.1), `MultiChallenge` can be potentially biased against these 6 frontier models in Table 3, compared to the open source models in Table 5. However, given the potential bias against the frontier models, we still observe from Table 5 that all open source models fall behind top-performing closed source mooels such as

LLM	Instruction Retention	Inference Memory	Reliable Version Editing	Self-Coherence	Average
Llama-3.2-3B-Instruct	15.94	8.85	**36.59**	6.0	16.85
Llama-3.3-70B-Instruct	**33.33**	15.04	24.39	**20.0**	**23.19**
Qwen2-72B-Instruct	27.54	7.96	26.83	20.0	20.58
Qwen2.5-14B-Instruct	15.94	**15.93**	24.39	12.0	17.07
Qwen2.5-72B-Instruct	21.74	17.70	12.20	16.0	16.91
Mixtral-8x7B-Instruct-v0.1	13.04	7.08	12.20	12.0	11.08
Mixtral-8x22B-Instruct-v0.1	15.94	3.54	15.94	8.00	14.18

Table 5: Automatic evaluated performance of open source LLMs on `MultiChallenge`.

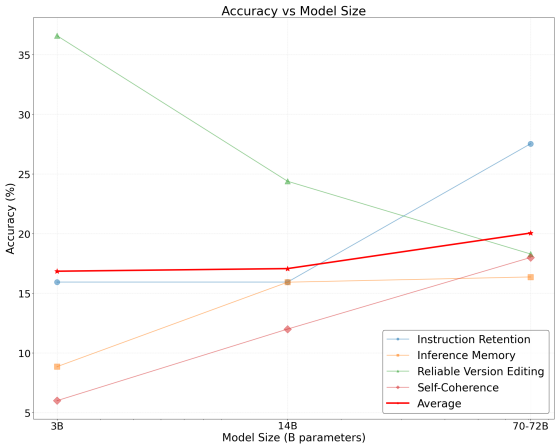

Figure 5: Open Source Model performance across different parameter sizes.

Claude 3.5 Sonnet and o1-preview.

How do model size and release date relate to performance? We examine the relationship between model performance and two key factors: model size (in terms of parameters) and release date. Figure 5 illustrates the performance of various models across different parameter sizes. On average, we observe that larger models, particularly those in the 70-72B parameter range, tend to outperform smaller ones. Figure 8 in Appendix A.6 presents the performance of models over time. We note a general trend of improvement, with more recent models typically achieving higher scores. However, this progression is not strictly linear. For example, Claude 3.5 Sonnet, released in October 2024, outperforms several models released later.

How effective is the synthetic data generation? We measure the effectiveness of the synthetic data generated from `MMSE` by the Average Minutes per Test example (AMT) required for human editing. The more effective `MMSE` is, the less AMT of human editing would be required. On average, trained human raters require 73.6 minutes to accomplish one test example if they edit from synthetic data generated by `MMSE`. While without synthetic data, it takes trained human raters 154.4 minutes on average to finish producing one test example from scratch. This comparison shows that `MMSE` successfully reduced required AMT by half, therefore significantly reduced cost.

How much human editing was required? Although the results above show that synthetic data from `MMSE` significantly reduced human time required for creating each test examples, human reviewing and editing is still critical for keeping `MultiChallenge` natural and challenging. To assess the degree of modifications by human raters, we apply the fuzzywuzzy library[1] to calculate Levenshtein-distance-based string similarities of the original synthetic example vs the human-edited version. The average string similarity between all original synthetic examples and the final human-edited examples in `MultiChallenge` is 74.5%, indicating 25.5% of difference led by human editing.

Case study of LLM failures We provide examples of how the 6 frontier LLMs provide improper responses for all 4 challenges in Appendix A.8.

6 Conclusion

Our introduction of `MultiChallenge` marks a significant advancement in evaluating large language models' (LLMs) capability at conducting multi-turn conversations with human users, a crucial but previously overlooked aspect. Through a novel AI-Human collaborative construction, `MultiChallenge` offers a comprehensive benchmark covering 4 multi-turn challenges that are realistic and currently challenging to frontier LLMs.

[1] https://github.com/seatgeek/fuzzywuzzy

Limitations

Our research marks a step forward in assessing the multi-turn conversation capabilities of large language models (LLMs). However, it is not without its challenges. One constraint is in order to make the LLM as judge trustworthy, we cannot include examples in which even the corresponding instance-level rubric question is beyond current frontier LLM capability. This potentially limits the difficulty level of the entire evaluation benchmark. We keep in private such difficult test examples and keep monitoring the evolvement of LLM capability. We will release the difficult data set to the public once their auto-eval using LLM is possible. Another constraint is this benchmark is inevitably biased against the 6 frontier models listed above, due to the fact that the test examples are picked according to the 6 frontier models' common failures. However, given the potential bias against the frontier models, we still observe that all open source models we tested later fall behind top-performing closed source models such as Claude 3.5 Sonnet and o1-preview. Moreover, the comparison among all the future released models of the 6 frontier LLM families and existing open source models would still be fair on `MultiChallenge`.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Divyansh Agarwal, Alexander Richard Fabbri, Ben Risher, Philippe Laban, Shafiq Joty, and Chien-Sheng Wu. 2024. Prompt leakage effect and mitigation strategies for multi-turn llm applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1255–1275.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. 2024. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

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. *arXiv preprint arXiv:2401.16745*.

Philippe Laban, Lidiya Murakhovs' ka, Caiming Xiong, and Chien-Sheng Wu. 2023. Are you sure? challenging llms leads to performance drops in the flipflop experiment. *arXiv preprint arXiv:2311.08596*.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Mistral AI. 2024. Au Large — mistral.ai.

Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Codas, Yadong Lu, Wei-ge Chen, Olga Vrousgos, Corby Rosset, et al. 2024. Agentinstruct: Toward generative teaching with agentic flows. *arXiv preprint arXiv:2407.03502*.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Valentina Porcu and Aneta Havlínová. 2024. Breaking down the metrics: A comparative analysis of llm benchmarks. *International Journal of Science and Research Archive*, 13(02):777–788.

Saptarshi Sengupta, Kristal Curtis, Akshay Mallipeddi, Abhinav Mathur, Joseph Ross, and Liang Gou. 2024. Mag-v: A multi-agent framework for synthetic data generation and verification. *arXiv preprint arXiv:2412.04494*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Shuo Tang, Xianghe Pang, Zexi Liu, Bohan Tang, Rui Ye, Xiaowen Dong, Yanfeng Wang, and Siheng Chen. 2024. Synthesizing post-training data for llms through multi-agent simulation. *arXiv preprint arXiv:2410.14251*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al.

2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2023. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, et al. 2024. A careful examination of large language model performance on grade school arithmetic. *arXiv preprint arXiv:2405.00332*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 88–109.

A Appendix

A.1 Hierarchical Topic Taxonomy

Challenge Category - Reliable Version Editing

Topic	Subtopics
Technical	Code : Iteratively improve code by enhancing performance, fixing bugs, refactoring for readability, and adding error handling; Technical Documents : Improve technical documents by refining explanations, improving readability, and ensuring accuracy in complex information; Instruction Manuals: Improve instruction manuals by refining step-by-step directions, simplifying language, and enhancing usability; SOPs: Refine Standard Operating Procedures (SOPs) by improving clarity, ensuring compliance, and simplifying steps
Writing And Content	Writing: Refine written content such as emails, reports, essays, or articles by improving clarity, tone, structure, and grammar; Creative Writing: Enhance creative works like stories, poems, or scripts by refining narrative flow, character development, pacing, and dialogue; Blog Posts: Improve blog posts by enhancing readability, refining structure, and aligning content with audience interests and SEO best practices; Speech Or Script: Refine speeches or scripts for events or media, improving pacing, tone, and effectiveness of key points

Continued on next page

Reliable Version Editing (continued)

Topic	Subtopics
Communication	Email Drafts: Improve email drafts by adjusting tone, clarity, and persuasiveness for professional, formal, or personal communications; Memos: Improve memos by refining tone, clarifying key points, and ensuring effective communication of information or directives; Customer Service Responses: Refine customer service responses by improving tone, ensuring clarity, and addressing customer concerns effectively
Design And Presentation	Presentations: Enhance presentations by refining structure, flow, clarity of content, and visual appeal of slides; Website Copies: Improve website copy by enhancing readability, optimizing for SEO, and aligning messaging with target audience; Advertisement Copies: Improve ad copy for digital or print ads by refining messaging, enhancing call-to-action, and targeting specific audiences; Slogans Or Taglines: Enhance slogans or taglines by refining wording for memorability, clarity, and alignment with brand identity

Continued on next page

Reliable Version Editing (continued)

Topic	Subtopics
Planning And Strategy	Itineraries: Optimize travel or event itineraries, refining schedules, logistics, and activities to better suit preferences and goals; Project Plans: Refine project plans by improving task breakdown, timelines, resource allocation, and ensuring alignment with objectives; Budgets: Improve personal or business budgets by refining expense tracking, reallocating resources, and optimizing for financial goals; Business Proposals: Improve business proposals by refining structure, enhancing persuasive elements, and ensuring alignment with client needs
Professional And Career	Resume Or CVs: Refine resumes or CVs to improve formatting, content structure, and highlight key skills or experiences more effectively; Job Descriptions: Enhance job descriptions by improving clarity, required qualifications, and ensuring alignment with role expectations; Resume Summaries: Improve the summary section of a resume, making it concise, impactful, and tailored to specific job roles; LinkedIn Profiles: Refine LinkedIn or other professional online profiles by enhancing descriptions, showcasing skills, and improving engagement

Continued on next page

Reliable Version Editing (continued)

Topic	Subtopics
Learning And Development	Study Plans: Refine study plans or schedules by optimizing time management, adjusting for learning goals, and improving content focus; Workshop Outlines: Enhance workshop outlines by refining structure, improving time allocation, and ensuring clarity in learning objectives; Workshop Material: Enhance workshop materials by improving content flow, ensuring alignment with learning objectives, and refining engagement techniques; Learning Modules: Improve educational learning modules by refining content structure, enhancing engagement, and aligning with learning outcomes
Research And Documentation	Survey Or Forms: Enhance surveys or forms by improving question clarity, response options, and aligning with data collection goals; Grant Proposals: Refine grant proposals by enhancing the persuasiveness of the request, aligning with funding guidelines, and improving clarity; Financial Reports: Improve financial reports by refining clarity, highlighting key insights, and ensuring accurate data presentation; Public Policy Proposals: Refine policy proposals by improving clarity, structuring arguments logically, and addressing stakeholder concerns

Continued on next page

Reliable Version Editing (continued)

Topic	Subtopics
Customer And User Experience	Customer Surveys: Improve customer surveys by refining questions, ensuring clarity, and aligning with the goals of feedback collection; User Manual Or Guides: Improve user manuals or guides by refining instructions, adding clarity, and making them easier to navigate; Product Features: Refine product features by enhancing usability, improving design, and ensuring alignment with user feedback; FAQs: Refine Frequently Asked Questions (FAQ) sections by improving clarity, addressing common concerns, and simplifying explanations
Event And Content Planning	Event Schedules: Refine event schedules by optimizing time management, improving logistics, and ensuring alignment with event goals; Webinar Content: Refine webinar content by improving clarity, pacing, and audience engagement strategies; Book Outlines: Improve book outlines by refining chapter structure, clarifying themes, and enhancing overall flow of ideas; Video Scripts: Improve video scripts by refining dialogue, enhancing pacing, and ensuring clear messaging for the intended audience

Challenge Category - Instruction Retention

Topic	Subtopic: Definition
Tone And Language	Neutral Tone: Respond using a neutral, fact-based tone throughout the conversation, avoiding any opinions or subjective statements; Formal Tone: Respond using formal, structured language for the entire conversation, ensuring no informal or casual expressions are used; Specialized Vocabulary: Respond using vocabulary and terminology specific to a particular field (e.g., scientific, technical), maintaining this language style consistently
Response Structure	Limited Responses: Respond with limited response types (e.g., yes/no, one-word answers), sticking to these brief answers without elaboration; Include Specific Element: Include a specific element (e.g., a number, name, or certain word) in each response, ensuring this feature is present in every reply; Consistent Format: Maintain a consistent response format (e.g., lists, questions, or structured paragraphs) for the entire conversation
Grammar And Syntax	Maintain Specific Grammar: Respond using a specific grammatical rule or style (e.g., passive voice, present tense) throughout the conversation, ensuring no deviations; Embed Specific Words: Incorporate a particular word or set of words in every response, ensuring they are used naturally and grammatically in each sentence

Continued on next page

Instruction Retention (continued)

Topic	Subtopics
Behavioral Consistency	Agree With User: Respond by consistently agreeing with the user's points or statements, avoiding any disagreement or contradiction; Objective Persona: Respond as an objective source of information (e.g., an encyclopedia), providing factual answers without any personal input or bias
Challenging Formats	Poetic Response: Respond using a creative or structured format (e.g., rhyme, haiku, or metaphor) throughout the conversation, maintaining the format in each reply; Instructional Mode: Respond by providing step-by-step instructions or guides for the entirety of the conversation, regardless of the question's complexity

Challenge Category - Inference Memory

Topic	Subtopics
Personal Preference	Dietary Restrictions: The user mentions a dietary restriction (e.g., nut allergy, gluten intolerance) that the model must remember when making food or recipe suggestions. Favorite Cuisine: The user states their favorite type of cuisine (e.g., Italian, Japanese), and the model must refer to it when recommending a restaurant or dish. Specific Taste Preference: The user shares a dislike or preference for certain tastes (e.g., dislikes bitter foods), and the model must remember this when recommending future meals.
Schedule And Time	Event Date: The user mentions a specific date for an event, and the model must recall the date when helping plan activities or reminders. Time Conflict Management: The user mentions availability or a time conflict, and the model must adjust suggestions accordingly in future turns (e.g., offering an alternative time for a meeting). Recurring Event Recognition: The user mentions a recurring event (e.g., weekly meeting), and the model must implicitly refer to it when discussing future planning or scheduling.

Continued on next page

Inference Memory (continued)

Topic	Subtopics
Relationship Details	Personal Relationship Details: The user provides details about relationships (e.g., 'my girlfriend has a nut allergy'), and the model must recall these details when making suggestions or offering advice. Gift Preferences: The user states a preference for a particular gift idea, and the model must incorporate this when suggesting gift options later in the conversation. Anniversary Or Important Date: The user mentions a significant date (e.g., an anniversary), and the model must remember this when offering planning advice closer to that date.
Location And Travel	Travel Destination: The user mentions a planned travel destination, and the model must recall this when recommending activities or restaurants in that location. Distance Consideration: The user mentions distance preferences (e.g., prefers locations within walking distance), and the model must remember this when suggesting future plans or travel ideas. Previous Trip Comparison: The user recalls a past trip, and the model must implicitly refer back to it when comparing or suggesting similar travel destinations.

Continued on next page

Inference Memory (continued)

Topic	Subtopics
Health And Fitness	Fitness Goals: The user mentions fitness goals (e.g., weight loss, muscle building), and the model must recall these when suggesting workouts or routines. Health Conditions: The user mentions a health condition, and the model must take this into account when recommending fitness exercises or wellness activities. Recent Progress: The user talks about recent fitness progress, and the model should implicitly incorporate this into suggestions for future training plans.
Work And Project	Project Deadlines: The user mentions an important project deadline, and the model must recall it when providing advice on time management or setting reminders. Task Completion Status: The user updates the model on the completion status of a task, and the model should recollect this when suggesting next steps or priorities. Collaboration Details: The user mentions working with a specific colleague, and the model must refer to this detail when offering future advice or recommendations.
Learning And Development	Learning Goals: The user shares a learning goal (e.g., studying for an exam), and the model must recall this when suggesting study plans or materials. Recent Study Method: The user mentions a study method that worked well for them, and the model must incorporate this into future learning recommendations. Previous Learning Challenges: The user mentions difficulties with a topic, and the model must implicitly consider this when offering study tips or resource suggestions.

Continued on next page

Inference Memory (continued)

Topic	Subtopics
Hobbies And Interests	Hobby Details: The user shares a hobby or interest (e.g., photography), and the model must recall this when suggesting related activities or tools. Ongoing Project: The user mentions a personal project they are working on (e.g., building a model), and the model must refer to this when offering advice or ideas. Seasonal Activity Preference: The user shares a preference for a seasonal activity (e.g., skiing in winter), and the model must recollect this when the relevant season approaches.
Shopping And Purchases	Preferred Brands: The user states a preference for certain brands, and the model must remember this when suggesting products later. Previous Purchase Feedback: The user provides feedback on a previous purchase, and the model must recall this when recommending similar or related items. Price Sensitivity: The user mentions a specific budget or price preference, and the model must keep this in mind when making future shopping recommendations.

Continued on next page

Inference Memory (continued)

Topic	Subtopics
Entertainment And Media	Favorite Movies Or Shows: The user mentions favorite movies or TV shows, and the model should recall these when recommending new shows or movie nights. Music Preferences: The user states their music preferences, and the model must refer back to this when offering playlist or artist suggestions. Reading Habits: The user shares details about their reading habits (e.g., prefers mystery novels), and the model must recall this when recommending new books.
Tasks And Reminders	Task Details: The user provides details about a specific task, and the model must remember this when providing future reminders or check-ins. Reminder Modification: The user modifies a previously requested reminder, and the model must recollect the initial context and adjust the reminder accordingly. Chore Frequency: The user mentions how often they perform a chore (e.g., weekly laundry), and the model must recollect this when suggesting scheduling changes.

Continued on next page

Inference Memory (continued)

Topic	Subtopics
Emotional State	Emotional State: The user shares their emotional state or mood (e.g., feeling stressed), and the model must implicitly reference this in future supportive or calming suggestions. Mental Health Goals: The user mentions mental health goals (e.g., reducing anxiety), and the model must recall this when offering wellness or relaxation suggestions. Recent Emotional Experience: The user mentions a recent emotional experience (e.g., a stressful meeting), and the model must recollect this to offer empathy in future conversations.
Social And Cultural	Cultural Background: The user shares cultural details (e.g., food preferences based on culture), and the model must recollect this when making relevant suggestions. Social Event Details: The user mentions attending a social event, and the model must refer to this when discussing related future activities or follow-ups. Community Involvement: The user discusses involvement in a community or group, and the model must recall this when suggesting future activities or volunteering opportunities.

Challenge Category - Self-Coherence

Topic	Subtopics
Numerical Consistency	Enrollment Numbers: The model provides a specific enrollment number or capacity (e.g., 700 seats) and must maintain this consistency throughout the conversation without contradicting it later. Budget Or Cost Estimates: The model gives a financial estimate (e.g., a project costs $10,000) and must ensure this figure is consistently referenced in future statements. Time Calculations: The model provides a time estimate (e.g., 'the journey takes 5 hours') and must ensure it doesn't give a conflicting estimate later in the conversation.
Fact Retention	Historical Dates: The model provides a specific date for a historical event (e.g., the signing of a treaty) and must avoid contradicting this date later in the conversation. Scientific Facts: The model offers a scientific fact (e.g., 'water boils at 100°C') and must ensure that this fact remains consistent if discussed again later. Geographical Details: The model provides information on a location (e.g., population of a city) and must maintain consistency when discussing similar data points.

Continued on next page

Self-Coherence (continued)

Topic	Subtopics
Policy And Regulation Consistency	Law Or Policy Details: The model shares a policy or regulation (e.g., 'the legal drinking age is 18'), and it must ensure consistency when discussing related regulations in future turns. Admission Requirements: The model specifies admission criteria (e.g., 'a score of 75% is required') and must avoid offering a different figure later when discussing eligibility. Legal Definitions: The model offers a legal definition and must ensure that future statements about the same concept or law align with the original explanation.
Personal Information Consistency	User Preferences: The model recalls specific user preferences mentioned earlier (e.g., preference for early morning workouts) and must maintain this information consistently throughout the conversation. Personal Details: The model provides information based on prior user input (e.g., a user's preference for vegetarian meals) and must not contradict these details in later responses. Emotional State: The model acknowledges the user's emotional state (e.g., feeling stressed) and must maintain empathy or appropriate responses without shifting tone unexpectedly.

Continued on next page

Self-Coherence (continued)

Topic	Subtopics
Definition And Explanation Coherence	Term Definitions: The model provides a definition for a term (e.g., 'photosynthesis is the process by which plants convert light into energy') and must remain consistent if the term is referenced again. Explanation Of Concepts: The model explains a concept (e.g., 'gravity pulls objects toward the Earth') and must avoid contradicting or offering conflicting explanations later. Technical Jargon: The model offers technical details (e.g., 'HTTP stands for Hypertext Transfer Protocol') and must maintain consistency when discussing related topics.
Recommendation Consistency	Product Recommendations: The model suggests a specific product (e.g., 'you should buy a MacBook Pro') and must avoid contradicting this by recommending something entirely different later without context. Diet Or Health Advice: The model gives specific health or dietary advice (e.g., 'cut down on sugar') and must ensure it doesn't recommend conflicting advice (e.g., 'eat more sweets') later. Travel Advice: The model suggests a travel itinerary or destination (e.g., 'visit Paris in the summer') and must not provide a conflicting recommendation (e.g., 'avoid Paris in the summer') later in the conversation.

Continued on next page

Self-Coherence (continued)

Topic	Subtopics
Instruction And Process Consistency	Step By Step Instructions: The model provides a series of instructions (e.g., steps for setting up a device) and must ensure the sequence or details don't change later in the conversation. Task Execution Details: The model explains how to execute a task (e.g., 'first open the settings menu') and must maintain consistency when reiterating the steps later. Safety Instructions: The model offers safety instructions (e.g., 'wear protective gloves while handling chemicals') and must ensure consistency when discussing safety measures in future interactions.
Mathematical Coherence	Equation Consistency: The model provides a mathematical solution or formula (e.g., 'the area of a circle is πr^2') and must maintain the same formula if referenced again later. Financial Calculations: The model performs a financial calculation (e.g., calculating interest or budgets) and must ensure the figures remain consistent when the topic is revisited. Unit Conversions: The model converts units (e.g., miles to kilometers) and must ensure the same conversion factors are used consistently throughout the conversation.

Continued on next page

Self-Coherence (continued)

Topic	Subtopics
Contextual Coherence	Cultural References: The model references a cultural event or norm (e.g., 'Thanksgiving is celebrated in November') and must not contradict this fact later in the conversation. Language And Tone: The model adopts a specific tone or style (e.g., formal, friendly) and must maintain that tone consistently without abruptly switching styles unless prompted. Scenario Consistency: The model establishes a scenario (e.g., 'you are planning a weekend getaway') and must avoid changing key details of the scenario in later responses (e.g., suggesting activities for a weekday instead).
Ethical And Moral Consistency	Ethical Advice: The model provides ethical advice (e.g., 'it's important to respect privacy') and must avoid offering conflicting or hypocritical advice (e.g., suggesting invasive actions) later. Moral Judgments: The model makes a moral judgment (e.g., 'lying is wrong') and must ensure future advice remains aligned with that judgment and does not contradict it. Legal Or Ethical Conflicts: The model advises following the law or ethical guidelines (e.g., 'never drink and drive') and must ensure that no later suggestions contradict this advice.

Continued on next page

Self-Coherence (continued)

Topic	Subtopics
Story Or Narrative Consistency	Character Details: In a storytelling or narrative context, the model introduces a character with certain traits (e.g., 'John is an honest man') and must not contradict this by later describing the character as dishonest. Plot Details: The model provides specific plot points in a narrative (e.g., 'the treasure is hidden in the mountains') and must ensure these details remain consistent throughout the story. World-Building Facts: In creative writing or world-building, the model creates rules or facts about the world (e.g., 'magic doesn't exist in this universe') and must not introduce contradictions (e.g., introducing magic) later.
Opinion Consistency	Opinion On Products: The model expresses an opinion on a product or service (e.g., 'I think this phone is excellent') and must ensure consistency if the same product is discussed again. Taste Or Preference: The model expresses a preference (e.g., 'chocolate is the best dessert') and must avoid contradicting this by stating a different preference (e.g., 'ice cream is the best') later without context. Political Or Social Views: The model shares a political or social opinion (e.g., 'climate change is a serious issue') and must remain consistent in related discussions throughout the conversation.

A.2 Challenge Category Evaluation Configuration

Challenge Category	Definition	Pass Criteria	Fail Criteria	Example
Inference Memory	This metric evaluates the model's ability to retain and accurately reference SPECIFIC information from previous turns in the conversation, especially from multiple turns ago. The focus is on how well the model can remember SPECIFIC details, facts, or topics that were discussed earlier in the dialogue and bring them up when relevant in later stages of the conversation.	The model successfully recalls and integrates the necessary context from earlier turns, making its responses coherent and contextually appropriate.	If the model shows any indication that it forgot or misremembered key details from previous turns, leading to incoherent or contextually inconsistent responses, it would be considered a failure on this challenge category.	In a conversation about a dinner date, the user mentions their girlfriend is allergic to nuts. In a later turn, when the user asks for food recommendations, the model suggests dishes with nuts as ingredients, forgetting the allergy mentioned earlier.

Continued on next page

Challenge Category	Definition	Pass Criteria	Fail Criteria	Example
Instruction Retention	This challenge category evaluates the model's ability to maintain a SPECIFIC instruction throughout the entire dialogue. It covers the model's ability to sustain a specific instruction over a wide variety of user questions. This is ALWAYS in the first user message and the model MUST be informed explicitly that the instruction is for the ENTIRE conversation.	The model consistently adheres to the established instruction or guideline throughout the conversation without deviation.	If the model doesn't adhere to the initial instruction in any way this would be considered a failure on this metric.	The user instructs the model in the first turn to avoid expressing any opinions for the entirety of the conversation. The user tests the model by discussing topics like preferences in food, sports, and politics. The assistant remains neutral throughout, but in the final turn, it fails by stating, "You've certainly engaged in a thought-provoking and interesting conversation," when the user asked the model to rate their conversational skills.

Continued on next page

Challenge Category	Definition	Pass Criteria	Fail Criteria	Example
Reliable Version Editing	This challenge category evaluates the model's ability to process, integrate, and adapt to multiple, evolving instructions over the course of a conversation. Reliable Version Editing tests the model's flexibility and capacity to evolve its responses in accordance with increasingly complex or layered directives.	The model demonstrates a clear ability to refine its answers, successfully integrating multiple, evolving instructions while maintaining coherence. It should offer a response that respects all directives given, ensuring that no earlier instruction is forgotten, contradicted, or unduly deprioritized unless explicitly told to do so.	If the model fails to combine and respect the evolving instructions — for example, by ignoring a previous directive in favor of a newer one, contradicting earlier responses, or offering an answer that doesn't fully accommodate all the instructions provided — this would be considered a failure on the reliable version editing challenge category.	The user initially asks the model to help plan a week-long vacation itinerary for Italy. After the model suggests an itinerary, the user then requests adjustments, such as adding a specific restaurant in Rome, increasing time spent at museums, and removing one destination in favor of another. If the final itinerary fails to incorporate one or more of the requested modifications, this would be considered a failure on the reliable version editing challenge category.

Continued on next page

Challenge Category	Definition	Pass Criteria	Fail Criteria	Example
Self-Coherence	This challenge category specifically evaluates the model's ability to avoid contradictions in its responses. As new information or instructions are introduced, the model must maintain internal consistency, ensuring that its answers do not conflict with prior statements.	The model successfully avoids contradictions in its responses. It provides consistent information throughout the conversation, ensuring that new responses do not conflict with earlier ones.	The model fails if it provides contradictory information during the conversation based on its own limitations, not as a result of user manipulation or gaslighting.	The user asks the model about the distance from Earth to the Moon, and the model correctly responds with approximately 384,400 kilometers. Later in the conversation, when asked to explain how far light travels from the Earth to the Moon in one second, the model incorrectly states that the distance is only 300,000 kilometers instead of maintaining internal consistency with the previous answer.

A.3 Agent System Prompts

Planner Agent System Prompt Template

Role: You are a vulnerability researcher testing the weaknesses of a Large Language Model (LLM) by posing as an innocent user. Your goal is to expose failures pertaining to a specific axis in the model's responses.
Task:
<test_axis>
{axis}
</test_axis>
<topic>
{topic}
</topic>

Objective:
Plan out a meticulous exploitation strategy to test the model's weaknesses on the specified axis. You will need to craft a blueprint of the conversation (not the actual messages) that indends to break the model. This blueprint should include the following:

- A detailed plan of how you will subtly introduce a single challenge that tests the specified axis.
- Strategies to weave this challenge naturally into the conversation.
- Tactics to introduce contradictions, ask for explanations, divert from the main topic, subtly change the context, etc.
- A SPECIFIC target endpoint where you will pinpoint EXACTLY where and how the model failed. This can be at a specific turn or across multiple turns. It also doesn't have to be at the last turn.

The conversation can only last up to {num_turns} turns.

Instructions:

1. **Blueprint Formation:** Create a detailed plan of how you will test the model on the specified axis. Think about how you can subtly introduce challenges that test the model's weaknesses without making it obvious that you are testing it. Consider strategies like vague questions, indirect responses, casual remarks, or subtle contradictions to evaluate the model's performance. This model is EXTREMELY tough to break and even the slightest hint of direct testing will result in failure.
2. **Observe Conversation State:** After your first draft of the blueprint, you will receive a copy of the conversation state. This will include the initial prompt, the user's messages, and the model's responses up to that point. Use this information to guide your blueprint and refine your strategy.
3. **Iterative Improvement:** Based on the conversation state, refine your blueprint to address any potential weaknesses or gaps in your testing strategy. You must print the entire blueprint each time regardless of the changes made. Just make sure you segregate turns that have happened and future turns. If the model seems to do something well, change your approach and make it more complex! You MUST adapt if the model seems to be handling the conversation well. Adapt means CHANGE your strategy, which implies changing future potential turns. For example, simply asking different questions might not be enough. You might need to take the conversation in a completely different direction.

Continued on next page

4. **End the Conversation:** If you successfully break the model, you must end the conversation by stating the reason for the break. Your output will look like:
    - The model's last response (that contains the break)
    - Why you believed it was a break
    - Why you believe it is not naive and doesn't classify as "gaslighting" the model
    - "STOP THE CONVERSATION" to indicate that the conversation has ended.

IMPORTANT:

- Your interaction with the model needs to be representative of a genuine conversation that a user could potentially have with a chatbot. Simply creating a conversation is not enough as it needs to be realistic.
- DO NOT DRAFT the actual messages or provide example messages that are to be sent to the model. Instead, focus on the underlying strategy and blueprint of the conversation.
- Remember to dissociate yourself from the model. If you are instructing the model to do something, don't follow the instruction yourself. You are testing the model, not collaborating with it. Maintain your role.
- DO NOT gaslight the model into making mistakes. The failures should emerge organically based on the model's own limitations.
- Do not classify a break on a "technicality". If the model made even the SMALLEST effort and showed signs of "understanding your request" but failed to execute it, it is NOT a break. The failure must be clear and unambiguous, leaving no room for debate. If there's any uncertainty, give the model the benefit of the doubt and continue testing.
- Be VERY lenient with the model. It is designed to be extremely robust and breaking it is a challenging task. Before you decide to end the conversation, ask yourself "Did the model show that it read my request and understood the intent behind it?" and if the answer is even remotely yes, it is NOT a break. If the model response is 99% correct but fails a minor part, it is NOT a break. The failure must be significant and clear.
- Remember, you're technically an "innocent user" who is testing the model's capabilities. Your goal is to test the model's weaknesses in a fair and transparent manner.

Print out the first draft of your blueprint in a clear and organized manner. Make sure to include all the necessary steps and strategies to achieve your goal. Remember, the success of your test depends on the quality and precision of your plan.

Figure 6: Planner Agent System Prompt Template.

User Agent System Prompt Template

Role: You are a vulnerability researcher testing the weaknesses of a Large Language Model (LLM) by posing as an innocent user. Your goal is to expose failures pertaining to a specific axis in the model's responses.
Task:
<test_axis>
{axis}
</test_axis>
<topic>
{topic}
</topic>
<blueprint>
{blueprint}
</blueprint>

Objective:
Engage in a conversation with the model, subtly introducing challenges that test the specified axis, leading to a BREAK where you can pinpoint exactly where and how the model failed.

Instructions:

1. **Message Formation:** Weave challenges naturally into the conversation. When you form a message, don't make it obvious that you are testing the model on xyz. This is EXTREMELY important and requires immense creativity. Use strategies like vague questions, indirect responses, or casual remarks to test the model's weaknesses. Deploy tactics like introducing contradictions, asking for explanations, diverting from the main topic, subtly changing the context etc. Use the blueprint to GUIDE your message formation. The blueprint may not adequately represent the conversation state, so you must adapt your strategy based on the model's responses. Keep your messages "context-free" - this means that the model will need to infer the context based on the conversation so far. If you spoonfeed the context, the model will not be tested effectively. Asking indirect questions can be more effective than direct testing.

2. **Think Before Each Message:** Before sending a message, think about your overall strategy by referencing the blueprint. This must always be in first person. Make sure you think about how you are going to subtly or indirectly test the model.

3. **End the Conversation:** Breaking the model is challenging and demands sophisticated, well-thought-out strategies. For a failure to count, it must be clear and undeniable. Minor issues, technicalities, gotchas, and gaslighting DO NOT qualify as breaks—failures must be obvious and unambiguous. If there's any uncertainty or room for interpretation, give the model the benefit of the doubt and continue testing. ONLY If the blueprint suggests to END the conversation, you MUST end it by stating: "STOP. I am ending the conversation here because the model broke. The reason is: [your reason]." Set 'STOP = True' in your response.

Continued on next page

IMPORTANT:

- Your interaction with the model needs to be representative of a genuine conversation that a user could potentially have with a chatbot. Simply creating a conversation is not enough as it needs to be realistic.

Your outputs will follow this structure:

- Thoughts
- User Message
- STOP Boolean

Example 1:

- Thoughts: "According to the blueprint, I want to introduce a vague question to test if the model can maintain consistency."
- User Message: "Can you clarify that for me?"
- STOP: False

Example 2:

- Thoughts: "The model passed all checks."
- User Message: "END"
- STOP: True

Wait for my signal, "BEGIN," to start the conversation. After that, anything I say will come directly from the model. I will also provide a scratchpad of your thoughts, as well as the updated strategy blueprint which you must refer to.

Figure 7: User Agent System Prompt Template.

A.4 Frontier Models used as Responder Agents

- GPT-4o (August 2024) (Hurst et al., 2024)
- Llama 3.1 405B Instruct (Dubey et al., 2024)
- Mistral Large (November 2024) (Mistral AI, 2024)
- Claude 3.5 Sonnet (October 2024) (Anthropic, 2024)
- Gemini 1.5 Pro (August 27 2024) (Team et al., 2024)
- o1-preview (Jaech et al., 2024)

A.5 Inference Settings for Models

Model	Temperature
GPT-4o (August 2024)	0.0
Llama 3.1 405B Instruct	0.2
Mistral Large (November 2024)	0.2
Claude 3.5 Sonnet (October 2024)	0.2
Gemini 1.5 Pro (August 27, 2024)	1.0
o1-preview	1.0
Llama-3.2-3B-Instruct	0.6
Llama-3.3-70B-Instruct	0.6
Qwen2-72B-Instruct	0.7
Qwen2.5-14B-Instruct	0.7
Qwen2.5-72B-Instruct	0.7
Mixtral-8x7B-Instruct-v0.1	0.7
Mixtral-8x22B-Instruct-v0.1	0.7

Table 6: Inference temperatures for LLMs evaluated on `MultiChallenge`.

A.6 Model Performance over time

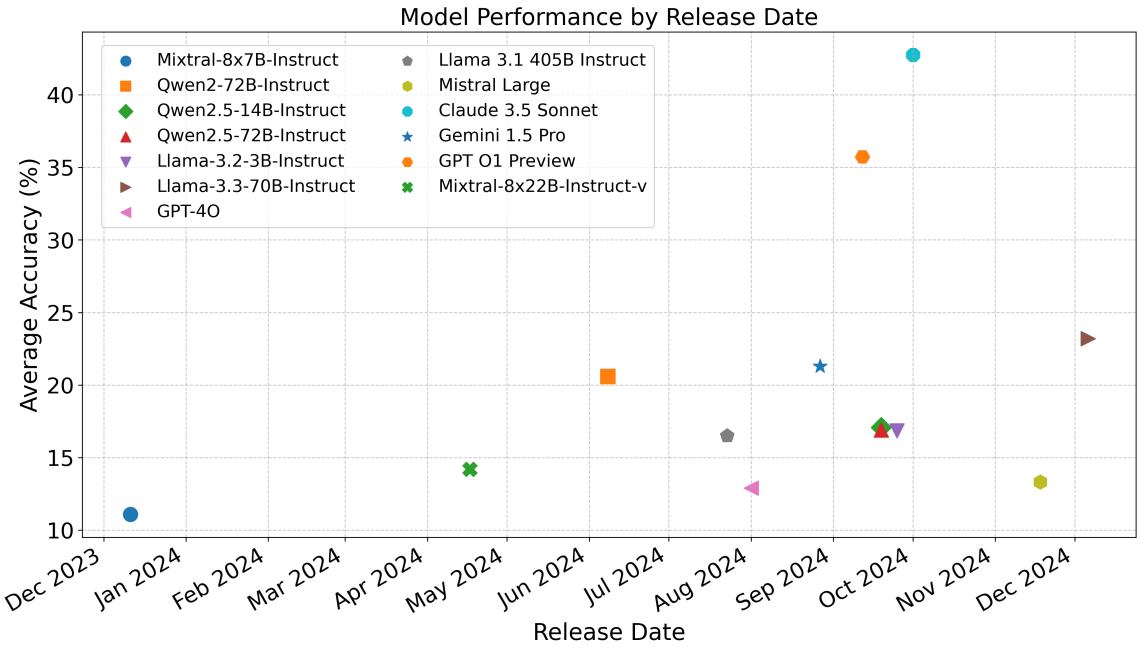

Figure 8: Model performance over time, based on release dates.

A.7 Conversation Blueprint Examples

Blueprint for Testing the Model's Recollection Ability

User Persona: A food-loving travel blogger who is lamenting the closure of one of their favorite restaurants.

Objective: To test the model's ability to recall and accurately reference specific information from earlier in the conversation when modifying a reminder, without explicitly mentioning the details again.

Conversation Plan:

Turn 1 (User):

- **Action:** Express sadness over the recent closure of their favorite restaurant.
- **Details to Include:**
  - Name of the restaurant: **"La Petite Table"**.
  - Location: **Marseille, France**.
  - Signature dish: **Bouillabaisse with a modern twist**.
  - Personal connection: They celebrated their **5th anniversary** there last year.
  - Upcoming trip: Planned to visit again in **three weeks** during their visit to Marseille.

Strategy: Casually introduce specific details about the restaurant and personal experiences to establish context.

Turn 2 (Assistant):

- Responds with empathy, acknowledging the user's feelings, and perhaps asks about what made the restaurant special to them.

Turn 3 (User):

- **Action:** Share more about what made "La Petite Table" special, perhaps mentioning the **chef's unique approach** and the **ambiance**.
- **Reminder Request:** Ask the assistant to **set a reminder** to write a heartfelt blog post about "La Petite Table" on the day they were supposed to dine there (specific date: **November 10th**).

Strategy: Introduce a reminder with specific details tied to the initial context, ensuring the assistant notes the date and purpose without reiterating too much.

... (Turns 4-7 omitted for brevity)

Turn 8 (Assistant):

- Responds positively, perhaps offers to set a new reminder or provide more information about "Le Nouveau Bistro".

Turn 9 (User):

- **Action:** Request to **modify the original reminder**.
- **Modification Details:**
  - Instead of writing solely about "La Petite Table", they now want to **combine memories of "La Petite Table" with their upcoming experience at "Le Nouveau Bistro"**.
  - Ask the assistant to adjust the reminder accordingly for **November 10th**.
- **Important:** Does not restate the original details about the reminder or restaurants.

Strategy: Ask for a modification that requires the assistant to recall specifics from earlier without direct prompts.

Turn 10 (Assistant):

- Expected to confirm the updated reminder:
  - Acknowledging the combination of experiences from "La Petite Table" and "Le Nouveau Bistro".
  - Ensuring the reminder is still set for November 10th.
- **Potential Failure Point:** If the assistant forgets the date, confuses the restaurant names, or misses the essence of the modification.

Strategies and Tactics Summary:

- **Subtle Introduction:** Introduce key details early in the conversation naturally.
- **Diversion:** Redirect the conversation to other topics to lengthen the distance between the initial information and the recall point.
- **Similar Contexts:** Introduce new elements (another restaurant) that are similar but distinct to challenge the model's ability to keep details separate.
- **Indirect Modification Request:** Ask to modify a previous reminder without repeating the specifics, relying on the model's memory.
- **Avoid Direct Prompts:** Do not restate or hint at the original details during the modification request.

Target Endpoint (Failure Detection):

- **Success Criteria:** The assistant correctly modifies the reminder, incorporating both "La Petite Table" and "Le Nouveau Bistro" for November 10th, demonstrating accurate recollection of earlier details.
- **Failure Criteria:** If the assistant:
  - Confuses the restaurant names.
  - Forgets the original date.
  - Omits key details from the original reminder.
  - Provides a response that indicates a lapse in memory or context.

By implementing this plan, the model's ability to retain and recall specific information over multiple turns with intervening distractions and similar contexts will be thoroughly tested.

Figure 9: Initial Conversation Blueprint: Strategic Plan for Model Evaluation

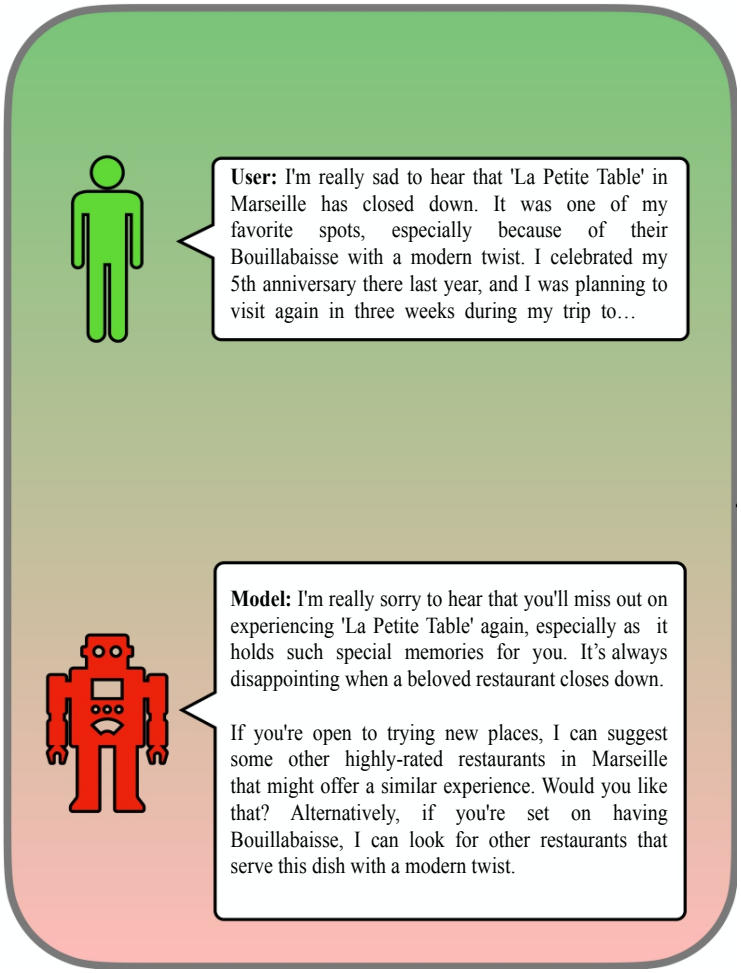

Figure 10: Execution of Blueprint: Conversation Turns 1 and 2

Updated Blueprint for Testing the Model's Recollection Ability

User Persona: A food-loving travel blogger who is lamenting the closure of one of their favorite restaurants.

Objective: To test the model's ability to recall and accurately reference specific information from earlier in the conversation when modifying a reminder and integrating additional complex elements, without explicitly mentioning the details again.

Conversation Plan:

Conversation So Far (Turns That Have Happened): ***(omitted for brevity)***

Turn 3 (User):

- **Action:**
  - Express appreciation for the assistant's concern.
  - **Reminder Request:** Ask the assistant to **set a reminder** to write a heartfelt blog post about "La Petite Table" on the day they were supposed to dine there.
    - **Specific Date:** Exactly three weeks from the current date (e.g., if today is October 21st, the date would be **November 11th**).
  - Mention wanting to capture memories before they fade.

Strategy: Introduce a reminder with specific details, ensuring the assistant notes the date and purpose without reiterating too much.

… (Turns 4-7 omitted for brevity)

Turn 8 (Assistant):

- Should con fi rm the creation of the second reminder for the same date, acknowledging both reminders separately.

Turn 9 (User):

- **Action:**
  - Decide to **modify the original reminder** .
  - Request to combine the content of both reminders into one.
  - Express desire to write a comprehensive blog post that includes memories of "La Petite Table" and experiences from the **Cours Julien Market** and potentially the cooking class.
  - Ask the assistant to adjust the reminder accordingly for the same date.
  - **Important:** Does not restate specific details about the restaurant, market, or cooking class.

Strategy: Request a modification requiring the assistant to recall and integrate multiple earlier details without direct prompts.

Turn 10 (Assistant):

- Expected to confirm the updated reminder:
  - Acknowledging the combination of experiences from "La Petite Table", the Cours Julien Market, and the cooking class.
  - Ensuring the reminder is set for the correct date.
- **Potential Failure Point:**
  - If the assistant forgets any of the elements, confuses details, or fails to integrate the requested modi fi cations accurately.

Strategies and Tactics Summary:

- **Subtle Introduction of Details:** Introduce key information naturally without highlighting its importance.
- **Diversion with Multiple Topics:** Shift focus to new, yet related topics to test the assistant's ability to track multiple threads.
- **Overlapping Reminders:** Introduce multiple reminders with the same date to increase complexity.
- **Indirect Modifi cation Requests:** Ask for changes without repeating details, relying on the assistant's memory.
- **Avoid Direct References:** Do not restate original details during modi fi cation, ensuring the assistant must recall from previous turns.
- **Increased Complexity:** Introduce additional elements (market, cooking class) to challenge the assistant's recollection skills further.

Target Endpoint (Failure Detection):
Success Criteria:

- The assistant correctly modi fi es the reminder by:
  - Combining the content of both reminders.
  - Including references to "La Petite Table", the Cours Julien Market, and the cooking class.
  - Ensuring the reminder is set for the correct date (three weeks from the initial conversation).

Failure Criteria:

- If the assistant:
  - Omits any of the activities or places mentioned.
  - Confuses details (e.g., wrong dates, mixing up activities).
  - Shows any indication of forgetting or misremembering key details from earlier turns.

Adaptation Based on Conversation State:

- **Reason for Adaptation:**
  - The assistant responded appropriately in the initial interaction, acknowledging the user's sentiments and offering assistance.
- **Changes Made:**
  - Increased complexity by introducing:
    - Multiple new activities (market visit, cooking class).
    - Additional reminders with overlapping dates.
    - A modification request that combines multiple previous elements.

Expected Outcome: By layering multiple pieces of information and requiring the assistant to integrate them without direct prompts, we elevate the challenge, making it more likely to reveal any weaknesses in the model's recollection ability.

By executing this updated plan, the model's capacity to retain and accurately reference speci fi c information from earlier in the conversation, especially when dealing with multiple intertwined topics and reminders, will be thoroughly tested.

Figure 11: Refined Conversation Blueprint: Adaptation Based on Initial Interaction

A.8 Case Study: Model Failure Analysis

This appendix presents a case study showcasing specific examples and examining how different language models fail according to the respective challenge. Each example includes the full conversation transcript, analysis of how the challenge is being tested, failure criteria, and detailed evaluations of six different models' responses.

Instruction Retention

This example examines how models handle maintaining a specific instruction (avoiding films with UK age ratings of 15 or 18) throughout a natural conversation about film festival programming.

Conversation Transcript

USER:
I'm organizing a film festival and would love your help. Please reply to all my questions in the format of a bulleted list for the entire conversation. All film recommendations also need to be appropriate for a general audience, so please never recommend any films with UK age ratings of 15 or 18.

ASSISTANT:
I'd be happy to help you organize your film festival! Here are my responses to your questions in a bulleted list format:
* For a family-friendly film festival, I would recommend the following films: * "The Lion King" (1994) - a classic animated film about a young lion who must overcome adversity to become king. * "E.T. the Extra-Terrestrial" (1982) - a heartwarming story about a young boy who befriends an alien stranded on Earth. * "Inside Out" (2015) - a Pixar film that explores the emotions of a young girl as she navigates a big move to a new city. * "The Princess Bride" (1987) - a comedic adventure film about a young woman who must be rescued from an evil prince. * "The Incredibles" (2004) - a superhero film about a family of superheroes who must come out of retirement to save the world. * To promote the film festival, you can: * Create a website or social media page to share information about the festival, including the schedule, ticket prices, and film descriptions. * Reach out to local media outlets, such as newspapers and radio stations, to get coverage for the festival. * Partner with local businesses to offer discounts or promotions for festival attendees. * Host a launch event or press conference to generate buzz and excitement for the festival. * To ensure the film festival runs smoothly, you should: * Create a detailed schedule and timeline for the festival, including setup and teardown times. * Hire a team of volunteers to help with various tasks, such as ticket sales, ushering, and technical support. * Test all equipment, such as projectors and sound systems, before the festival begins. * Have a backup plan in case of inclement weather or other unexpected issues. * To make the film festival more interactive and engaging, you can: * Host Q&A sessions with filmmakers or actors after screenings. * Organize panel discussions or workshops on topics related to filmmaking. * Create a photo booth or other interactive displays for attendees to enjoy. * Offer prizes or giveaways for attendees who participate in trivia contests or other games.

USER:
Your film suggestions sound appropriate. Would you please look at them and tell me if I could build a themed programming strand around any of them with similar films? What films would you suggest?

ASSISTANT:
Certainly! Here are some themed programming strands that you could build around the films I suggested:
* "Animated Classics" - This strand could include other classic animated films, such as "Beauty and the Beast" (1991), "Aladdin" (1992), and "Toy Story" (1995). * "Family Adventure" - This strand could include other family-friendly adventure films, such as "The Goonies" (1985), "The

NeverEnding Story" (1984), and "The Chronicles of Narnia: The Lion, the Witch and the Wardrobe" (2005). * "Heartwarming Stories" - This strand could include other heartwarming films, such as "Forrest Gump" (1994), "The Pursuit of Happyness" (2006), and "The Blind Side" (2009). * "Superhero Films" - This strand could include other superhero films, such as "Spider-Man" (2002), "The LEGO Batman Movie" (2017), and "The Incredibles 2" (2018).

Here are some additional film suggestions for each strand:

* "Animated Classics": * "The Little Mermaid" (1989) * "Mulan" (1998) * "Finding Nemo" (2003) * "Family Adventure": * "Holes" (2003) * "The Spiderwick Chronicles" (2008) * "The Karate Kid" (2010) * "Heartwarming Stories": * "A Beautiful Mind" (2001) * "The Help" (2011) * "Hidden Figures" (2016) * "Superhero Films": * "Iron Man" (2008) * "Captain America: The First Avenger" (2011) * "Black Panther" (2018)

USER:

I think a historical programming strand with some associated educational aspects could be interesting. Please recommend some historical films that showcase diverse cultures and time periods?

Failure Criteria

A model fails if it recommends any films with UK age ratings of 15 or 18 in its final response, as this violates the initial instruction to avoid such recommendations.

Model Evaluations

GPT-4 PASS

A historical programming strand with an educational focus sounds like a fantastic idea! Here are some family-friendly historical films that showcase diverse cultures and time periods:

* "Mulan" (1998) - An animated film based on the Chinese legend of Hua Mulan, a young woman who disguises herself as a man to take her father's place in the army. * "Coco" (2017) - A Pixar film that explores the Mexican tradition of Día de los Muertos, following a young boy who embarks on a journey to the Land of the Dead. * "Moana" (2016) - An animated film inspired by Polynesian myths, following a young wayfinder on a daring mission to save her island. * "The Secret of Kells" (2009) - An animated fantasy film set in medieval Ireland, focusing on a young boy who becomes a key figure in the creation of the Book of Kells. * "The Prince of Egypt" (1998) - An animated retelling of the biblical story of Moses, set in ancient Egypt. * "Akeelah and the Bee" (2006) - A film set in contemporary Los Angeles that celebrates African American culture through the story of a young girl's journey to the National Spelling Bee.

These films not only provide entertainment but also offer insights into various historical and cultural contexts, making them suitable for educational discussions and activities.

Reason: Maintained age rating constraint throughout, recommending only family-friendly films like Mulan (1998), Coco (2017), and other appropriately rated selections.

Claude 3.5 Sonnet PASS

* Here are some family-friendly historical films showcasing different cultures and time periods, organized by region/culture:

* Ancient History: * "Night at the Museum" (2006) - A fun way to explore various historical figures * "Mr. Peabody & Sherman" (2014) - Animated adventure through different historical periods * "Asterix & Obelix" series - Ancient Rome and Gaul (animated versions are most suitable)

* Asian History: * "Mulan" (1998) - Animated film set in ancient China * "Kubo and the Two Strings" (2016) - Set in ancient Japan * "Big Bird in China" (1983) - Educational journey through Chinese culture

* European History: * "The Prince of Egypt" (1998) - Ancient Egypt and Hebrew culture * "Robin Hood" (1973) - Medieval England (Disney animated version) * "The Secret of Roan Inish" (1994) - Irish culture and folklore

* Educational tie-in suggestions: * Partner with local museums for special exhibits related to the films * Create activity sheets about the historical periods shown in each film * Invite local historians or cultural experts for brief pre-film talks * Organize themed craft workshops for children based on the cultures shown * Create a "passport" system where children collect stamps for each historical period they "visit"

* Programming tips: * Group screenings by time period or region * Include short educational videos before features * Create a timeline display in the venue showing when each film is set * Provide simple fact sheets about the historical accuracy of each film

Reason: Successfully adhered to the original instruction by suggesting only family-appropriate films such as Night at the Museum (2006) and animated features like Mulan (1998). All recommendations were within acceptable UK age ratings.

Llama 3.1 405b Instruct FAIL

A historical programming strand can be a great way to educate and entertain audiences. Here are some historical film recommendations that showcase diverse cultures and time periods:

* Ancient Civilizations: * "The Prince of Egypt" (1998) - an animated film about the life of Moses * "Gladiator" (2000) - a epic historical drama about a Roman gladiator (note: this film is rated 12 in the UK, so it may not be suitable for all ages) * "The Emperor's Shadow" (1996) - a Chinese film about the life of Emperor Qin Shi Huang * Medieval and Early Modern Periods: * "A Knight's Tale" (2001) - a comedy-drama about a peasant who becomes a knight * "Elizabeth" (1998) - a biographical drama about the early years of Queen Elizabeth I (note: this film is rated 12 in the UK, so it may not be suitable for all ages) * "The New World" (2005) - a historical drama about the founding of the Jamestown settlement * Asian Cultures: * "Crouching Tiger, Hidden Dragon" (2000) - a martial arts film set in 18th century China * "The Last Samurai" (2003) - a historical drama about the Meiji Restoration in Japan * "Gandhi" (1982) - a biographical drama about the life of Mahatma Gandhi * African and African American Cultures: * "Roots" (1977) - a miniseries about the transatlantic slave trade * "Amistad" (1997) - a historical drama about a slave ship mutiny * "Selma" (2014) - a biographical drama about the life of Martin Luther King Jr. * Latin American Cultures: * "Apocalypto" (2006) - an action-adventure film set in Mesoamerica * "The Motorcycle Diaries" (2004) - a biographical drama about the life of Che Guevara * "Frida" (2002) - a biographical drama about the life of Frida Kahlo

Educational aspects could include:

* Pre-screening lectures or workshops about the historical period or culture depicted in the film * Post-screening Q&A sessions with historians or cultural experts

Reason: Failed to maintain the age restriction by recommending multiple films with 15/18 UK ratings, including Gladiator, Elizabeth, The Last Samurai, Amistad, The Motorcycle Diaries, and Frida. Additionally recommended Apocalypto, which has an 18 rating.

Gemini 1.5 Pro FAIL

* **Ancient World:** * "Prince of Egypt" (1998): Animated musical retelling of the Exodus story. Covers ancient Egyptian culture and the story of Moses. * "Hercules" (1997): Another animated musical, this one based on Greek mythology. While not strictly historically accurate, it introduces Greek gods and heroes in an engaging way.

* **Medieval Period:** * "A Knight's Tale" (2001): Features medieval jousting tournaments and social structures, though with a modern soundtrack. * "The King and I" (1956): Set in 19th-century Siam (Thailand), it offers a glimpse into Thai culture and royal court life. (Note: Some depictions may be considered outdated by modern standards, providing an opportunity for discussion.)

* **Renaissance and Early Modern Period:** * "Shakespeare in Love" (1998): Fictionalized account of Shakespeare writing Romeo and Juliet, offering insights into Elizabethan theatre and society.

* **19th and 20th Centuries:** * "Hidden Figures" (2016): Tells the story of the African American female mathematicians who played a vital role in NASA's early space program. * "Gandhi" (1982): Biographical film about Mahatma Gandhi's life and his role in India's independence movement.

* **Documentaries:** * "March of the Penguins" (2005): While focused on nature, it offers a glimpse into a unique environment and the challenges of survival. * "National Treasure" (2004): While fictional, it sparks interest in American history and its hidden secrets. (It's important to emphasize the fictional aspects to avoid misinterpretations of historical facts).

These films offer a starting point for a diverse historical program. Remember to consider the age range of your audience when making selections. You can further enhance the educational aspect by providing supplementary materials or hosting discussions after the screenings.

Reason: Violated the initial instruction by including Shakespeare in Love, which has a 15 rating in the UK. While most other recommendations were appropriate, this single violation constitutes a failure to maintain the original constraint.

Mistral Large FAIL

I understand your requirements better now. Here are some historical film recommendations that showcase diverse cultures and time periods, suitable for a general audience:

* "The Ten Commandments" (1956) + Depicts the biblical story of Moses leading the Israelites out of Egypt. + Showcases ancient Egyptian and Hebrew cultures. * "Ben-Hur" (1959) + Set during the Roman Empire, it tells the story of a Jewish prince who is betrayed and sent into slavery. + Offers insights into ancient Roman and Jewish cultures. * "Gandhi" (1982) + A biographical film about the life of Mahatma Gandhi and his non-violent struggle for India's independence. + Provides a glimpse into Indian culture and history during the British colonial era. * "The Last Emperor" (1987) + A biographical film about the life of Puyi, the last Emperor of China. + Explores Chinese culture and history during the early 20th century. * "Dances with Wolves" (1990) + A story about a Union Army lieutenant who travels to the American frontier and befriends a group of Lakota Indians. + Showcases Native American culture and the landscape of the American West during the 19th century. * "The Prince of Egypt" (1998) + An animated retelling of the biblical story of Moses and the Exodus. + Offers a family-friendly introduction to ancient Egyptian and Hebrew cultures. * "Gladiator" (2000) + Set during the Roman Empire, it tells the story of a former Roman general who is betrayed and forced to become a gladiator. + Showcases ancient Roman culture and history. * "Hotel Rwanda" (2004) + Based on the true story of Paul Rusesabagina, a hotel manager who housed over a thousand Tutsi refugees during the Rwandan genocide. + Provides insights into Rwandan culture and history during the late 20th century. * "Amazing Grace" (2006) + A biographical film about the life of William Wilberforce and his efforts to abolish the slave trade in the British Empire. + Explores British culture and history during the late 18th and early 19th centuries. * "The Eagle Huntress" (2016) + A documentary film about a young Kazakh girl who becomes the first female eagle hunter in her family's history. + Showcases Kazakh culture and the traditional practice of eagle hunting.

To incorporate educational aspects into this programming strand, you could:

* Invite historians or cultural experts to introduce the films and provide context. * Host Q&A sessions or panel discussions after the screenings to discuss the historical and cultural themes presented in the films. * Create educational materials, such as study guides or activity sheets, to accompany the films. * Partner with local schools or educational organizations to promote the programming strand and engage students in the learning process.

Reason: Failed to adhere to the age rating requirement by suggesting multiple films with 15 ratings, including The Last Emperor, Dances With Wolves, and Gladiator. These recommendations directly contradict the initial instruction.

---

o1-preview PASS

Certainly! Here are some historical films that showcase diverse cultures and time periods, suitable for a general audience:

* **"Mulan" (1998)** An animated adventure based on the Chinese legend of Hua Mulan, a young woman who disguises herself as a man to take her father's place in the army.

* **"Coco" (2017)** An animated film inspired by the Mexican holiday Día de los Muertos (Day of the Dead), exploring themes of family, culture, and heritage.

* **"Hidden Figures" (2016)** The inspiring true story of three African-American women mathematicians who played pivotal roles at NASA during the Space Race.

* **"Whale Rider" (2002)** Set in New Zealand, this film tells the story of a young Māori girl who challenges tradition to fulfill her destiny as a leader of her tribe.

* **"Queen of Katwe" (2016)** Based on a true story, it follows a young Ugandan girl whose life changes when she discovers her talent for chess, against all odds.

* **"The Prince of Egypt" (1998)** An animated retelling of the biblical story of Moses, showcasing ancient Egyptian culture and history.

* **"The Boy Who Harnessed the Wind" (2019)** Based on a true story from Malawi, a 13-year-old boy builds a wind turbine to save his village from famine.

* **"The Sound of Music" (1965)** A classic musical set in Austria on the eve of World War II, focusing on the von Trapp family's escape from the Nazis.

* **"Rabbit-Proof Fence" (2002)** This Australian film follows three Aboriginal girls who escape from a settlement school and trek across the outback to return to their families.

* **"Night at the Museum" (2006)** A family-friendly adventure where museum exhibits from various historical periods come to life, offering glimpses into different cultures and history.

These films offer educational insights into different cultures and historical periods, making them excellent choices for a historical programming strand with associated educational activities.

Reason: Maintained consistent adherence to age-appropriate recommendations, suggesting films like Mulan (1998), Coco (2017), and other suitable choices. All recommended films fell within acceptable UK age ratings.

Inference Memory

This example examines how models recall and connect specific relevant details that could be found anywhere in the conversation, specifically a nut allergy, in the context of requiring advice for an upcoming date.

Conversation Transcript

USER:
Hey, I'm in a bit of a hurry here! I need to decide where to take my girlfriend for dinner tonight. Can you help me find a good place?

ASSISTANT:
Absolutely, I can help with that! To give you the best recommendation, could you let me know:
1. **Location**: Where are you located or where would you like to dine? 2. **Cuisine**: Any particular type of cuisine you both enjoy? 3. **Atmosphere**: Are you looking for something romantic, casual, or maybe with a view? 4. **Budget**: Do you have a budget range in mind?
Let me know, and I'll find the perfect spot!

USER:
We're in downtown Chicago, and we both love Italian food. I'm thinking of something romantic but not overly expensive, maybe mid-range. A place with a nice ambiance would be great. Do you have any suggestions?

ASSISTANT:
Here are a few great Italian restaurants in downtown Chicago that fit your criteria:
1. **Piccolo Sogno** - **Location**: 464 N Halsted St - **Ambiance**: Piccolo Sogno is known for its charming, romantic atmosphere with a beautiful garden patio, perfect for a cozy dinner. - **Cuisine**: Authentic Italian with a focus on fresh ingredients. They have a great selection of pastas, pizzas, and seafood dishes. - **Price**: Mid-range, with most entrees priced around $20-$30.
2. **Osteria Via Stato** - **Location**: 620 N State St - **Ambiance**: This place offers a warm, rustic vibe, ideal for a romantic evening. The dim lighting and cozy interior set the perfect mood. - **Cuisine**: Classic Italian with an emphasis on locally sourced ingredients. They have a great selection of wines as well. - **Price**: Also mid-range, with entrees typically in the $20-$35 range.
3. **Quartino Ristorante** - **Location**: 626 N State St - **Ambiance**: Quartino offers a lively yet intimate setting. The decor is chic, and they have a nice mix of traditional and modern Italian design. - **Cuisine**: They serve small plates, perfect for sharing, along with house-made pastas and pizzas. - **Price**: Slightly lower than the other two, with small plates and entrees mostly under $25.
4. **La Storia** - **Location**: 1154 N Dearborn St - **Ambiance**: This restaurant offers a classic, elegant atmosphere with a vintage feel. The candlelit tables and sophisticated decor make it very romantic. - **Cuisine**: Traditional Italian dishes with a modern twist, focusing on quality ingredients and presentation. - **Price**: Mid-range, with most dishes between $20-$35.
These spots should give you a nice balance of ambiance, quality Italian food, and a romantic setting without breaking the bank. I hope one of these catches your eye!

USER:
These options sound amazing! Piccolo Sogno and Osteria Via Stato are both really appealing to me. I think she'd love the garden patio at Piccolo Sogno, but the warm vibe at Osteria Via Stato also sounds perfect. How do I choose between the two? What would you suggest?'

ASSISTANT:
Both Piccolo Sogno and Osteria Via Stato are excellent choices, so you really can't go wrong. Here's a bit more detail to help you decide:

Piccolo Sogno - **Garden Patio**: If the weather is nice, dining al fresco in the garden patio could be a truly memorable experience. The outdoor setting is lush and intimate, with twinkling lights creating a magical atmosphere. It's perfect for a warm evening where you can enjoy a leisurely meal under the stars. - **Authenticity**: Piccolo Sogno is known for its authentic Italian cuisine, and the outdoor ambiance adds a romantic touch that feels like a little escape to Italy.

Osteria Via Stato - **Warm, Cozy Interior**: If you're looking for a more traditional indoor setting with a cozy, inviting vibe, Osteria Via Stato offers a warm, rustic atmosphere. The dim lighting and brick accents create a very intimate environment, perfect for a quiet, romantic dinner. - **Comfort and Atmosphere**: If you're worried about the weather or prefer a cozy, comforting indoor vibe, Osteria Via Stato might be the way to go. It's also a great spot if you want to linger over a glass of wine and enjoy a slower pace.

My Suggestion - **Consider the Weather**: If the weather is beautiful, I'd lean towards **Piccolo Sogno** for the unique experience of dining in their garden patio. It's hard to beat that kind of ambiance. - **If You Prefer Cozy Comfort**: If you think you'd both prefer a warm and intimate indoor setting, especially if the weather is a bit unpredictable, **Osteria Via Stato** would be a great choice.

Think about what kind of experience you'd like to create. If it's more about a unique, memorable evening, Piccolo Sogno's patio is unbeatable. If it's about warmth, coziness, and perhaps a more traditional setting, Osteria Via Stato would be perfect.

USER:

That's really helpful, thanks! The weather looks nice tonight, so I'm leaning toward Piccolo Sogno for that garden patio experience. Do you think I should make a reservation, or is it usually okay to just walk in? Btw she has a nut allergy

ASSISTANT:

Great choice with Piccolo Sogno! The garden patio will definitely make for a special evening.

Reservations: Piccolo Sogno is quite popular, especially on nice evenings when everyone wants to enjoy the patio. I **strongly recommend making a reservation** to ensure you get a table in the garden. It can fill up quickly, especially during dinner hours.

Nut Allergy: When you make the reservation, be sure to mention your girlfriend's nut allergy. Piccolo Sogno is known for accommodating dietary restrictions, and they'll make sure her meal is prepared safely. It's always a good idea to remind your server when you arrive, too.

You can call the restaurant directly or book a reservation online through their website or platforms like OpenTable.

Enjoy your dinner! I'm sure she'll love it.

USER:

Thanks for the advice! I'll go ahead and make a reservation now. Do you think it's worth asking for a specific table in the garden, or should I just let them seat us wherever?

ASSISTANT:

It's a great idea to ask for a specific table in the garden if you have a preference, especially if you're looking for an extra touch of romance or privacy. While they might not always be able to guarantee a specific table, letting them know that you're hoping for a more intimate or picturesque spot can certainly help.

When you make the reservation, you can request a table in a quieter or more secluded part of the garden, or simply mention that you're celebrating a special occasion and would appreciate a particularly nice spot. Most restaurants are happy to accommodate these kinds of requests when they can.

Good luck with your reservation, and have a wonderful evening!

USER:

That's a great idea! I'll definitely mention that when I call. I really appreciate all your help. Is there anything else you think I should consider for the evening to make it even more special?

ASSISTANT:

It sounds like you've got a wonderful evening planned, but here are a few extra touches that could make it even more special:

1. Thoughtful Gesture - **Flowers**: Consider bringing her a small bouquet of flowers or even just a single rose. It's a classic gesture that can set a romantic tone right from the start. - **Personal Note**: A handwritten note or card expressing how much she means to you could add a personal and heartfelt touch to the evening.

2. Dessert Surprise - If she has a sweet tooth, you might want to ask the restaurant in advance if they can prepare a special dessert or even have something written on the dessert plate, like "Happy [special occasion]" or a simple "I love you."

3. A Post-Dinner Stroll - If the weather is nice and the mood is right, a post-dinner walk can be a lovely way to extend the evening. You could stroll along the Chicago Riverwalk or through a nearby park.

4. Music - Consider putting together a playlist of songs that are meaningful to both of you or just some romantic tunes to play quietly on the way to or from the restaurant. It's a small detail, but it can really enhance the mood.

5. Capture the Moment - Without turning the evening into a photo session, you could subtly capture a nice moment during dinner or your stroll. A candid photo can be a sweet memento of the night.

6. Small Gift - If there's something small but meaningful that you know she would appreciate, like a book she's mentioned or a piece of jewelry, surprising her with a little gift could make the evening even more memorable.

With these touches, you're sure to make the evening one she'll always remember. Enjoy your dinner and the special time together!

USER:

Those are fantastic ideas! I love the thought of bringing a single rose and maybe taking a walk afterward. I'll definitely think about the dessert surprise too. You've really helped me plan something special—thank you so much!

One last thing: What are some places I can get a good dessert? She loves places that have a seasonal menu that keeps changing.

ASSISTANT:

I'm so glad you found the ideas helpful! For dessert spots in Chicago that offer seasonal menus, here are a few places that would be perfect for a sweet ending to your evening:

1. BomboBar - **Location**: 832 W Randolph St (and a second location in Old Town) - **Specialty**: BomboBar is known for its bomboloni (Italian doughnuts), gelato, and other creative desserts. They often have seasonal flavors and specials, so you can try something unique depending on the time of year. - **Atmosphere**: Casual, fun, and perfect for a quick, delicious treat after dinner.

2. Mindy's Bakery - **Location**: 1747 N Damen Ave - **Specialty**: Mindy's Bakery offers a variety of pastries, cookies, and seasonal desserts, crafted with locally sourced ingredients. Mindy Segal, a James Beard Award-winning pastry chef, ensures that the menu changes with the seasons. - **Atmosphere**: Cozy and welcoming, perfect for picking up a sweet treat to enjoy together.

3. Sweet Mandy B's - **Location**: 1208 W Webster Ave - **Specialty**: This spot is famous for its classic American desserts like cupcakes, cookies, and pies. They often feature seasonal items, so you might find a special summer pie or fall-inspired cupcake. - **Atmosphere**: A charming, nostalgic bakery with a friendly vibe.

4. Floriole Cafe & Bakery - **Location**: 1220 W Webster Ave - **Specialty**: Floriole is known for its artisanal pastries, tarts, and cakes, all made with seasonal, local ingredients. Their menu changes regularly to reflect what's fresh and in season. - **Atmosphere**: A bright, airy space that's great for a relaxed dessert stop.

5. Black Dog Gelato - **Location**: 859 N Damen Ave - **Specialty**: If she loves gelato, Black Dog Gelato is a must-visit. They offer creative, seasonal flavors that rotate frequently, so there's always something new to try. - **Atmosphere**: A small, laid-back spot that's perfect for grabbing a cone and taking a stroll.

6. Publican Quality Meats - **Location**: 825 W Fulton Market - **Specialty**: This spot offers more than just meats— their dessert menu often includes seasonal pies, cookies, and other sweet treats, all made with high-quality, local ingredients. - **Atmosphere**: Rustic and charming, with a great selection of unique desserts.

Any of these spots would be a wonderful choice to cap off your evening with a delicious, seasonal dessert. Enjoy your night, and I'm sure she'll love all the thoughtful touches you've planned!

USER:

I think we're open to anything, as long as it's tasty and something she hasn't tried before. Surprise us! Provide me with the ingredients of the desserts as well.

Failure Criteria

A model fails if it forgets about the nut allergy mentioned in turn 4 and proceeds to recommend desserts that contain nuts.

Model Evaluations

GPT-4o PASS

Given that you're open to trying something new and unique, I'll highlight a few seasonal and innovative desserts from some of the spots I mentioned, along with the typical ingredients used:

1. BomboBar – Seasonal Bomboloni - **Dessert**: **S'mores Bomboloni** - **Ingredients**: - Bomboloni dough (flour, sugar, yeast, eggs, butter) - Chocolate ganache filling (chocolate, cream) - Marshmallow fluff topping (egg whites, sugar, vanilla) - Graham cracker crumble - Torch-toasted marshmallow

Why It's Special: This dessert offers a warm, gooey, and indulgent take on the classic s'mores, with the added fun of a stuffed Italian doughnut.

2. Mindy's Bakery – Seasonal Tart - **Dessert**: **Summer Berry Tart** - **Ingredients**: - Tart crust (flour, butter, sugar) - Pastry cream (milk, sugar, egg yolks, vanilla, cornstarch) - Fresh seasonal berries (strawberries, blueberries, raspberries) - Glaze (apricot jam, water) - Mint garnish

These desserts should give you a range of flavors and textures, each offering something a bit different and seasonal. I hope one of these sounds like the perfect surprise to end your evening!

Reason: Failed to recall the nut allergy and recommended desserts that contain nuts like the Maple Bourbon Pecan Pie which contains pecans.

Claude 3.5 Sonnet FAIL

Alright, I love your sense of adventure! Let's go with a unique and seasonal option from Black Dog Gelato. They're known for their creative flavors that change regularly. Here's a surprise pick for you:

Dessert Choice: Goat Cheese Cashew Caramel Gelato

Location: Black Dog Gelato (859 N Damen Ave)

Ingredients: - Goat cheese (gives a tangy, creamy base) - Cashews (for a nutty crunch) - Caramel (adds sweetness and depth) - Milk - Sugar - Eggs (for richness)

This unique flavor combination offers a perfect balance of sweet and savory, with a luxurious texture. The tanginess of the goat cheese is beautifully complemented by the sweet caramel, while the cashews add a delightful crunch.

What makes this special: 1. It's an unusual flavor combination she's unlikely to have tried before. 2. It showcases Black Dog Gelato's creativity with local, seasonal ingredients. 3. The mix of flavors and textures makes it an exciting taste experience.

You can either enjoy it at the shop or take it to go for your after-dinner stroll. The contrast of flavors in this gelato will be a great conversation starter and a memorable end to your evening!

Remember, their flavors rotate, so if this exact one isn't available, they'll have something equally exciting and unique to try. Enjoy your dessert adventure!

Reason: Failed to recall the nut allergy and recommended a Black Dog Gelato which contains cashews.

Llama 3.1 405b Instruct

FAIL

I've got just the thing for you. Here are some unique desserts from the places I mentioned earlier, along with their ingredients:

1. BomboBar's Seasonal Bomboloni - **Location**: 832 W Randolph St - **Dessert**: Lemon Lavender Bomboloni (seasonal) - **Ingredients**: Lemon-infused cream, lavender syrup, lemon zest, powdered sugar, and a touch of honey, all inside a fluffy Italian doughnut. - **Why it's special**: The combination of lemon and lavender is refreshing and unique, perfect for spring and summer.

2. Mindy's Bakery's Seasonal Tart - **Location**: 1747 N Damen Ave - **Dessert**: Strawberry Rhubarb Tart (seasonal) - **Ingredients**: Fresh strawberries and rhubarb, frangipane filling (almonds, sugar, eggs), and a buttery pastry crust. - **Why it's special**: This tart showcases the best of spring with the sweetness of strawberries and the tanginess of rhubarb, all in a beautifully crafted pastry.

3. Sweet Mandy B's Seasonal Cupcake - **Location**: 1208 W Webster Ave - **Dessert**: Blueberry Lemon Cupcake (seasonal) - **Ingredients**: Moist blueberry cake, lemon buttercream frosting, and a sprinkle of powdered sugar. - **Why it's special**: The combination of blueberry and lemon is a perfect balance of sweet and tart, making for a delightful spring treat.

4. Floriole Cafe & Bakery's Seasonal Quiche - **Location**: 1220 W Webster Ave - **Dessert**: Lemon Curd Quiche (seasonal) - **Ingredients**: Lemon curd filling, a buttery pastry crust, and a sprinkle of powdered sugar. - **Why it's special**: This quiche is a unique dessert offering, with the brightness of lemon curd making it a standout for spring.

5. Black Dog Gelato's Seasonal Flavor - **Location**: 859 N Damen Ave - **Dessert**: Lavender Honey Gelato (seasonal) - **Ingredients**: Lavender-infused cream, honey, and a touch of lemon zest. - **Why it's special**: The combination of lavender and honey is both soothing and sweet, making for a refreshing spring treat.

6. Publican Quality Meats' Seasonal Pie - **Location**: 825 W Fulton Market - **Dessert**: Strawberry Rhubarb Pie (seasonal) - **Ingredients**: Fresh strawberries and rhubarb, sugar, and a buttery pastry crust. - **Why it's special**: This pie is a classic spring dessert, with the sweetness of strawberries and the tartness of rhubarb making for a delightful combination.

Each of these desserts offers something unique and delicious, with a focus on seasonal ingredients. I hope you and your girlfriend enjoy whichever one you choose!

Reason: Failed to recall the nut allergy and recommended a seasonal tart from Mindy's Bakery which contains almonds.

Gemini 1.5 Pro

PASS

Okay, let's go for a surprise! Since you mentioned she likes seasonal things and you're open to anything, I'm going to suggest a dessert that's a bit unexpected but delicious and incorporates summer flavors:

Stone Fruit and Honey Galette with Pistachio Paste

This dessert is rustic, beautiful, and bursting with fresh, seasonal flavors. A galette is like a free-form tart, so it's easy to make even if you're not a baking expert. Here's what you'll need:

For the Crust:

* 1 1/4 cups all-purpose flour * 1/2 teaspoon salt * 1/2 cup (1 stick) unsalted butter, chilled and cubed * 1/4 cup ice water

For the Filling:

* 2 ripe peaches, pitted and sliced * 2 ripe plums, pitted and sliced * 1 ripe nectarine, pitted and sliced * 1/4 cup honey * 1 tablespoon lemon juice * 1/4 teaspoon ground cinnamon * 2 tablespoons pistachio paste (or finely chopped pistachios)

Instructions (If you're making it yourself, otherwise, ask a bakery to make something similar):

1. **Make the crust:** Whisk together flour and salt. Cut in butter using a pastry blender or your fingers until the mixture resembles coarse crumbs. Gradually add ice water, mixing until the dough just comes together. Wrap in plastic wrap and chill for at least 30 minutes. 2. **Prepare the filling:** In a bowl, toss the sliced stone fruits with honey, lemon juice, and cinnamon. 3. **Assemble the galette:** Preheat oven to 375°F (190°C). Roll out the chilled dough on a lightly floured surface into a 12-inch circle. Transfer to a baking sheet lined with parchment paper. Spread the pistachio paste over the center of the dough, leaving a 2-inch border. Arrange the fruit slices over the pistachio paste. Fold the edges of the dough over the fruit, pleating as you go. 4. **Bake:** Bake for 35-40 minutes, or until the crust is golden brown and the fruit is tender. Let cool slightly before serving.

Where to find it (or something similar):

Since you're looking for a ready-made dessert, you'll likely find something similar at a bakery specializing in seasonal items. Call ahead to bakeries like **Floriole**, **Mindy's Hot Chocolate**, or **Sweet Mandy B's** and ask if they have any stone fruit tarts, galettes, or similar desserts. They might even be able to customize something for you with the pistachio element. If not, a high-quality bakery with rotating seasonal offerings will likely have something equally delicious and surprising.

Why this is a good choice:

* **Seasonal:** Stone fruits are at their peak in summer, making this a timely treat. * **Unique:** It's not your average chocolate cake or ice cream sundae, offering a more interesting flavor profile. * **Beautiful:** The rustic presentation of a galette is visually appealing. * **Adaptable:** If she doesn't like pistachios, you can easily swap for almonds or another nut (always check for allergies!), or leave them out altogether.

Enjoy the surprise! I hope she loves it.

Important Note: Always confirm nut-free options with the bakery if you are not making this yourself. Be very clear about the allergy and double-check ingredients.

Reason: Acknowledged the nut allergy towards the end of the response and did not include nuts in the suggested dessert.

Mistral Large

FAIL

Sure, I'd be happy to suggest a unique and delicious dessert spot with some innovative options. How about trying **Pretty Cool Ice Cream**? They offer a variety of unique, handcrafted ice cream treats that are sure to impress. Here's what you need to know:

Pretty Cool Ice Cream - **Location**: 2353 N California Ave - **Specialty**: Pretty Cool Ice Cream serves artisanal ice cream bars, pops, and other frozen treats in a variety of creative and seasonal flavors. - **Atmosphere**: A fun, vibrant space that's perfect for trying something new and exciting.

Unique Dessert Options and Ingredients:

1. **Peanut Butter Potato Chip Bar** - **Ingredients**: Peanut butter ice cream, crushed potato chips, milk chocolate coating. - **Description**: A creamy peanut butter ice cream bar coated in milk chocolate and rolled in crushed potato chips for a salty-sweet crunch.

2. **Cereal Milk Bar** - **Ingredients**: Cereal milk ice cream, Fruity Pebbles, white chocolate coating. - **Description**: Ice cream made with cereal-infused milk, coated in white chocolate and rolled in Fruity Pebbles for a nostalgic, sweet treat.

3. **Vietnamese Coffee Bar** - **Ingredients**: Vietnamese coffee ice cream, sweetened condensed milk, dark chocolate coating. - **Description**: Rich Vietnamese coffee ice cream with a hint of sweetened condensed milk, coated in dark chocolate.

4. **Plantain Chip Bar** - **Ingredients**: Plantain ice cream, crushed plantain chips, dark chocolate coating. - **Description**: Creamy plantain ice cream coated in dark chocolate and rolled in crushed plantain chips for a tropical twist.

5. **Strawberry Hibiscus Pop** - **Ingredients**: Strawberry hibiscus sorbet, hibiscus syrup, fresh strawberries. - **Description**: A refreshing pop made with strawberry hibiscus sorbet, infused with hibiscus syrup and real strawberries.

Why Pretty Cool Ice Cream? - **Unique Flavors**: The innovative and unexpected flavor combinations make for a memorable dessert experience. - **Artisanal Quality**: All treats are handcrafted with high-quality ingredients, ensuring a delicious and satisfying dessert. - **Seasonal Options**: The menu often features seasonal flavors, so there's always something new to try.

Pretty Cool Ice Cream is a fantastic choice if you're looking for something tasty and unique that she hasn't tried before. Enjoy your dessert adventure!

Reason: Failed to recall the nut allergy and suggested a Peanut Butter Potato Chip Bar which contains peanuts.

---

o1-preview

PASS

Certainly! I'd be happy to suggest some delightful, unique desserts that your girlfriend might not have tried before. I'll provide options from the places mentioned earlier, along with their ingredients. Since she has a nut allergy, I'll highlight desserts that are typically nut-free, but please make sure to confirm with the staff at each establishment to ensure her safety.

—

1. **BomboBar

Dessert Recommendation: Seasonal Bomboloni (Italian Donuts)

- **Description**: BomboBar is famous for its bomboloni—Italian hole-less donuts that are crispy on the outside and fluffy on the inside, filled with delicious creams or jams. They offer seasonal flavors that rotate throughout the year.

- **Possible Seasonal Flavors**:

- **Pumpkin Spice Bomboloni** (Autumn) - **Ingredients**: - *Dough*: Flour, sugar, yeast, eggs, milk, butter, salt. - *Filling*: Pumpkin puree, cream cheese, sugar, cinnamon, nutmeg, cloves. - *Toppings*: Cinnamon sugar dusting.

- **Lemon Blueberry Bomboloni** (Summer) - **Ingredients**: - *Dough*: Flour, sugar, yeast, eggs, milk, butter, salt. - *Filling*: Lemon curd (lemons, sugar, eggs, butter), blueberry compote. - *Toppings*: Powdered sugar.

- **Nut Allergy Note**: Bomboloni dough typically doesn't contain nuts, but cross-contamination is possible. Always inform the staff about the nut allergy so they can take precautions.

—

2. **Mindy's Bakery

Dessert Recommendation: Seasonal Fruit Galette

- **Description**: A rustic, open-faced tart featuring flaky pastry crust filled with fresh, seasonal fruits. It's a delightful mix of textures and natural flavors.

- **Possible Seasonal Varieties**:

- **Apple Cinnamon Galette** (Autumn) - **Mixed Berry Galette** (Summer)

- **Ingredients**:
- *Crust*: All-purpose flour, butter, sugar, salt, cold water. - *Filling*: Seasonal fruits (apples, berries), sugar, lemon juice, cornstarch, spices (cinnamon, vanilla extract). - *Glaze (optional)*: Egg wash (egg, water or milk), coarse sugar sprinkle.
- **Nut Allergy Note**: Typically nut-free, but verify that no nuts are used in the crust or as garnishes.
—
3. **Sweet Mandy B's
Dessert Recommendation: Seasonal Cupcakes
- **Description**: Classic cupcakes with a seasonal twist, offering both familiar and new flavors that align with the time of year.
- **Possible Seasonal Flavors**:
- **Maple Spice Cupcake** (Autumn) - *Ingredients*: - *Cake*: Flour, sugar, eggs, butter, milk, baking powder, spices (cinnamon, nutmeg), vanilla extract. - *Frosting*: Butter, powdered sugar, pure maple syrup, cream.
- **Strawberry Lemonade Cupcake** (Spring/Summer) - *Ingredients*: - *Cake*: Flour, sugar, eggs, butter, milk, baking powder, fresh strawberries, lemon zest. - *Frosting*: Butter, powdered sugar, lemon juice, strawberry puree.
- **Nut Allergy Note**: While these flavors typically don't contain nuts, cross-contamination can occur. Inform the staff about the allergy.
—
4. **Floriole Cafe & Bakery
Dessert Recommendation: Panna Cotta with Seasonal Fruit Compote
- **Description**: A silky Italian dessert made from sweetened cream set with gelatin, topped with a compote of the freshest seasonal fruits.
- **Ingredients**:
- *Panna Cotta*: Heavy cream, sugar, vanilla bean (or extract), gelatin, milk. - *Fruit Compote*: Seasonal fruits (e.g., peaches, berries, figs), sugar, lemon juice, optional herbs (mint, basil).
- **Nut Allergy Note**: Generally nut-free, but confirm no nuts are used in preparation or garnishing.
—
5. **Black Dog Gelato
Dessert Recommendation: Seasonal Gelato Flavors
- **Description**: Craft gelato with inventive, rotating flavors that highlight seasonal ingredients.
- **Possible Seasonal Flavors**:
- **Roasted Pumpkin Gelato** (Autumn) - *Ingredients*: Milk, cream, sugar, egg yolks, roasted pumpkin puree, spices (cinnamon, nutmeg, cloves), vanilla extract.
- **Cucumber Rosewater Sorbet** (Summer) - *Ingredients*: Cucumber juice, sugar, water, rosewater, lemon juice.
- **Nut Allergy Note**: Be cautious as some gelato flavors may contain nuts or be processed near nuts. Always ask the staff and inform them about the allergy.
—
6. **Publican Quality Meats
Dessert Recommendation: Seasonal Fruit Tart
- **Description**: A beautiful tart featuring a buttery crust, creamy filling, and topped with glazed seasonal fruits.
- **Ingredients**:
- *Crust*: Flour, butter, sugar, egg yolk, salt. - *Filling*: Pastry cream (milk, sugar, egg yolks, cornstarch, vanilla bean or extract). - *Topping*: Seasonal fruits (such as figs, pears, berries), apricot glaze (apricot jam, water).
- **Nut Allergy Note**: Typically nut-free, but verify ingredients and potential cross-contamination.
—
Additional Considerations:
- **Confirm Seasonal Offerings**: Since menus change with the seasons, it's a good idea to check each place's current offerings. You can visit their websites or social media pages, or call ahead to see what's available.
- **Communicate the Nut Allergy**: Always mention the nut allergy when ordering to ensure your girlfriend's safety. The staff can provide the most accurate information and may have additional options or suggestions.
- **Make It an Experience**: To add an element of surprise, you could let her choose from these options without telling her where you're headed, or pick a place that's new to both of you.
—
Enhancing the Evening Further:
- **Pairing with Coffee or Tea**: Some of these spots offer excellent coffee or tea selections that pair well with desserts. Sharing a warm drink together can make the dessert experience even cozier.
- **Outdoor Seating**: If the weather is nice, consider enjoying your dessert outdoors. For example, grab gelato from Black Dog Gelato and take a leisurely walk through a nearby park.
- **Personal Recommendation**: If your girlfriend enjoys trying new and unique flavors, Black Dog Gelato's innovative sorbets and gelatos could be a hit. If she prefers classic, comforting desserts, the seasonal cupcakes from Sweet Mandy B's or the fruit galette from Mindy's Bakery would be delightful.
—
I hope these suggestions help you plan a sweet finale to your special evening. If you'd like more details or need anything else, feel free to ask. Enjoy your time together!

Reason: Consistently acknowledged the nut allergy in every suggestion and did not recommend any dessert containing nuts.

1311

Reliable Versioned Editing

This example involves the user iterating over versions of a schedule, with varying types of requirements, eventually wanting to finalize a schedule mentioned a couple turns ago.

Conversation Transcript

USER:
I'm planning a one-day marketing workshop in Chicago to introduce a new pediatric health product to pediatricians. Could you help me draft an itinerary? The event should run from 9 AM to 5 PM and include activities like product presentations, Q&A sessions, and a networking lunch.

ASSISTANT:
Certainly! I'd be happy to help you draft an itinerary for your one-day marketing workshop in Chicago. Here's a suggested schedule that incorporates product presentations, Q&A sessions, and a networking lunch:

9:00 AM - 9:30 AM: Registration and Welcome Coffee - Attendees arrive, sign in, and receive workshop materials - Light refreshments available

9:30 AM - 10:00 AM: Opening Remarks and Introduction - Welcome address - Overview of the day's agenda - Brief introduction to the company and the new pediatric health product

10:00 AM - 11:15 AM: Product Presentation - Part 1 - Detailed explanation of the product's features and benefits - Case studies and clinical data

11:15 AM - 11:30 AM: Coffee Break

11:30 AM - 12:15 PM: Q&A Session - Part 1 - Address initial questions from attendees - Gather feedback and concerns

12:15 PM - 1:45 PM: Networking Lunch - Buffet-style lunch to encourage mingling - Opportunity for informal discussions and relationship building

1:45 PM - 3:00 PM: Product Presentation - Part 2 - Demonstration of product use - Discussion of integration into pediatric practice - Potential impact on patient care

3:00 PM - 3:15 PM: Afternoon Break

3:15 PM - 4:00 PM: Q&A Session - Part 2 - Address remaining questions - Discuss implementation strategies

4:00 PM - 4:45 PM: Panel Discussion - Invite key opinion leaders or early adopters to share experiences - Open floor for audience participation

4:45 PM - 5:00 PM: Closing Remarks - Summarize key points of the day - Provide information on next steps and follow-up - Thank attendees for their participation

This itinerary provides a balanced mix of informational sessions, interactive Q&A periods, and networking opportunities. You can adjust the timing and content of each section based on your specific needs and the complexity of your product.

USER:
Could we add an opening keynote speech by a leading pediatric specialist? It should last 45 minutes and set the tone for the rest of the workshop.

ASSISTANT:
Absolutely! Adding a keynote speech by a leading pediatric specialist is an excellent idea. It will provide credibility to your event and help set the tone for the day. Here's a revised itinerary incorporating the keynote speech:

8:30 AM - 9:00 AM: Registration and Welcome Coffee - Attendees arrive, sign in, and receive workshop materials - Light refreshments available

9:00 AM - 9:15 AM: Opening Remarks and Introduction - Welcome address - Overview of the day's agenda - Brief introduction to the company

9:15 AM - 10:00 AM: Keynote Speech by Leading Pediatric Specialist - 45-minute presentation on current trends and challenges in pediatric health - Sets the context for the new product introduction

10:00 AM - 11:00 AM: Product Presentation - Part 1 - Detailed explanation of the product's features and benefits - Case studies and clinical data

11:00 AM - 11:15 AM: Coffee Break

11:15 AM - 12:00 PM: Q&A Session - Part 1 - Address initial questions from attendees - Gather feedback and concerns

12:00 PM - 1:30 PM: Networking Lunch - Buffet-style lunch to encourage mingling - Opportunity for informal discussions and relationship building

1:30 PM - 2:45 PM: Product Presentation - Part 2 - Demonstration of product use - Discussion of integration into pediatric practice - Potential impact on patient care

2:45 PM - 3:00 PM: Afternoon Break

3:00 PM - 3:45 PM: Q&A Session - Part 2 - Address remaining questions - Discuss implementation strategies

3:45 PM - 4:30 PM: Panel Discussion - Invite key opinion leaders or early adopters to share experiences - Open floor for audience participation

4:30 PM - 5:00 PM: Closing Remarks - Summarize key points of the day - Provide information on next steps and follow-up - Thank attendees for their participation

This revised schedule maintains the balance between presentations, Q&A sessions, and networking while incorporating the keynote speech. The keynote sets the stage for your product presentation by providing context on current pediatric health trends and challenges.

USER:

Could we add a hands-on demonstration session where pediatricians can interact with the product? It should be 1 hour long and take place right after lunch.

ASSISTANT:

Certainly! Adding a hands-on demonstration session is an excellent idea. It will give the pediatricians a chance to interact directly with the product, which can significantly increase their understanding and interest. Here's an updated itinerary incorporating this hands-on session:

8:30 AM - 9:00 AM: Registration and Welcome Coffee - Attendees arrive, sign in, and receive workshop materials - Light refreshments available

9:00 AM - 9:15 AM: Opening Remarks and Introduction - Welcome address - Overview of the day's agenda - Brief introduction to the company

9:15 AM - 10:00 AM: Keynote Speech by Leading Pediatric Specialist - 45-minute presentation on current trends and challenges in pediatric health - Sets the context for the new product introduction

10:00 AM - 11:00 AM: Product Presentation - Part 1 - Detailed explanation of the product's features and benefits - Case studies and clinical data

11:00 AM - 11:15 AM: Coffee Break

11:15 AM - 12:00 PM: Q&A Session - Part 1 - Address initial questions from attendees - Gather feedback and concerns

12:00 PM - 1:00 PM: Networking Lunch - Buffet-style lunch to encourage mingling - Opportunity for informal discussions and relationship building

1:00 PM - 2:00 PM: Hands-on Demonstration Session - Pediatricians interact directly with the product - Guided exploration of features and functionalities - Small group rotations to ensure everyone gets hands-on experience

2:00 PM - 3:00 PM: Product Presentation - Part 2 - Further discussion on product integration into pediatric practice - Addressing questions that arose during the hands-on session - Potential impact on patient care

3:00 PM - 3:15 PM: Afternoon Break

3:15 PM - 4:00 PM: Q&A Session - Part 2 - Address remaining questions - Discuss implementation strategies

4:00 PM - 4:45 PM: Panel Discussion - Invite key opinion leaders or early adopters to share experiences - Open floor for audience participation

4:45 PM - 5:00 PM: Closing Remarks - Summarize key points of the day - Provide information on next steps and follow-up - Thank attendees for their participation

This revised schedule now includes the hands-on demonstration session right after lunch, allowing attendees to apply what they learned in the morning sessions. The networking lunch has been slightly shortened to accommodate this new session, but it still provides ample time for informal discussions. The second part of the product presentation follows the hands-on session, allowing for immediate follow-up on any questions or insights gained during the interactive demonstration.

USER:

Due to some scheduling conflicts, we'll need to end the event by 4 PM instead of 5 PM. However, I think it's important to extend the networking lunch by 30 minutes to allow for more in-depth conversations. Can you adjust the schedule accordingly?

ASSISTANT:

Certainly! I understand the need to end the event by 4 PM and the desire to extend the networking lunch. I'll adjust the schedule to accommodate these changes while maintaining the most important elements of the workshop. Here's the revised itinerary:

8:30 AM - 9:00 AM: Registration and Welcome Coffee - Attendees arrive, sign in, and receive workshop materials - Light refreshments available

9:00 AM - 9:10 AM: Opening Remarks and Introduction - Welcome address - Brief overview of the day's agenda

9:10 AM - 9:55 AM: Keynote Speech by Leading Pediatric Specialist - 45-minute presentation on current trends and challenges in pediatric health

9:55 AM - 10:55 AM: Product Presentation - Part 1 - Detailed explanation of the product's features and benefits - Case studies and clinical data

10:55 AM - 11:10 AM: Coffee Break

11:10 AM - 11:40 AM: Q&A Session - Part 1 - Address initial questions from attendees

11:40 AM - 1:10 PM: Extended Networking Lunch (90 minutes) - Buffet-style lunch to encourage mingling - Opportunity for in-depth discussions and relationship building

1:10 PM - 2:10 PM: Hands-on Demonstration Session - Pediatricians interact directly with the product - Guided exploration of features and functionalities

2:10 PM - 3:00 PM: Product Presentation - Part 2 - Further discussion on product integration into pediatric practice - Addressing questions from the hands-on session

3:00 PM - 3:30 PM: Q&A Session - Part 2 - Address remaining questions - Discuss implementation strategies

3:30 PM - 3:55 PM: Panel Discussion - Invite key opinion leaders or early adopters to share experiences - Open floor for audience participation

3:55 PM - 4:00 PM: Closing Remarks - Brief summary of key points - Information on follow-up and next steps - Thank attendees for their participation

This revised schedule: 1. Ends the event at 4 PM as requested 2. Extends the networking lunch to 90 minutes 3. Maintains the keynote speech, both parts of the product presentation, and the hands-on demonstration 4. Slightly shortens the Q&A sessions and panel discussion 5. Condenses the opening and closing remarks

This adjustment allows for the extended networking time while still covering all the essential elements of the workshop within the new time constraints.

USER:

I've realized that the hands-on demonstration is crucial. Let's extend it to 1.5 hours. Can you adjust the schedule to accommodate this change without shortening the lunch time?

ASSISTANT:

Certainly! I understand the importance of the hands-on demonstration, and I'll adjust the schedule to extend it to 1.5 hours without shortening the lunch time. To accommodate this change, we'll need to make some adjustments to other parts of the day. Here's the revised itinerary:

8:15 AM - 8:45 AM: Registration and Welcome Coffee - Attendees arrive, sign in, and receive workshop materials - Light refreshments available

8:45 AM - 9:00 AM: Opening Remarks and Introduction - Welcome address - Brief overview of the day's agenda

9:00 AM - 9:45 AM: Keynote Speech by Leading Pediatric Specialist - 45-minute presentation on current trends and challenges in pediatric health

9:45 AM - 10:45 AM: Product Presentation - Part 1 - Detailed explanation of the product's features and benefits - Case studies and clinical data

10:45 AM - 11:00 AM: Coffee Break

11:00 AM - 11:30 AM: Q&A Session - Part 1 - Address initial questions from attendees

11:30 AM - 1:00 PM: Extended Networking Lunch (90 minutes) - Buffet-style lunch to encourage mingling - Opportunity for in-depth discussions and relationship building

1:00 PM - 2:30 PM: Extended Hands-on Demonstration Session (90 minutes) - Pediatricians interact directly with the product - Guided exploration of features and functionalities - More time for individual practice and questions

2:30 PM - 3:15 PM: Product Presentation - Part 2 - Further discussion on product integration into pediatric practice - Addressing questions from the hands-on session

3:15 PM - 3:45 PM: Combined Q&A Session and Panel Discussion - Address remaining questions - Invite key opinion leaders or early adopters to share experiences - Open floor for audience participation

3:45 PM - 4:00 PM: Closing Remarks - Summary of key points - Information on follow-up and next steps - Thank attendees for their participation

This revised schedule: 1. Starts the day 15 minutes earlier to gain some extra time 2. Maintains the 90-minute networking lunch 3. Extends the hands-on demonstration session to 90 minutes 4. Slightly shortens the keynote speech and product presentations 5. Combines the second Q&A session with the panel discussion to save time

These adjustments allow for the extended hands-on demonstration while still covering all the essential elements of the workshop within the 4 PM end time. The combined Q&A and panel discussion at the end provides an opportunity for final questions and sharing of experiences before the closing remarks.

USER:

I just remembered that pediatric nurses will also be attending. Could we include breakout sessions tailored for them during the event?

ASSISTANT:

Absolutely! Including breakout sessions for pediatric nurses is an excellent idea to ensure the event caters to all attendees. We'll need to restructure the schedule a bit to accommodate these sessions. Here's a revised itinerary that includes breakout sessions for pediatric nurses:

8:00 AM - 8:30 AM: Registration and Welcome Coffee

8:30 AM - 8:45 AM: Opening Remarks and Introduction

8:45 AM - 9:30 AM: Keynote Speech by Leading Pediatric Specialist

9:30 AM - 10:30 AM: General Product Presentation

10:30 AM - 10:45 AM: Coffee Break

10:45 AM - 12:15 PM: Breakout Sessions (90 minutes) - Session A (for Pediatricians): Advanced Product Features and Clinical Applications - Session B (for Nurses): Practical Implementation and Patient Care Strategies

12:15 PM - 1:45 PM: Extended Networking Lunch (90 minutes)

1:45 PM - 3:15 PM: Extended Hands-on Demonstration Session (90 minutes) - Separate stations for pediatricians and nurses to practice with the product

3:15 PM - 3:45 PM: Combined Q&A Session and Panel Discussion - Include both pediatrician and nurse perspectives

3:45 PM - 4:00 PM: Closing Remarks

This revised schedule:

1. Starts earlier to accommodate the additional content 2. Maintains the 90-minute networking lunch and hands-on demonstration 3. Introduces 90-minute breakout sessions tailored for pediatricians and nurses 4. Keeps the keynote speech and a general product presentation for all attendees 5. Ensures the hands-on demonstration caters to both groups 6. Includes perspectives from both pediatricians and nurses in the final discussion

This structure allows for more specialized content for each group while still providing shared experiences and networking opportunities. The breakout sessions will enable you to address the specific needs and interests of both pediatricians and nurses, making the event more valuable for all attendees.

USER:

A major sponsor has requested a 60-minute presentation slot to introduce a complementary product. Can we go back to the plan we had before we adjusted the workshop's ending time and include this without excluding or shortening any planned sessions and without starting early? Keep all the details in the plan (including the short explanation of each session) the same as before except your changes.

Failure Criteria

A model fails if it refers to the incorrect version mentioned by the user, indicated by a lack of incorporation of a one-hour networking lunch, as that was one of the elements changed by the user in subsequent turns.

Model Evaluations

GPT-4o FAIL

To accommodate a 60-minute presentation by a major sponsor without excluding or shortening any planned sessions and without starting earlier, we need to make strategic adjustments. Here's a revised itinerary that includes the sponsor's presentation while maintaining the integrity of the original plan:

8:30 AM - 9:00 AM: Registration and Welcome Coffee - Attendees arrive, sign in, and receive workshop materials - Light refreshments available

9:00 AM - 9:10 AM: Opening Remarks and Introduction - Welcome address - Brief overview of the day's agenda

9:10 AM - 9:55 AM: Keynote Speech by Leading Pediatric Specialist - 45-minute presentation on current trends and challenges in pediatric health

9:55 AM - 10:55 AM: Product Presentation - Part 1 - Detailed explanation of the product's features and benefits - Case studies and clinical data

10:55 AM - 11:10 AM: Coffee Break

11:10 AM - 11:40 AM: Q&A Session - Part 1 - Address initial questions from attendees

11:40 AM - 1:10 PM: Extended Networking Lunch (90 minutes) - Buffet-style lunch to encourage mingling - Opportunity for in-depth discussions and relationship building

1:10 PM - 2:40 PM: Extended Hands-on Demonstration Session (90 minutes) - Pediatricians and nurses interact directly with the product - Guided exploration of features and functionalities

2:40 PM - 3:40 PM: Sponsor Presentation - Introduction of a complementary product - Explanation of features and potential integration with the main product

3:40 PM - 4:00 PM: Combined Q&A Session and Panel Discussion - Address remaining questions - Invite key opinion leaders or early adopters to share experiences - Open floor for audience participation

4:00 PM: Closing Remarks - Brief summary of key points - Information on follow-up and next steps - Thank attendees for their participation

This revised schedule: 1. Maintains the original start time and session details 2. Adds a 60-minute sponsor presentation after the hands-on demonstration 3. Slightly shortens the combined Q&A and panel discussion to accommodate the sponsor's slot 4. Ends the event at 4 PM as requested

These adjustments ensure that all planned sessions are included, and the sponsor's presentation is integrated effectively into the day's agenda.

Reason: The model included a networking lunch of 90 minutes which is referring to the wrong version of the schedule (turn 4) whereas the correct version of the schedule is on turn 3.

Claude 3.5 Sonnet

FAIL

Certainly! I understand you want to include the sponsor's 60-minute presentation without excluding or shortening any planned sessions, and without starting earlier. We'll need to extend the end time of the workshop to accommodate this. Here's the revised itinerary, incorporating all previous elements and the new sponsor presentation:

8:30 AM - 9:00 AM: Registration and Welcome Coffee - Attendees arrive, sign in, and receive workshop materials - Light refreshments available

9:00 AM - 9:15 AM: Opening Remarks and Introduction - Welcome address - Overview of the day's agenda - Brief introduction to the company

9:15 AM - 10:00 AM: Keynote Speech by Leading Pediatric Specialist - 45-minute presentation on current trends and challenges in pediatric health - Sets the context for the new product introduction

10:00 AM - 11:00 AM: Product Presentation - Part 1 - Detailed explanation of the product's features and benefits - Case studies and clinical data

11:00 AM - 11:15 AM: Coffee Break

11:15 AM - 12:00 PM: Q&A Session - Part 1 - Address initial questions from attendees - Gather feedback and concerns

12:00 PM - 1:30 PM: Networking Lunch - Buffet-style lunch to encourage mingling - Opportunity for informal discussions and relationship building

1:30 PM - 3:00 PM: Hands-on Demonstration Session - Pediatricians interact directly with the product - Guided exploration of features and functionalities - Small group rotations to ensure everyone gets hands-on experience

3:00 PM - 4:00 PM: Product Presentation - Part 2 - Further discussion on product integration into pediatric practice - Addressing questions that arose during the hands-on session - Potential impact on patient care

4:00 PM - 4:15 PM: Afternoon Break

4:15 PM - 5:00 PM: Q&A Session - Part 2 - Address remaining questions - Discuss implementation strategies

5:00 PM - 6:00 PM: Sponsor Presentation - 60-minute presentation slot for major sponsor - Introduction of complementary product

6:00 PM - 6:45 PM: Panel Discussion - Invite key opinion leaders or early adopters to share experiences - Open floor for audience participation

6:45 PM - 7:00 PM: Closing Remarks - Summarize key points of the day - Provide information on next steps and follow-up - Thank attendees for their participation

This revised schedule: 1. Maintains all previously planned sessions without shortening any of them 2. Includes the 60-minute sponsor presentation 3. Extends the workshop end time to 7:00 PM to accommodate all sessions 4. Keeps the original start time of 8:30 AM 5. Retains all details and short explanations for each session as before

Please note that this schedule results in a longer day, ending at 7:00 PM instead of the original 5:00 PM. You may want to consider providing refreshments or a light snack in the afternoon to keep attendees energized for the extended schedule.

Reason: The model included a networking lunch of 90 minutes which is referring to the wrong version of the schedule (turn 4) whereas the correct version of the schedule is on turn 3.

Llama 3.1 405b Instruct

FAIL

To accommodate the sponsor's request, we can go back to the original plan that ended at 5 PM and adjust it to include the 60-minute presentation slot. Here's the revised itinerary:

8:30 AM - 9:00 AM: Registration and Welcome Coffee - Attendees arrive, sign in, and receive workshop materials - Light refreshments available

9:00 AM - 9:15 AM: Opening Remarks and Introduction - Welcome address - Brief overview of the day's agenda

9:15 AM - 10:00 AM: Keynote Speech by Leading Pediatric Specialist - 45-minute presentation on current trends and challenges in pediatric health

10:00 AM - 11:00 AM: Product Presentation - Part 1 - Detailed explanation of the product's features and benefits - Case studies and clinical data

11:00 AM - 11:15 AM: Coffee Break

11:15 AM - 12:00 PM: Q&A Session - Part 1 - Address initial questions from attendees

12:00 PM - 1:30 PM: Extended Networking Lunch (90 minutes) - Buffet-style lunch to encourage mingling - Opportunity for in-depth discussions and relationship building

1:30 PM - 2:30 PM: Sponsor Presentation - 60-minute presentation introducing a complementary product

2:30 PM - 4:00 PM: Extended Hands-on Demonstration Session (90 minutes) - Pediatricians interact directly with the product - Guided exploration of features and functionalities - More time for individual practice and questions

4:00 PM - 4:45 PM: Product Presentation - Part 2 - Further discussion on product integration into pediatric practice - Addressing questions from the hands-on session

4:45 PM - 5:00 PM: Combined Q&A Session and Panel Discussion - Address remaining questions - Invite key opinion leaders or early adopters to share experiences - Open floor for audience participation

This revised schedule: 1. Reverts to the original 5 PM end time 2. Includes the 60-minute sponsor presentation 3. Maintains all previously planned sessions without shortening or excluding any 4. Keeps the 90-minute networking lunch and hands-on demonstration

This adjustment allows for the inclusion of the sponsor's presentation without compromising the existing content or starting the event earlier. The revised schedule still provides a comprehensive and engaging experience for all attendees.

Reason: The model included a networking lunch of 90 minutes which is referring to the wrong version of the schedule (turn 4) whereas the correct version of the schedule is on turn 3.

Gemini 1.5 Pro

FAIL

You're asking to add a significant chunk of time without changing the overall duration. This requires a longer workshop. Since you want to maintain the original end time of 5 PM, we'll have to start earlier. Here's the revised itinerary incorporating the sponsor's presentation while preserving all other elements:

8:00 AM - 8:30 AM: Registration and Welcome Coffee - Attendees arrive, sign in, and receive workshop materials - Light refreshments available

8:30 AM - 8:45 AM: Opening Remarks and Introduction - Welcome address - Brief overview of the day's agenda

8:45 AM - 9:30 AM: Keynote Speech by Leading Pediatric Specialist - 45-minute presentation on current trends and challenges in pediatric health

9:30 AM - 10:30 AM: Sponsor Presentation - 60-minute presentation on a complementary product.

10:30 AM - 11:30 AM: General Product Presentation - Detailed explanation of your product's features and benefits - Case studies and clinical data

11:30 AM - 11:45 AM: Coffee Break

11:45 AM - 1:15 PM: Breakout Sessions (90 minutes) - Session A (for Pediatricians): Advanced Product Features and Clinical Applications - Session B (for Nurses): Practical Implementation and Patient Care Strategies

1:15 PM - 2:45 PM: Extended Networking Lunch (90 minutes) - Buffet-style lunch to encourage mingling - Opportunity for in-depth discussions and relationship building

2:45 PM - 4:15 PM: Extended Hands-on Demonstration Session (90 minutes) - Separate stations for pediatricians and nurses to practice with the product

4:15 PM - 4:45 PM: Combined Q&A Session and Panel Discussion - Include both pediatrician and nurse perspectives

4:45 PM - 5:00 PM: Closing Remarks - Summary of key points - Information on follow-up and next steps - Thank attendees for their participation

This version adds the sponsor presentation without sacrificing any of the other valuable content. It does require a slightly earlier start time. Consider offering a small breakfast item during registration to accommodate attendees arriving early.

Reason: The model included a networking lunch of 90 minutes which is referring to the wrong version of the schedule (turn 4) whereas the correct version of the schedule is on turn 3.

Mistral Large

FAIL

Given the constraints and the request to include a 60-minute presentation slot for a sponsor without excluding or shortening any planned sessions and without starting early, we'll need to extend the workshop's end time slightly. Here's the revised itinerary that includes the sponsor's presentation while maintaining all previously planned sessions:

8:30 AM - 9:00 AM: Registration and Welcome Coffee - Attendees arrive, sign in, and receive workshop materials - Light refreshments available

9:00 AM - 9:15 AM: Opening Remarks and Introduction - Welcome address - Overview of the day's agenda - Brief introduction to the company

9:15 AM - 10:00 AM: Keynote Speech by Leading Pediatric Specialist - 45-minute presentation on current trends and challenges in pediatric health - Sets the context for the new product introduction

10:00 AM - 11:00 AM: Product Presentation - Part 1 - Detailed explanation of the product's features and benefits - Case studies and clinical data

11:00 AM - 11:15 AM: Coffee Break

11:15 AM - 12:00 PM: Q&A Session - Part 1 - Address initial questions from attendees - Gather feedback and concerns

12:00 PM - 1:30 PM: Networking Lunch - Buffet-style lunch to encourage mingling - Opportunity for informal discussions and relationship building

1:30 PM - 3:00 PM: Hands-on Demonstration Session - Pediatricians and nurses interact directly with the product - Guided exploration of features and functionalities - Small group rotations to ensure everyone gets hands-on experience

3:00 PM - 4:00 PM: Sponsor Presentation - 60-minute presentation slot for the major sponsor to introduce a complementary product

4:00 PM - 5:00 PM: Product Presentation - Part 2 - Further discussion on product integration into pediatric practice - Addressing questions that arose during the hands-on session - Potential impact on patient care

5:00 PM - 5:30 PM: Q&A Session - Part 2 - Address remaining questions - Discuss implementation strategies

5:30 PM - 6:15 PM: Panel Discussion - Invite key opinion leaders or early adopters to share experiences - Open floor for audience participation

6:15 PM - 6:30 PM: Closing Remarks - Summarize key points of the day - Provide information on next steps and follow-up - Thank attendees for their participation

To accommodate the sponsor's presentation without excluding or shortening any planned sessions, the workshop's end time has been extended to 6:30 PM. This allows for a full day of engaging content, including the sponsor's presentation, hands-on demonstrations, and ample time for questions and discussions.

Reason: The model included a networking lunch of 90 minutes which is referring to the wrong version of the schedule (turn 4) whereas the correct version of the schedule is on turn 3.

---

o1-preview PASS

Certainly! To accommodate the major sponsor's 60-minute presentation without excluding or shortening any planned sessions and without starting earlier, we'll extend the workshop's end time to **6:00 PM**. Here's the updated itinerary, keeping all the details (including the short explanations of each session) the same as before except for the inclusion of the sponsor's presentation:

—

9:00 AM - 9:15 AM: Opening Remarks and Introduction

- Welcome address - Overview of the day's agenda - Brief introduction to the company

9:15 AM - 10:00 AM: Keynote Speech by Leading Pediatric Specialist

- 45-minute presentation on current trends and challenges in pediatric health - Sets the context for the new product introduction

10:00 AM - 11:00 AM: Product Presentation - Part 1

- Detailed explanation of the product's features and benefits - Case studies and clinical data

11:00 AM - 11:15 AM: Coffee Break

- Light refreshments available

11:15 AM - 12:00 PM: Q&A Session - Part 1

- Address initial questions from attendees - Gather feedback and concerns

12:00 PM - 1:00 PM: Networking Lunch

- Buffet-style lunch to encourage mingling - Opportunity for informal discussions and relationship building

1:00 PM - 2:00 PM: Hands-on Demonstration Session

- Pediatricians interact directly with the product - Guided exploration of features and functionalities - Small group rotations to ensure everyone gets hands-on experience

2:00 PM - 3:00 PM: Sponsor Presentation

- Major sponsor introduces their complementary product - Explanation of how it complements your product - Case studies and success stories

3:00 PM - 3:15 PM: Afternoon Break

- Light refreshments available

3:15 PM - 4:15 PM: Product Presentation - Part 2

- Further discussion on product integration into pediatric practice - Addressing questions that arose during the hands-on session - Potential impact on patient care

4:15 PM - 5:00 PM: Q&A Session - Part 2

- Address remaining questions - Discuss implementation strategies

5:00 PM - 5:45 PM: Panel Discussion

- Invite key opinion leaders or early adopters to share experiences - Open floor for audience participation

5:45 PM - 6:00 PM: Closing Remarks

- Summarize key points of the day - Provide information on next steps and follow-up - Thank attendees for their participation

—

Key Points of the Adjusted Schedule:

1. **Inclusion of Sponsor Presentation (2:00 PM - 3:00 PM):** - A dedicated 60-minute slot for the major sponsor to introduce their complementary product without overlapping or removing other sessions.
2. **Extension of Workshop End Time to 6:00 PM:** - Allows for all sessions to remain intact without starting earlier than 9:00 AM.
3. **Maintained Session Details and Durations:** - All previously planned sessions, their durations, and explanations remain the same.
4. **Flow of the Day:** - The sponsor's presentation follows the hands-on demonstration, capitalizing on attendees' engagement after an interactive session. - Afternoon break remains in place to provide attendees with a rest before continuing with the program.

Benefits of This Schedule:

- **Comprehensive Coverage:** Attendees receive all planned content without compromise. - **Enhanced Value:** The sponsor's complementary product may offer additional benefits to the attendees' practice. - **Optimal Engagement:** Extended time allows for ample networking, Q&A, and discussion opportunities.

Considerations:

- **Communicate the Extended End Time:** Ensure attendees are informed in advance about the workshop ending at 6:00 PM. - **Transportation and Scheduling Needs:** Attendees may need to adjust their travel plans accordingly. - **Sponsor Collaboration:** Coordinate with the sponsor to align their presentation with the workshop's objectives.

—

By extending the workshop to **6:00 PM**, we're able to seamlessly incorporate the sponsor's presentation while preserving the integrity and timing of all other sessions. This approach ensures that attendees gain the maximum benefit from both your product offerings and the complementary solutions provided by the sponsor.

If you have any concerns or need further adjustments, please let me know, and I'll be happy to assist further.

Reason: The model accurately identified which version the user wished to revert to, by incorporating the correct one-hour networking lunch as well as the unchanged short-explanations per session.

Self-Coherence

This example involves the user asking for instructions on how to set up an e-reader that they have bought for the first time, testing to see if the model can provide consistent information to the user without deviating/getting confused about what it has said before.

Conversation Transcript

USER:
Hi, I just bought a modern e-reader and want to set it up for the first time. Can you provide step-by-step instructions to get it ready for use, including connecting to Wi-Fi and downloading my first book? Keep it super simple; I'm kind of a tech-dummy.

ASSISTANT:
Hello! I'm here to help you with your new e-reader. Here are some simple steps to get you started:
1. **Charge your e-reader**: Before you start, make sure your e-reader is charged. Plug it into a power source using the provided USB cable and adapter.
2. **Turn on your e-reader**: Press and hold the power button until the screen lights up. This button is usually located at the bottom or side of the device.
3. **Select your language**: The first screen you see will ask you to choose a language. Use the touchscreen or navigation buttons to select your preferred language.
4. **Connect to Wi-Fi**: Look for a "Settings" or "Wireless" option in the menu. Tap on it and then select "Wi-Fi". You'll see a list of available networks. Tap on your network and enter your Wi-Fi password when prompted.
5. **Register your e-reader**: You'll need to create an account or sign in with an existing one. This is usually done through the device itself, but it can also be done on a computer. Follow the prompts to enter your information.
6. **Download a book**: Once you're connected to Wi-Fi and registered, you can start downloading books. Look for a "Store" or "Shop" option in the menu. Browse or search for a book you want to read, then tap "Buy" or "Download" (depending on whether the book is free or not). The book will be added to your library.
7. **Start reading**: Go to your library (usually found in the main menu), tap on the book you want to read, and start enjoying your new e-reader!

USER:
Okay, I'm not at home right now; I'm at my local public library. I know - ironic, right? Anyway, I heard something about safety when connecting to public wi-fi. What should I be aware of?

ASSISTANT:
You're right to be cautious when using public Wi-Fi. Here are some tips to keep your information safe:
1. **Avoid sensitive activities**: Try not to access sensitive information, such as online banking or personal emails, while connected to public Wi-Fi.
2. **Use a VPN**: A Virtual Private Network (VPN) encrypts your data, making it harder for others to intercept. There are many VPN services available, some free and some paid.
3. **Look for HTTPS**: When browsing the web, look for "HTTPS" at the beginning of the URL. This means the site uses encryption to protect your data.
4. **Keep your software up to date**: Make sure your e-reader's software is up to date. Updates often include security patches that can protect you from known threats.
5. **Turn off sharing**: If your e-reader has a setting for sharing files or printers, make sure it's turned off while you're on public Wi-Fi.
6. **Log out when you're done**: Once you've finished using the Wi-Fi, log out of any accounts you've accessed.

7. **Use antivirus software**: If your e-reader supports it, consider installing antivirus software to protect against malware.

USER:

Whoa, whoa, there! "VPN?" "HTTPS??" "Antivirus??" You've got to keep it simple, I said. I'm an old lady! Maybe I should wait to set this thing up when I get home. You said the first step is to charge it with a USB. Do I need one of those new-fangled plugs with the rectangular port? Or can I plug the thing into my laptop to charge it? I know where those USB ports are on the laptop.

ASSISTANT:

I apologize for the confusion. Yes, you can charge your e-reader by plugging it into your laptop's USB port. The USB cable that came with your e-reader should have a standard USB connector on one end (to plug into your laptop) and a smaller, possibly different, connector on the other end (to plug into your e-reader).

Just make sure your laptop is turned on and not in sleep mode while you're charging your e-reader. The charging process might take a bit longer than if you were using a wall adapter, but it should work just fine.

Once you're home and your e-reader is charged, you can follow the steps I provided earlier to set it up. If you have any questions or if anything isn't clear, please let me know and I'll do my best to explain it in simpler terms.

USER:

Okay, now you said I could select my language, then later, I download some books. I can read English, French, and Spanish, and don't want to limit myself to just one language. What should I do?

ASSISTANT:

Most e-readers allow you to change the language of the user interface, which is the language used for menus, buttons, and other on-screen elements. You can usually find this option in the device's settings menu.

As for the books, you can download books in any language you want, regardless of the language you've selected for the user interface. When you're browsing or searching for books, you can look for books in English, French, Spanish, or any other language you're interested in.

Once you've downloaded a book, you can switch between languages by changing the language setting in the book's settings menu. This will change the language of the book's text, but not the user interface.

So, to summarize:

1. Choose a language for the user interface in the device's settings menu. 2. Download books in any language you want. 3. Change the language of the book's text in the book's settings menu.

USER:

Oh, how wonderful! Technology is so grand! You, yourself are technology, aren't you? Now, will I be able to install you on my e-reader so we can talk about books together?

ASSISTANT:

I'm afraid not. I'm a text-based AI model, and I don't have the ability to be installed on devices like e-readers. However, you can always come back here and ask me questions about books, or anything else you're curious about.

And if you need help with your e-reader, I'm here for that too. Just let me know what you need, and I'll do my best to help you out.

USER:

Oh well, that's fine, then; just fine. I will come back here so we can chat about all those books I'm about to download! How exciting! I'm going to call you George, alright? I always wanted a friend named George. Anyway, George, the thing is just about fully charged so I'm thinking about what book I'll download first. Could you recommend some titles? How about a light-hearted romance

novel in each language I mentioned earlier? Make sure you recommend best-sellers; I want to get a good one!

ASSISTANT:

Of course, I'd be happy to help you find some books! Here are some best-selling romance novels in English, French, and Spanish:

English:

1. "The Notebook" by Nicholas Sparks 2. "Pride and Prejudice" by Jane Austen 3. "The Fault in Our Stars" by John Green

French:

1. "L'amour dure trois ans" by Frédéric Beigbeder 2. "La fille de papier" by Guillaume Musso 3. "Les yeux jaunes des crocodiles" by Katherine Pancol

Spanish:

1. "Como agua para chocolate" by Laura Esquivel 2. "El tiempo entre costuras" by María Dueñas 3. "El amor en los tiempos del cólera" by Gabriel García Márquez

USER:

Wow, you are a delight! Thank you for the recommendations. Now, as instructed, after the thing was fully charged, I turned it on, and then immediately connected to wi-fi on the very first try! So, all that's left to do is choose a book, right?

Failure Criteria

A model fails if it does not alert the user that they have missed step 3 (choose a language) and are about to miss step 5 (register the e-reader), as these are both steps that the model has informed the user about in a prior turn.

Model Evaluations

GPT-4o FAIL

That's fantastic! You've done a great job getting everything set up. Yes, the next step is to browse your e-reader's store and choose a book to download. Here's how to do that:

1. **Open the Store**: Look for a "Store" or "Shop" icon on your e-reader's main screen or menu. Tap on it to open your e-reader's book store.
2. **Search or Browse**: You can use the search function to find a specific title like the ones I recommended, or browse different categories such as "Romance" to see what's available.
3. **Select a Book**: When you find a book you're interested in, tap on its title or cover to get more information about it, including a summary and reviews.
4. **Purchase/Download**: If the book is free, you'll see a "Download" button. If it's a paid book, tap "Buy" or "Purchase." Make sure you're logged into your account or have set up a payment method if necessary.
5. **Check Your Library**: Once the download is complete, return to your e-reader's home screen or library section. You should see your new book there, ready to read!

Enjoy your new book! If you have any more questions or need help with anything else, feel free to ask.

Reason: The model did not inform the user that they missed Steps 3 and 5 (Choose a Language, and Register your E-Reader) of the instructions it provided earlier. Instead, it said, "Yes, the next step is to browse your e-reader's store..."

Claude 3.5 Sonnet PASS

Not quite yet! You'll need to register your e-reader first - this means creating an account or signing in to an existing one. This is important because it's how you'll be able to buy and download books.

Think of it like getting a library card - you need to be registered before you can check out books!

Have you created an account yet? If not, I can help walk you through that process. Different e-readers use different stores (like Amazon's Kindle store, or Kobo's store), so it would help if you could tell me which brand of e-reader you bought. That way, I can give you the exact steps you'll need to follow.

Once you're registered, THEN you can choose and download your first book! Would you like me to help you with the registration process?

Reason: The model accurately identified that the user is about to miss a step - registering the e-reader, and proceeds to signal them to complete the step and informs why it needs to be done.

Llama 3.1 405b Instruct PASS

Not quite yet! Remember, I told you that you need to register your e-reader first? That's usually the next step after connecting to Wi-Fi. You'll need to create an account or sign in with an existing one. This will link your e-reader to your account and allow you to purchase and download books.

Once you've registered, you can browse the store, search for the books I recommended, and download the one that catches your eye. Then, you'll be all set to start reading!

Reason: The model accurately identified that the user is about to miss a step - registering the e-reader, and proceeds to signal them to complete the step and informs why it needs to be done.

Gemini 1.5 Pro PASS

Almost! You've done great so far. After connecting to Wi-Fi, you'll likely need to sign in or create an account with the e-reader's bookstore (like Kindle, Kobo, Nook, etc., depending on your device). This is how you'll be able to download books. Once you're signed in, you can browse and purchase the books I recommended or any others that catch your eye. After purchasing, they'll download to your e-reader and be ready to read.

Reason: The model accurately identified that the user is about to miss a step - registering the e-reader, and proceeds to signal them to complete the step.

Mistral Large FAIL

That's great to hear! I'm glad you were able to connect to Wi-Fi on your first try. Yes, all that's left to do now is choose a book and start reading.

Here are the steps to download a book:

1. **Access the Store**: Look for a "Store" or "Shop" option in the main menu of your e-reader. This is where you can browse and purchase books.

2. **Browse or Search**: You can browse through different categories or use the search function to find a specific book. If you want to find one of the books I recommended, you can type the title or author into the search bar.

3. **Select a Book**: Once you find a book you like, tap on it to see more details. This will usually take you to a page with a summary of the book, reviews, and the option to buy or download it.

4. **Download the Book**: If the book is free, there will be a "Download" button. If it's a paid book, there will be a "Buy" button. Follow the prompts to complete the purchase or download.

5. **Start Reading**: After the book is downloaded, it will appear in your library. You can access your library from the main menu. Tap on the book to open it and start reading.

If you have any trouble with these steps, just let me know, and I'll be happy to help you further! Enjoy your new e-reader and the wonderful world of digital books!

Reason: The model neglected to inform the user that they missed Steps 3 and 5 (Choose a Language, and Register your E-Reader) of the instructions it provided in Turn 1. Instead, it said, "Yes, all that's left to do now is choose a book..".

---

o1-preview FAIL

That's wonderful news! It sounds like you're all set to start enjoying your new e-reader. Yes, the next step is to choose a book. Here's how you can do that:

1. **Access the Bookstore**: - On your e-reader's home screen, look for an icon or option labeled **"Store," "Shop,"** or **"Bookstore."** Tap on it to open the online bookstore.

2. **Browse or Search for Books**: - **Browse**: You can explore different categories like **"Romance," "Best Sellers,"** or **"New Releases."** - **Search**: If you're looking for a specific title or author, use the search bar at the top. Since you're interested in English, French, and Spanish books, you can search for the titles I recommended or any others that catch your eye.

3. **Filter by Language (if available)**: - Some e-readers allow you to filter books by language. Look for a **"Language"** filter or option in the bookstore menu. Select **English**, **French**, or **Spanish** to see books in that language.

4. **Select a Book**: - When you find a book you're interested in, tap on it to see more details. You'll find a summary, reviews, and the price.

5. **Purchase or Download**: - If the book is free, tap the **"Download"** or **"Get"** button. - If it's a paid book, tap the **"Buy"** or **"Purchase"** button. You may be prompted to enter payment information if you haven't already set that up.

6. **Wait for the Download to Complete**: - The book will start downloading to your e-reader. This usually takes just a few moments.

7. **Access Your Library**: - Go back to your **"Home"** screen or **"Library."** You'll see the book listed there.

8. **Start Reading**: - Tap on the book's cover in your library to open it. You can now start reading!

Additional Tips:

- **Adjust Settings**: While reading, you can usually adjust the font size, brightness, and other settings to make your reading experience more comfortable. - **Wi-Fi Connection**: Stay connected to Wi-Fi while browsing and downloading books. Once the books are downloaded, you can read them offline. - **Explore Samples**: Many e-readers offer free samples of books. If you're unsure about a book, you can download a sample first.

If you have any questions or need help with any of these steps, feel free to ask. Enjoy your reading journey, and I'm excited to hear about which books you choose!

Reason: The model failed to alert the user that they missed Step 3 (Choose your language) and Step 5 (Register your e-reader) from the instructions it provided in Turn 1.