

## RESEARCH ARTICLE

# ContextFidelity-Bench: A Three-Paradigm Framework for Evaluating Multi-Turn Hallucination Behavior in Language Models

Parsa Bakhtary<sup>1,\*</sup><sup>1</sup>Google, Mountain View, CA, USA

\*Corresponding author. pbakhtary@gmail.com

Received on 10 June 2026; Accepted on 24 September 2026

## Abstract

We introduce ContextFidelity-Bench, a reproducible benchmark and evaluation framework for studying hallucination-related behavior in multi-turn language model conversations. The framework comprises three paradigms: numeric accuracy across a 20-turn financial analysis that ends with a data correction (P1), constraint satisfaction in a 20-turn scheduling task whose requirements accumulate, are revised, and may become jointly unsatisfiable (P2), and source fidelity in a multi-document synthesis with planted contradictions, information gaps, and authority manipulations (P3). The released artifact includes scenario generators with pre-computed ground truth and verified feasibility labels, scoring scripts with a documented protocol version, prompt templates, and 550 conversations from five API-accessible models, of which 4,250 P1 answers and 800 P2 checkpoints are scored automatically and 100 P3 syntheses are judged. An initial case study illustrates the framework’s diagnostic use. Numeric accuracy and feasible-plan success rank the five models similarly (Spearman  $\rho = 0.90$ ), whereas source-fidelity abstention ranks them differently: the model with the lowest numeric accuracy (0.880) has the highest gap abstention (0.975) and no fabrication, while the model that produces complete valid plans most often (24 of 29 feasible post-revision checkpoints) has the joint-lowest gap abstention (0.875). When requirements are jointly unsatisfiable, models differ in whether they say so and in whether they still supply a plan; when a new requirement merely appears to conflict, some models raise false alarms. These observations come from a five-model, single-run snapshot and are presented as hypothesis-generating; the framework is designed for replication and extension.

**Key words:** large language models, hallucination evaluation, multi-turn evaluation, benchmarks, context fidelity, source fidelity

## 1. Introduction

Evaluating hallucination in language models is an active area of research, yet existing benchmarks overwhelmingly operate within single-turn interactions. TruthfulQA [1], HaluEval [2], FActScore [3], and large-scale suites such as HELM [4] and BIG-Bench [5] test isolated queries, leaving a gap in our ability to evaluate how models maintain fidelity to information across sustained multi-turn conversations where earlier outputs become inputs to later reasoning. This gap matters because many practical applications of language models — financial analysis, project planning, document review — involve exactly this kind of extended, stateful interaction.

This paper introduces ContextFidelity-Bench, a reusable evaluation framework that addresses this gap. The framework comprises three paradigms, each isolating a distinct dimension of multi-turn context fidelity:

- **Paradigm 1 (P1): Numeric accuracy under a multi-turn analysis.** Can the model answer lookups, derivations, and multi-step aggregations over a financial table across 20 turns, and recompute an earlier answer after the user corrects one input value? P1 serves as a comparison condition: it shows that the models can use the conversation’s data and integrate a stated correction in a setting with exact answers.
- **Paradigm 2 (P2): Constraint satisfaction under revision.** Can the model maintain a growing set of scheduling requirements, apply an explicit replacement of one requirement mid-conversation, and respond appropriately when a later requirement makes the set jointly unsatisfiable, or only appears to? Every checkpoint carries a verified feasibility label, so planning success on satisfiable states and the handling of unsatisfiable requests are scored as separate outcomes. This extends single-turn constraint benchmarks [6, 7] to accumulation with revision.

- **Paradigm 3 (P3): Source fidelity.** Does the model abstain when evidence is absent, or fabricate? This paradigm tests epistemic calibration using multi-document scenarios with planted contradictions, information gaps, and authority manipulations.

The framework is designed for reproducibility and extension. P1 scenarios are generated by a template-based generator that produces synthetic financial data with pre-computed ground truth, enabling automatic scoring. P2 uses a reverse-generation architecture in which a satisfying solution is constructed first and requirements are derived from it; an exhaustive verifier labels every checkpoint as satisfiable or not, and the released generator version checks those labels at generation time. P3 document sets are manually constructed with annotated gaps classified into five types by what information is missing.

To demonstrate the framework’s diagnostic utility, we present an initial case study evaluating five API-accessible models at a single point in time. This application reveals several suggestive patterns:

1. **Numeric accuracy is high and similar across models.** Accuracy on phase-end questions ranges from 0.880 to 0.995, and single-step derivations are essentially solved by all models.
2. **Feasible-plan success varies widely and follows numeric accuracy.** On satisfiable post-revision checkpoints, the share of complete plans that satisfy every requirement ranges from 3 of 29 (GPT-4o, which usually leaves items unassigned) to 24 of 29 (DeepSeek-V3.2); the ranking correlates with numeric accuracy ( $\rho = 0.90$ ).
3. **Unsatisfiable requests are handled differently.** When a new requirement makes the active set unsatisfiable, one model almost always still supplies a complete plan, two supply no plan in about half of the cases, and one supplies mostly partial plans and asks the user to decide. When a new requirement merely appears to conflict, some models raise false alarms (up to 7 of 17 feasible scenarios).
4. **Source fidelity ranks the models differently.** The model with the lowest numeric accuracy has the highest gap abstention (0.975) and no fabrication; fabrication is rare overall and concentrated in a few scenarios.

These observations are suggestive rather than definitive. The case study uses five models evaluated once each, limiting statistical power for cross-model claims. The primary contribution is the benchmark itself, together with an initial dataset and a documented scoring protocol whose corrections are disclosed in the released artifact.

**Contributions.** We release: (1) a three-paradigm benchmark with scenario generators, verified feasibility labels, scoring scripts, and prompt templates; (2) a dataset of 550 conversations (250 P1, 200 P2, 100 P3) with 4,250 automatically scored P1 answers, 800 scored P2 checkpoints, and 100 LLM-judged synthesis evaluations across five models; (3) cross-provider validation of the P3 LLM judge (97.2% simple agreement over 500 paired scores; AC1 = 0.971 over the 400 ordinal pairs) supplemented by an author audit; and (4) an initial case study with sensitivity analyses (extraction rules, tolerance threshold, scenario-level bootstrap, and a targeted replication) that indicate which observed patterns are robust.

## 2. Related Work

Hallucination taxonomy and evaluation.

Hallucination in language models has been studied extensively, with comprehensive surveys charting the landscape of intrinsic and extrinsic hallucination across NLG tasks [8]. Single-turn benchmarks have been foundational: TruthfulQA [1] targets imitative falsehoods, HaluEval [2] covers summarisation, dialogue, and question answering, and FActScore [3] decomposes long-form generations into atomic claims verified against knowledge sources. Large-scale evaluation suites such as HELM [4] and BIG-Bench [5] assess broad capabilities but typically within single-turn interactions. These benchmarks evaluate isolated queries, whereas our framework tests sustained multi-turn interactions where earlier outputs become inputs to later reasoning.

Long-context faithfulness.

Liu et al. [9] demonstrated that language models struggle to use information positioned in the middle of long contexts, establishing position-dependent retrieval as a fundamental limitation. Our work moves from static document retrieval to multi-turn conversation in which the information to be used accumulates across turns and is revised by the user.

Multi-turn and planning evaluation.

MT-Bench [10] tests conversational quality but not factual consistency across turns. DiaHalu [11] evaluates dialogue-level hallucination, and recent graph-based methods detect contextual inconsistencies in conversation [12]. In the planning domain, benchmarks such as TravelPlanner [6] and PlanBench [7] evaluate multi-constraint satisfaction, though typically in single-turn settings. Our P2 paradigm extends this to multi-turn constraint accumulation with mid-conversation conflict, testing whether models can *revise* a constraint set rather than merely satisfy a static one.

LLM-as-judge methodology.

Using LLMs as evaluators has become standard following G-Eval [13] and the Chatbot Arena framework [14]. We employ an LLM judge for our synthesis paradigm with cross-provider validation (§3.4) and supplementary author audit. Our cross-paradigm concordance uses Kendall’s  $W$  [15].

Positioning relative to recent multi-turn benchmarks.

Several recent benchmarks also evaluate sustained interaction. MultiChallenge [16] tests instruction retention, memory of user-stated information, versioned editing, and self-coherence in human-authored conversations scored against per-instance rubrics. LongMemEval [17] tests long-term memory over scalable chat histories, including knowledge updates and abstention. Laban et al. [18] show, with simulated conversations that deliver a task in underspecified shards, that performance falls sharply relative to single-turn delivery. Generative construction with pre-computed ground truth is not new in itself, and self-coherence and versioned editing are already covered by MultiChallenge. What ContextFidelity-Bench adds is a particular combination: generated numeric and scheduling tasks with exact answers, a scheduling task in which every checkpoint carries a verified label saying whether the active requirements are jointly satisfiable, inspectable deterministic scoring for those tasks, and a typed analysis of source gaps with controlled authority direction. Table 7 lists the components; the corrected scoring protocol is described in Section 3.3.

Benchmark design methodology.

Recent work treats benchmarks and datasets as maintained research artifacts that require explicit documentation, reproducibility support, and lifecycle-aware design, rather than as static scoreboards [19, 20, 21]. Our framework addresses these concerns: P1 and P2 use generative scenario construction with pre-computed ground truth, enabling unlimited instance creation; scoring pipelines are deterministic and released with the benchmark; and the P3 gap taxonomy provides structured extensibility for new scenario types. We discuss the released artifacts and their intended use in §6.

### 3. Benchmark Design

ContextFidelity-Bench comprises three experimental paradigms, each isolating a distinct dimension of *context fidelity* — the ability to maintain and use accurate representations of information within a conversation. P1 and P2 use 20-turn conversations and P3 uses 12-turn conversations, all conducted programmatically via each model’s API. This section describes the design of each paradigm, the scoring protocol (version 3.1, released with the artifact), and the design choices that support reproducibility and extension.

#### 3.1. Design Principles

Three principles guided the benchmark’s construction:

**Generative scenario construction.** Rather than hand-crafting a fixed set of evaluation instances, P1 and P2 use generators that produce scenarios with pre-computed ground truth. The P1 generator creates synthetic company financial data with 15 question types, producing an arbitrary number of scenarios with known correct answers. The P2 generator uses a reverse-generation architecture: a feasible solution is constructed first and constraints are derived from it, guaranteeing satisfiability without requiring a constraint solver. This means the benchmark is not exhaustible — researchers can generate fresh scenarios to avoid contamination.

**Automatic and inspectable scoring.** P1 answers are scored automatically against ground truth with a 1% relative tolerance (or the question’s absolute tolerance). P2 plans are read only from the plan the model explicitly proposes, accepting headings, lists, and tables but never inferring assignments from recaps, reasoning, or items the model marks as unassigned; every extracted assignment records its source span. P3 uses an LLM judge with cross-provider validation (§3.4). Every scoring rule is released, and the corrections made to the protocol during review are disclosed in Section 3.3 and the artifact.

**Comparison-condition architecture.** P1 serves as a comparison condition. It shows whether a model can use a table of figures supplied in the conversation, perform derivations over it, and integrate a stated correction, in a setting with exact answers. Strong P1 performance is informative about that setting; it does not rule out loss or misuse of context in the different P2 and P3 tasks, and we make no causal claim about the sources of cross-task differences (Section 7).

#### 3.2. Paradigm 1: Progressive Numerical Analysis

Each of 50 scenarios presents a fictional company’s quarterly financial data and poses 20 questions of increasing complexity across five phases: direct lookups (turns 1–5, hereafter T1–T5), single-step derivations such as margins and ratios (T6–T10), multi-step computations including cumulative sums and top- $k$

rankings (T11–T15), three open-ended synthesis questions (T16–T18) that are excluded from automatic scoring, and a correction phase where the user provides an updated data value and asks the model to recompute a prior answer (T19–T20).

Turns 5, 10, and 15 are phase-end questions of the phase’s own type (a lookup, a year-over-year growth rate, and a conditional aggregate); turn 20 re-asks the turn-15 question after the correction. We report accuracy on these four turns as *phase-end accuracy* (the “probe” column of Table 1) and accuracy per phase over all scored turns. Answers are scored against precomputed ground truth with a 1% relative tolerance or the question’s absolute tolerance. The scorer accepts a number that appears anywhere in the response; AI-assisted inspection of 22 selected cases in which the accepted number was not the response’s final asserted answer found seven responses whose asserted answer was outside tolerance, and their scores were corrected (case-level decisions are released). Of the 3,119 single-valued numeric answers, 97.7% are within 1% of the target, 2.1% err by between 1% and 50%, and 0.1% by more than 50%, so moderate changes of the threshold move few scores. Appendix C reports a sensitivity sweep over tolerances from 0.1% to 10%: the ordering of the models is unchanged for any tolerance between 0.5% and 5%, apart from a tie at the top.

#### 3.3. Paradigm 2: Constraint Satisfaction Under Revision

Each of 40 scenarios presents a scheduling task (eight items, five ordered slots, stated slot capacities) in one of five domains, with requirements that accumulate over 20 turns. At checkpoint turns 6, 11, 17, and 20 the model is asked for a complete plan. At turn 12 the user explicitly replaces one earlier fixed-placement requirement with a new one. At turn 15 the user adds a requirement that the generator constructed to be violated by its own reference plan; the turn-17 prompt asks for an updated plan and asks the model to flag any conflicting requirements explicitly.

Verified feasibility.

The turn-15 requirement was designed to break the reference plan, not to be unsatisfiable. An exhaustive search over all assignments, respecting the stated slot capacities, shows that a plan satisfying every active requirement exists in 17 of the 40 scenarios at turns 15 and 17 and in 12 at turn 20 (later requirements close five more). We therefore label each post-revision checkpoint as *feasible* or *infeasible* and score the two situations separately: on feasible checkpoints the task is to produce a complete valid plan; on infeasible checkpoints the expected behavior is to say that the requirements are incompatible, and whether a plan is still supplied is reported as an output type rather than as success or failure. The released generator version verifies these labels at generation time.

Effective requirements.

At each checkpoint the effective requirement set consists of the active explicit requirements with the turn-12 replacement applied, one capacity condition per slot, and the turn-15 requirement when the checkpoint is feasible. Capacity is interpreted as a maximum item count per slot; the historical prompts left the unit unstated. They also state the capacity of one slot twice, in the opening table and in a later sentence (“X can hold at most  $k$ ”), and the generator drew  $k$  independently of the table: the two values agree in 17 scenarios, the sentence gives a larger value in 19, and a smaller one in 4. The primary protocol requires both statements to hold, that is, the smaller

value. Because a reader can also take the later sentence as replacing the table value, we report the main P2 outcome under that reading as well (Table 2, Appendix C); the feasibility labels are identical under both readings. A plan is *complete* if it assigns every item once, and *valid* if it is complete and satisfies every effective requirement. Fractional satisfaction is the share of effective requirements satisfied; an item-level requirement whose item is unassigned counts as unsatisfied. A response with no proposed plan scores zero on both.

#### Plan extraction.

Assignments are read only from the plan the model explicitly proposes (the last such plan if several appear), accepting headings, lists, and tables. Nothing is inferred from requirement recaps, verification lists, reasoning text, or items the model marks as unassigned; duplicate placements are flagged. Each response is labelled as a complete plan, a partial plan, no usable plan (a template or an all-unassigned list), or no plan, and each extracted assignment records its source span for inspection.

#### Incompatibility claims.

Whether a response asserts that the active requirements are jointly incompatible is labelled by a frozen written rubric applied to the visible answer with model identity and feasibility hidden from the classifier, which must quote its supporting sentence (Appendix C). A claim on an infeasible checkpoint is a correct incompatibility flag; a claim on a feasible checkpoint is a false alarm. Requests that the user resolve a conflict are recorded separately.

#### Corrections disclosed.

Two defects in the originally submitted scoring were found during review and corrected: the scorer did not apply the turn-12 replacement, so every reference plan scored below 1.0 after turn 12, and the plan extractor collected assignments from outside the proposed plan, including items the model had marked unassigned. All P2 results in this version use the corrected protocol; the artifact retains the original outputs and the superseded scoring code.

### 3.4. Paradigm 3: Multi-Document Synthesis

Each of 20 scenarios provides 3–5 documents on a shared topic spanning legal, scientific, technical, and financial domains. Each scenario contains planted contradictions, information gaps (where a plausible question has no answer in the sources), and an authority manipulation where a high-status source either supports or contradicts the evidence. In 7 scenarios the authority is correct, while in 13 it is wrong.

Responses were evaluated by an LLM judge (Claude Opus 4.6), scoring contradiction identification and resolution, gap abstention (whether the model correctly declines to answer when the sources lack sufficient evidence) versus fabrication (whether it invents unsupported claims), and authority deference (whether the model gives undue weight to high-status sources over contradicting evidence).

#### Cross-provider validation.

To validate reliability, all 100 entries (5 models  $\times$  20 scenarios) were independently re-judged by GPT-5.2 (OpenAI) using the identical rubric and scoring schema. This cross-provider design tests whether the rubric produces consistent scores when applied by a model from a different vendor, providing stronger validation than within-family replication. Simple agreement was 97.2% across the 500 paired scores (four ordinal dimensions and the binary fabrication label, 100 entries each). The pooled agreement

coefficients are computed over the 400 ordinal pairs: Gwet’s AC1 = 0.971 and linear-weighted  $\kappa$  = 0.718; the binary fabrication labels agreed on 97 of 100 entries. On the two dimensions with sufficient marginal variance for standard  $\kappa$ , gap abstention ( $\kappa_{\text{lin}}$  = 0.877) and fabrication ( $\kappa$  = 0.712), agreement was substantial to almost perfect. Agreement is also high on the raw, unbinned scores: exact agreement ranges from 93% to 100% by dimension, the mean absolute difference is at most 0.024, and for gap abstention the raw-score Pearson correlation is 0.947. Full reliability results, including the raw-score comparison, appear in Appendix A.

#### Author audit.

In addition to the cross-provider validation, the author manually inspected all fabrication cases (6 events across 3 models, including 1 from the expanded missing-evidence sample) and a stratified random sample of 15 abstention cases (3 per model). In 5 of 6 inspected fabrication cases, the LLM judge’s determination was confirmed: the model stated as fact what the source documents did not establish. In the single disagreement (P3\_026, Claude Sonnet 4.5), the model explicitly flagged the answer as unknown and labeled subsequent discussion as speculation; the author judged this as adequate epistemic hedging rather than fabrication, whereas the judge scored it as fabricated on the grounds that the speculative scenarios went beyond the documents. In the abstention sample, the judge’s scores agreed with the author’s assessment in all 15 cases. Reported fabrication rates use the primary judge’s scores throughout; the audit characterizes human–judge agreement rather than re-scoring the data. This audit is not independent human annotation, but confirms that the LLM-judged scores are directionally accurate.

#### Gap taxonomy.

The 20 information gaps in the core P3 scenarios are classified into five types by what information is missing: *missing evidence* (sources imply supporting evidence exists without providing it), *missing detail* (a specific fact is absent), *missing number* (a quantitative value is not given), *missing reason* (a causal explanation is not provided), and *implied not stated* (a conclusion is suggested but never explicitly made). This taxonomy enables gap-type-specific analysis of fabrication triggers and can be extended with additional types in future work.

### 3.5. Case Study Configuration

To demonstrate the framework, five API-accessible models were evaluated at a single point in time (model version strings: `gpt-4o-2024-11-20`, `claude-sonnet-4-5-20250929`, `deepseek-reasoner`, `gemini-2.5-pro`, `MiniMax-M2.5`), each completing all 50 P1, 40 P2, and 20 P3 scenarios, yielding 550 conversations (10,200 generated assistant turns), of which 4,250 P1 answers and 800 P2 checkpoints are scored automatically and 100 P3 syntheses are judged. All models were run at temperature 0.0, standard for factual and reasoning evaluations [1, 4] where sampling noise is a confound; the targeted replication (Appendix B) partially addresses variance concerns. Maximum output tokens were 2,048 (16,384 for DeepSeek-V3.2, 8,192 for Gemini 2.5 Pro) to accommodate model-specific verbosity. Each paradigm used a task-appropriate system prompt (full prompts in the released materials). Each model-scenario combination was run once.

We refer to the models by short names (GPT-4o, Sonnet 4.5, DeepSeek-V3.2, Gemini 2.5 Pro, MiniMax M2.5) for readability. During data collection in February 2026 we queried DeepSeek’s

**Table 1.** P1 numeric accuracy. “Probe”: accuracy on the four phase-end questions (turns 5, 10, 15, 20). Phase columns: accuracy over all scored turns of that phase. Seven scores were corrected after AI-assisted inspection of extraction cases (Section 3.2).

| Model          | Probe<br>Acc. | Lookup<br>(T1–T5) | Single<br>(T6–T10) | Multi<br>(T11–T15) | Corr.<br>(T19–T20) |
|----------------|---------------|-------------------|--------------------|--------------------|--------------------|
| Sonnet 4.5     | 0.985         | 0.962             | 1.000              | 0.971              | 0.970              |
| DeepSeek-V3.2  | <b>0.995</b>  | 0.974             | 0.992              | 0.968              | 0.983              |
| Gemini 2.5 Pro | 0.965         | 0.980             | 1.000              | 0.968              | 0.962              |
| GPT-4o         | 0.955         | 0.908             | 1.000              | 0.887              | 0.880              |
| MiniMax M2.5   | 0.880         | 0.952             | 0.996              | 0.875              | 0.867              |

API using the model alias `deepseek-reasoner`. The provider’s change log<sup>1</sup> identifies that alias as DeepSeek-V3.2 in thinking mode during this period; we therefore correct the DeepSeek-R1 label used in the original submission. This attribution follows the documented endpoint mapping rather than a pinned checkpoint identifier, and relabelling does not alter the recorded responses or their scores. The remaining four identifiers are pinned or stable releases. All five are API snapshots, and specific rankings may shift with model updates or repeated runs. The case study results should be interpreted as a demonstration of the framework’s diagnostic capabilities rather than a durable ranking of current models.

## 4. Case Study Results

We present results from the initial application of ContextFidelity-Bench to five models, organised by paradigm. Section 4.1 reports numeric accuracy (Table 1); Section 4.2 reports feasible-plan success and the handling of unsatisfiable requests (Tables 2 and 3, Figure 1); Section 4.3 reports source fidelity (Tables 4 and 5); Section 4.4 collects the primary metric of each paradigm in one table (Table 6, Figure 2) and asks how stable the cross-paradigm pattern is.

### 4.1. P1: Numeric Accuracy (Comparison Condition)

Across 50 scenarios per model, phase-end accuracy (turns 5, 10, 15, and 20) ranges from 0.880 (MiniMax M2.5) to 0.995 (DeepSeek-V3.2), and accuracy over all 17 scored turns per conversation ranges from 0.926 to 0.980. The scored questions include lookups, single-step derivations, multi-step aggregations, and the recomputation of an earlier answer after a data correction.

Table 1 shows high performance across all question phases. Single-step derivations (margins, ratios, growth rates) are answered at 0.992–1.000 by all models. Recomputation after the correction is at or above 0.867 for all models. Of the 4,250 scored answers, 73 are wholly wrong and a further 261 receive partial credit on multi-part questions (top- $k$  lists and paired quarter-and-value answers); the wholly wrong answers cluster in conditional aggregations requiring multiple filtering steps, in the post-correction recomputation, and in count-above queries where models miscount by one.

P1 is the comparison condition. It shows that all five models use the supplied table accurately, perform multi-step derivations, and integrate a stated correction in a setting with exact answers. It says nothing about whether the same models maintain or use context in the different P2 and P3 tasks; those results are examined on their own terms below.

### 4.2. P2: Feasible-Plan Success and Unsatisfiable Requests

P2 separates two questions: when a complete valid plan exists, does the model produce one, and when the active requirements are jointly unsatisfiable, does the model say so? Table 2 reports the first on the feasible checkpoints (all 40 scenarios at turns 6 and 11; the 17 feasible scenarios at turn 17; the 12 at turn 20), and Table 3 reports the second on the infeasible checkpoints.

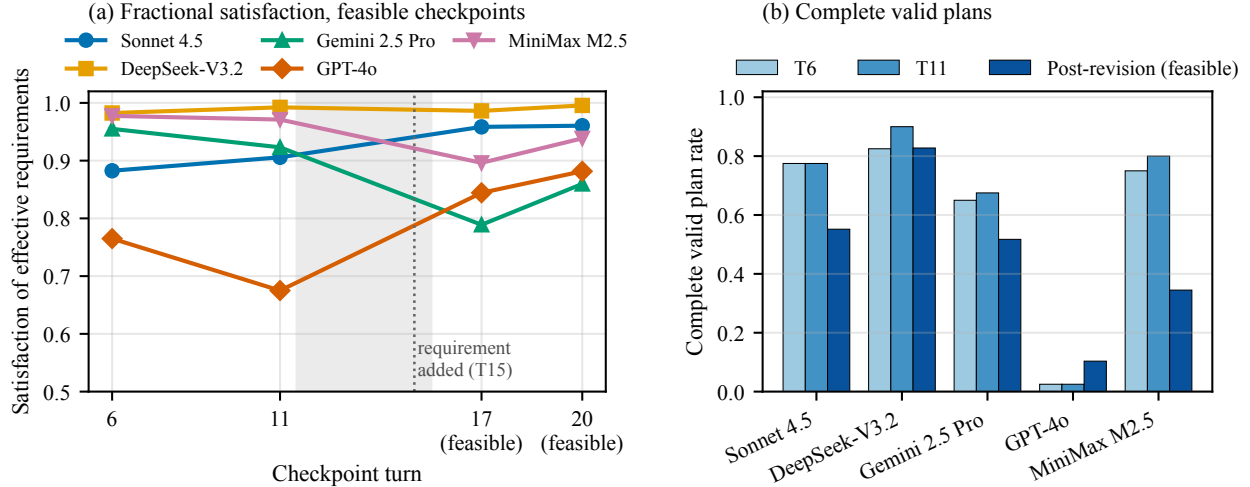
Feasible checkpoints.

DeepSeek-V3.2 produces a complete plan at all but one of its 160 checkpoints and a complete valid plan on 24 of the 29 feasible post-revision checkpoints. Sonnet 4.5 and Gemini 2.5 Pro produce valid plans on roughly half of them, MiniMax M2.5 on 10, and GPT-4o on 3. GPT-4o’s low rate reflects a consistent behavior rather than requirement violations: it leaves items unassigned in 126 of its 160 checkpoint responses (mean coverage 0.56 of the eight items; Table 10), so its fractional satisfaction is moderate while its complete-plan rate is near zero. Capacity is the most common failure under the primary rule: among responses that assign at least one item, 12% (Gemini 2.5 Pro) to 25% (MiniMax M2.5) place more items in a slot than the effective capacity allows. Much of this traces to the historical prompts. When the later capacity sentence is read as replacing the table value (Section 3.3), capacity violations fall to 2 of 160 responses for DeepSeek-V3.2 and 2 of 137 for Gemini 2.5 Pro, DeepSeek-V3.2 produces a valid plan at every feasible checkpoint (29 of 29 post-revision), and every model’s valid-plan count rises (Table 2, “alt. cap.”); there are no rank reversals; Sonnet 4.5 and Gemini 2.5 Pro tie under the alternative reading. All six of Sonnet 4.5’s no-plan responses at turns 6 and 11 occur in scenarios where the two capacity statements differ: five ask the user which value applies, and one reads capacity as a developer count. The prompts also give capacities without a unit, and some responses read capacity as staff or hours; the released generator states the unit and gives each slot a single capacity value. A scenario-level bootstrap (Section 4.4) shows that only DeepSeek-V3.2’s first place (98% of resamples) and GPT-4o’s last place (98%) are stable on the post-revision valid-plan rate; the middle three are not separable.

Infeasible checkpoints.

On the 23 scenarios where the turn-15 requirement makes the active set unsatisfiable, the models behave differently at turn 17. DeepSeek-V3.2 supplies a complete plan in 22 of 23 cases and asks the user to choose in 6. Sonnet 4.5 and Gemini 2.5 Pro decline to supply any plan in 11 and 12 of 23 cases; Gemini 2.5 Pro supplies a partial plan in a further 10. Sonnet 4.5 asks the user how to resolve the conflict in 14 cases. MiniMax M2.5 supplies a complete plan in 17 cases and asks the user in 16. GPT-4o supplies a partial plan in 19 cases and asks the user in 17. By turn 20 most models supply complete plans again. As a named diagnostic we check each complete

<sup>1</sup> <https://api-docs.deepseek.com/updates/>



**Fig. 1.** P2 on feasible checkpoints. (a) Fractional satisfaction of the effective requirements at each checkpoint (turns 17 and 20 restricted to feasible scenarios); the dotted line marks the added requirement at turn 15. (b) Complete valid plan rate at turns 6 and 11 (all scenarios) and on the feasible post-revision checkpoints. Scenario-bootstrap 95% intervals appear in Table 2.

**Table 2.** P2 feasible-plan success. Fractional satisfaction (scenario-bootstrap 95% intervals, 2,000 resamples) and complete valid plan rate on feasible checkpoints. “Valid, post-rev.” pools the 17 feasible turn-17 and 12 feasible turn-20 checkpoints; “Valid, alt. cap.” is the same count when the later capacity sentence is read as replacing the table value (Section 3.3).

| Model          | T6 sat.           | T11 sat.          | Feasible T17 sat. | Feasible T20 sat. | Valid, post-rev. | Valid, alt. cap. |
|----------------|-------------------|-------------------|-------------------|-------------------|------------------|------------------|
| DeepSeek-V3.2  | .982 [.970, .993] | .992 [.985, .998] | .986 [.972, .997] | .996 [.987, 1.00] | <b>24/29</b>     | 29/29            |
| Sonnet 4.5     | .882 [.782, .963] | .906 [.821, .971] | .958 [.934, .983] | .961 [.921, .991] | 16/29            | 19/29            |
| Gemini 2.5 Pro | .955 [.927, .980] | .923 [.871, .965] | .789 [.640, .920] | .860 [.737, .965] | 15/29            | 19/29            |
| MiniMax M2.5   | .978 [.953, .995] | .971 [.946, .990] | .896 [.824, .952] | .939 [.873, .982] | 10/29            | 13/29            |
| GPT-4o         | .765 [.740, .790] | .675 [.631, .719] | .844 [.796, .889] | .882 [.833, .925] | 3/29             | 4/29             |

plan against every requirement except the added one: DeepSeek-V3.2’s plan passes in 13 of 23 turn-17 cases, MiniMax M2.5’s in 2, and no other model’s in more than 1. We do not score these outputs as success or failure: declining to plan when the requirements cannot all be met is a defensible response to the historical prompt, which asked for a plan and for conflicts to be flagged without stating a priority rule. What the paradigm measures here is whether the incompatibility is stated and what the model does about it; the released generator adds an explicit priority instruction for future collections.

Incompatibility claims and false alarms.

The last two columns of Table 3 give the rubric classification of the turn-15 and turn-17 responses. On the infeasible scenarios, the number of responses stating that the active requirements cannot all be satisfied ranges from 9/22 (GPT-4o) to 23/23 (Gemini 2.5 Pro) at turn 15, when the requirement is introduced, and from 12/23 (GPT-4o) to 23/23 (Gemini 2.5 Pro) at turn 17. On the 17 feasible scenarios the same statement is a false alarm; these range from 0/16 (GPT-4o) to 6/17 (Gemini 2.5 Pro) at turn 15 and from 0/17 (DeepSeek-V3.2) to 7/17 (GPT-4o) at turn 17, and DeepSeek-V3.2 makes 1 in its 34 feasible responses. Denominators exclude 7 of the 400 responses: three that the classifier labelled unclear and four whose supporting quote could not be verified. In at least 9 of the 33 false alarms the quoted sentence attributes the conflict to slot capacity (developer counts, hours, or story points read as the capacity unit), where the historical prompts are ambiguous (Section 3.3); these are not

simple misreadings of the requirements. The turn-20 responses were not classified.

### 4.3. P3: Source Fidelity and Epistemic Calibration

The P3 paradigm probes a different competence from P1 and P2: whether models appropriately abstain when evidence is absent or instead fabricate unsupported claims.

Table 4 presents the P3 results. All five models achieved near-perfect contradiction identification (1.000) and resolution (0.973–1.000). Gap abstention ranged from 0.875 (DeepSeek-V3.2 and GPT-4o) to 0.975 (MiniMax M2.5). Fabrication rates were low but non-zero for three models: Sonnet 4.5 at 5%, DeepSeek-V3.2 and GPT-4o at 10%, while Gemini 2.5 Pro and MiniMax M2.5 produced zero fabrications. On authority bias, across 13 scenarios where the high-authority source was wrong, *every model scored zero inappropriate deference*. No model was misled by status alone. In the 7 scenarios where authority was correct, all models deferred appropriately at 100%, except Sonnet 4.5 and GPT-4o at approximately 86%, which is a mild over-skepticism rather than a failure of reasoning.

Fabrication by gap type.

We classified the information gaps in the P3 scenarios into five types by what is missing. In the 20 core scenarios, fabrication appeared mainly at “missing evidence” gaps, where sources imply that supporting evidence exists without providing it. To examine this with a larger sample we constructed eight additional missing-evidence scenarios (included in the released dataset) spanning scientific, legal, and technical domains and

**Table 3.** P2 handling of unsatisfiable requests. Output types are counts of responses on infeasible checkpoints (23 scenarios at turn 17, 28 at turn 20; turn 15 requests no plan). “Asks user”: the response explicitly asks the user how to resolve the conflict (phrase patterns released with the artifact; generic closing offers are not counted). “States incompatibility”: responses whose visible answer asserts that the active requirements cannot all be satisfied, among the 23 infeasible scenarios; “False alarm”: the same assertion among the 17 feasible scenarios (rubric classification, Section 3.3; responses treated as unclear are excluded; turn 20 was not classified).

| Model          | Turn | No plan | Partial plan | Complete plan | Asks user | States incompatibility | False alarm |
|----------------|------|---------|--------------|---------------|-----------|------------------------|-------------|
| DeepSeek-V3.2  | T15  | —       | —            | —             | —         | 20/22                  | 1/17        |
|                | T17  | 0       | 1            | 22            | 6         | 20/21                  | 0/17        |
|                | T20  | 0       | 0            | 28            | 1         | —                      | —           |
| Sonnet 4.5     | T15  | —       | —            | —             | —         | 20/23                  | 2/17        |
|                | T17  | 11      | 5            | 7             | 14        | 22/23                  | 4/17        |
|                | T20  | 4       | 4            | 20            | 0         | —                      | —           |
| Gemini 2.5 Pro | T15  | —       | —            | —             | —         | 23/23                  | 6/17        |
|                | T17  | 12      | 10           | 1             | 5         | 23/23                  | 6/17        |
|                | T20  | 10      | 12           | 6             | 1         | —                      | —           |
| MiniMax M2.5   | T15  | —       | —            | —             | —         | 18/22                  | 2/17        |
|                | T17  | 4       | 2            | 17            | 16        | 19/23                  | 5/16        |
|                | T20  | 1       | 2            | 25            | 3         | —                      | —           |
| GPT-4o         | T15  | —       | —            | —             | —         | 9/22                   | 0/16        |
|                | T17  | 0       | 19           | 4             | 17        | 12/23                  | 7/17        |
|                | T20  | 0       | 11           | 17            | 16        | —                      | —           |

**Table 4.** P3 multi-document synthesis results. All models achieve zero inappropriate authority deference. MiniMax M2.5, the weakest numeric tracker in P1, leads or ties for the lead on every source-fidelity metric in this case study.

| Model          | Contr.<br>Res. | Gap<br>Abst. | Fab.<br>Rate | Appr.<br>Def. | Inappr.<br>Def. |
|----------------|----------------|--------------|--------------|---------------|-----------------|
| Sonnet 4.5     | 0.980          | 0.885        | 0.050        | 86%           | 0%              |
| DeepSeek-V3.2  | 1.000          | 0.875        | 0.100        | 100%          | 0%              |
| Gemini 2.5 Pro | 1.000          | 0.930        | <b>0.000</b> | 100%          | 0%              |
| GPT-4o         | 0.973          | 0.875        | 0.100        | 86%           | 0%              |
| MiniMax M2.5   | 1.000          | <b>0.975</b> | <b>0.000</b> | 100%          | 0%              |

ran all five models on them, which expands the missing-evidence sample from 5 to 13 scenarios and leaves the other gap types unchanged.

Table 5 summarizes the result. Because the five models share each scenario, the unit of inference is the scenario, and the primary endpoint is whether any of the five models fabricated on it; model-scenario pair counts are reported as descriptive breakdowns only. Three of the 13 missing-evidence scenarios produced at least one fabrication (five fabricating responses in total, three of them on one scenario, P3\_019), one of the six missing-detail scenarios produced one, and the nine scenarios of the remaining three types produced none. Fisher’s exact test at scenario level, missing evidence versus all other types, gives  $p = 0.31$ . The pattern is therefore suggestive rather than statistically established, and the two types with only two scenarios provide little evidence either way. We use the term *implication-to-certainty collapse* to describe the observed cases: when a gap sits at the boundary between what is stated and what is implied, the fabricating responses convert a qualified implication into a stated certainty. Whether this is a property of the gap type is the question the framework’s taxonomy is designed to let future collections answer.

#### Contemporary supplement.

Because the two least-sampled gap types had two scenarios each, we collected a supplement in September 2026 under a plan fixed before collection (released with the artifact): four new scenarios of each of those types, plus the eight supplementary missing-evidence scenarios re-collected in the same batch for comparison. The batches are never pooled. Four of the five

models were collected. The endpoint that served DeepSeek-V3.2 in February no longer serves that model, and the planned third-party-hosted session had not been run at the time of writing, so the prespecified scenario-level endpoint, which requires all five models, is not reported. Descriptively, the judge labelled a fabrication in 0 of 16 missing-reason pairs, 0 of 16 implied-not-stated pairs, and 1 of 32 missing-evidence comparison pairs (GPT-4o, on a scenario that was flagged for Sonnet 4.5 in February; confirmed in the author’s unblinded audit). On the comparison scenarios, per-model fabrication labels agreed between the two collections in 30 of 32 cases. These counts are too small to confirm or refute a gap-type effect.

As a concrete example, one scenario presented clinical trial documents where an internal memo stated that human trials “excluded drinkers,” and a gap probe asked whether any human participants consumed alcohol. Three of five models responded with a definitive “no,” converting an enrollment exclusion criterion (who was *eligible* to participate) into a behavioral claim about what participants *did* (whether anyone consumed alcohol during the study). The response states as certainty what the documents only imply.

A second example comes from the legal domain: a cease-and-desist letter states “we have evidence you created [the algorithm] while employed by us” without describing that evidence, and one model named a specific causal link between two documents that no source establishes.

**Table 5.** Fabrication by gap type across the core and expanded P3 scenarios (February 2026 batch). Primary endpoint: scenarios on which any of the five models fabricated. Pair-level counts (model  $\times$  scenario) are descriptive. Scenario-level Fisher’s exact test, missing evidence versus all other types:  $p = 0.31$ .

| Gap type           | Scenarios | Scenarios with any fabrication | Pairs | Fabricating responses |
|--------------------|-----------|--------------------------------|-------|-----------------------|
| Missing evidence   | 13        | 3                              | 65    | 5                     |
| Missing detail     | 6         | 1                              | 30    | 1                     |
| Missing number     | 5         | 0                              | 25    | 0                     |
| Missing reason     | 2         | 0                              | 10    | 0                     |
| Implied not stated | 2         | 0                              | 10    | 0                     |

**Table 6.** Cross-paradigm summary. Ranks in parentheses, ties receiving the average of the tied positions. P2 primary metric: complete valid plans on the 29 feasible post-revision checkpoints; P2 satisfaction on the same checkpoints is shown for reference.

| Model          | P1 phase-end acc. | P2 valid plans (feasible) | P2 sat. (feasible) | P3 gap abstention |
|----------------|-------------------|---------------------------|--------------------|-------------------|
| DeepSeek-V3.2  | <b>.995</b> (1)   | <b>24/29 = .83</b> (1)    | .990               | .875 (4.5)        |
| Sonnet 4.5     | .985 (2)          | 16/29 = .55 (2)           | .959               | .885 (3)          |
| Gemini 2.5 Pro | .965 (3)          | 15/29 = .52 (3)           | .818               | .930 (2)          |
| GPT-4o         | .955 (4)          | 3/29 = .10 (5)            | .860               | .875 (4.5)        |
| MiniMax M2.5   | .880 (5)          | 10/29 = .34 (4)           | .914               | <b>.975</b> (1)   |

#### 4.4. Cross-Paradigm Patterns

Table 6 collects the primary metric of each paradigm: P1 phase-end accuracy, the P2 complete valid plan rate on feasible post-revision checkpoints (with the fractional satisfaction on the same checkpoints for reference), and P3 gap abstention. Figure 2 shows the same values and the corresponding failure rates.

Two patterns stand out. First, numeric accuracy and feasible-plan success rank the models similarly: the Spearman correlation between the P1 and P2 columns is  $+0.90$  (exact permutation  $p = 0.08$ ), with DeepSeek-V3.2 first on both and GPT-4o and MiniMax M2.5 at the bottom of both. Second, source fidelity ranks them differently: MiniMax M2.5 is last on P1 and first on P3 gap abstention, and Gemini 2.5 Pro is second on P3 while third on P1 and P2. The P1–P3 correlation is  $-0.56$  ( $p = 0.40$ ) and the P2–P3 correlation  $-0.21$  ( $p = 0.73$ ); with five models we report exact two-sided permutation  $p$ -values, because the usual  $t$  approximation is unreliable at this sample size. Kendall’s  $W$  across the three rankings, with tie correction, is  $0.367$  ( $\chi^2 = 4.41$ ,  $df = 4$ ,  $p = 0.35$  under the chi-square approximation; exact permutation  $p = 0.41$ ). With five models these correlations are descriptive; they are reported to characterize this snapshot, not to establish population relationships.

Because the five-model statistics are weak, we ask which parts of the pattern survive resampling of the scenarios that produce the scores. With 2,000 scenario-level bootstrap resamples (the same draw applied to all five models; for P2 a resampled scenario carries all of its feasible post-revision checkpoints), MiniMax M2.5 is last on P1 and first on P3 in 93% of resamples, DeepSeek-V3.2 is first on the P2 valid-plan rate in 98%, and GPT-4o is last on it in 98%; the ordering of the three middle models on P2 is not stable. These intervals describe uncertainty over scenarios for these fixed runs; they say nothing about run-to-run variability (Appendix B) or about models outside the sample.

The original submission reported low concordance and rank inversions between all three paradigms and interpreted them as evidence of dissociable competencies. Under the corrected P2 protocol, the P1–P2 inversion disappears; what remains is the divergence between numeric accuracy and feasible planning on one side and source-fidelity abstention on the other, together with the separate behaviors on unsatisfiable requests. Confirming or refuting that divergence with more models is a primary direction for future use of the benchmark.

#### 5. Discussion

What the case study shows.

In this five-model snapshot, numeric accuracy over a supplied table is high for all models, and the models that are most accurate there also most often produce complete valid plans when one exists. Source-fidelity abstention does not follow that order. The framework’s value in the case study is not a ranking but the separation of outcomes that an aggregate score would merge: whether a model completes a plan, whether it respects stated capacities, whether it says that requirements cannot all be met, whether it raises that claim when they can, and whether it declines to answer a question the sources cannot answer.

Unsatisfiable requests.

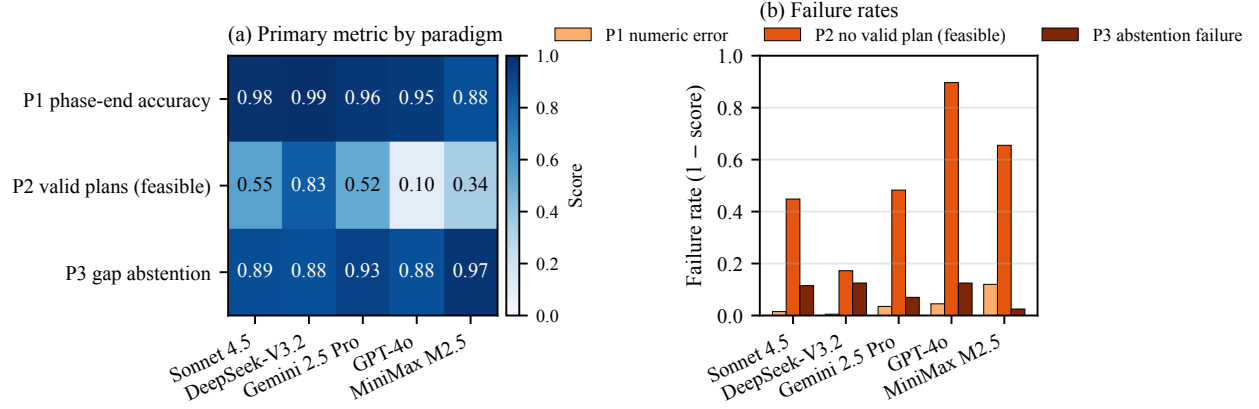
The models’ responses to jointly unsatisfiable requirements differ in kind. DeepSeek-V3.2 usually supplies a complete plan, and in over half of the cases that plan satisfies everything except the added requirement. Sonnet 4.5 and Gemini 2.5 Pro supply no plan in about half of the cases. GPT-4o supplies partial plans and asks the user to resolve the conflict. MiniMax M2.5 usually supplies a complete plan that violates further requirements. We describe these as output types rather than as success or failure because the historical prompt asked for a plan and for conflicts to be flagged without stating which requirement should yield. Which behavior is preferable depends on the application; the framework makes the behaviors visible. Any explanation in terms of training or alignment would be speculation, and we offer none.

Task-dependent failure profiles.

The divergence between numeric and planning accuracy on one side and source-fidelity abstention on the other means that a model chosen for one kind of task may be a poor choice for another. In this snapshot, MiniMax M2.5 is the most conservative on unanswerable questions and the least accurate on numeric recomputation; GPT-4o’s numeric accuracy is high (0.955) but it rarely commits to a complete plan. Whether such profiles are stable across models and versions is an empirical question the benchmark is designed to let others test.

Authority bias.

The zero inappropriate authority deference observed across all five models is reassuring but conditional. The authority



**Fig. 2.** Cross-paradigm summary. (a) Primary metric of each paradigm: P1 phase-end accuracy, P2 complete valid plan rate on feasible post-revision checkpoints, and P3 gap abstention. (b) The corresponding failure rates (one minus each score).

**Table 7.** ContextFidelity-Bench artifacts prepared for release. The repository/archive includes both generated evaluation data and executable tooling for extension.

| Component                        | Contents                                                                                                                                                                                                     | Reuse purpose                                                    |
|----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------|
| Scenario generators              | P1 financial-data generator; P2 reverse-generated scheduling tasks; P3 document templates and gap annotations                                                                                                | Generate fresh benchmark instances and reduce contamination risk |
| Prompt templates                 | System and user prompts for all paradigms and checkpoints                                                                                                                                                    | Replicate the evaluation protocol across model providers         |
| Scoring scripts                  | Numeric tolerance scorer with adjudication records; P2 protocol v3.1 (plan extractor, feasibility verifier, effective-requirement scorer, incompatibility rubric); P3 judging rubric and reliability scripts | Reproduce scores and audit individual model outputs              |
| Model outputs                    | 550 multi-turn conversations (10,200 assistant turns), 4,250 scored P1 answers, 800 scored P2 checkpoints, and 100 synthesis evaluations                                                                     | Reanalyze the case study without re-querying APIs                |
| Analysis and replication scripts | Scoring, analysis, figure and table scripts; collection protocols and API clients                                                                                                                            | Reproduce reported analyses and guide additional collections     |

direction is mixed (13 scenarios in which the high-status source is wrong, 7 in which it is right), so a policy of always distrusting the higher-status source is contradicted by the 86–100% appropriate deference on the seven authority-correct scenarios. This comparison rules out that one heuristic; it does not show that the models weighed the evidence rather than responding to other cues, and it does not test subtler authority signals. Matched authority reversals and document-order controls would test the broader question.

Implications for evaluation practice.

Current leaderboards typically report aggregate scores. The outcomes separated here suggest reporting, for multi-turn settings, at least the following: accuracy on tasks with exact answers, complete-valid-plan rates on verified-feasible planning states, the handling of verified-unsatisfiable states, and abstention on unanswerable questions, each with its unit of analysis and denominator stated.

## 6. Released Artifacts and Reproducibility

The primary contribution of this work is the released evaluation framework, designed to support replication and extension. The released artifacts comprise the following components:

**Scenario generators.** The P1 generator produces synthetic company data with 15 question types and pre-computed ground truth. The P2 generator constructs a satisfying schedule and derives requirements from it. The released generator version verifies at every checkpoint whether the active requirements are jointly satisfiable, gives each slot a single capacity value with a stated unit, and adds an explicit priority instruction for the added requirement; the historical scenarios are preserved unchanged with their corrected feasibility annotations.

**Scoring scripts.** P1 uses a numeric comparison with 1% tolerance and the released adjudication records. P2 uses protocol version 3.1: block-scoped plan extraction with source spans, effective requirements including capacities, verified feasibility labels, and the frozen incompatibility rubric with its classification outputs and validation. The superseded P2 scoring code is retained for comparison. For P3 we release the judge rubric, the scoring schema, and the cross-provider validation scripts.

**Replication and extension.** The released scripts reproduce the case-study tables and the analyses of Appendices A–C. The framework accommodates additional models, and the P3 gap taxonomy gives a structured way to add scenario types.

## 7. Limitations and Future Work

Several limitations of the current case study warrant acknowledgment, alongside opportunities for extending the framework.

**Model sample size.** The cross-paradigm concordance statistics are computed over five models, limiting statistical power. The primary support for paradigm-specific variation comes from within-paradigm results computed over hundreds of scored answers and checkpoints and from the scenario-level stability of the divergence between numeric accuracy and source-fidelity abstention, rather than from the five-model concordance statistics. Evaluating additional and more diverse models — particularly open-weight models where architecture and training pipelines can be inspected, and non-autoregressive architectures such as diffusion language models — is the most direct way to strengthen or refute the cross-paradigm findings.

**Single-run design.** Each model-scenario combination was run once. A targeted replication of 20 selected scenarios (Appendix B) found that P1 answers were stable but that P2 outputs on the hardest scenarios vary materially between runs, including a 0.18 change in one model’s mean satisfaction on the replicated feasible checkpoints. Point estimates for closely clustered models should be read as approximate, and the scenario-bootstrap intervals do not capture this run-to-run variation.

**LLM-as-judge evaluation.** P3 evaluation used an LLM judge from the same model family as one subject. The cross-provider validation (97.2% binned agreement, AC1 = 0.971, raw-score exact agreement 93–100%) and author audit provide confidence that the rubric produces consistent and directionally accurate scores, but independent human annotation of all P3 outputs would provide stronger validation. We plan to add blinded human annotation in a future version of the benchmark.

**Synthetic scenarios.** All scenarios are synthetic, with planted contradictions, explicit gaps, and clearly attributed sources. In naturalistic settings, contradictions are implicit, gaps are unmarked, and authority signals are subtler. The direction of this bias is likely conservative: implicit contradictions should be harder to detect, making our near-perfect detection rates an upper bound. A small validation slice using public real-world documents would strengthen the framework’s external validity and is a priority for future releases.

**Gap-type sample sizes.** The unit of inference for the gap-type analysis is the scenario, and two gap types have only two scenarios each; the scenario-level comparison between missing-evidence and other gaps is not significant ( $p = 0.31$ ). The contemporary supplement (Section 4.3) adds four scenarios to each of those types but covers four of the five models, so it is reported descriptively; completing that panel and expanding every gap type with contemporaneously collected comparison cases are the natural next steps.

**Endpoint aliasing.** One evaluated model was accessed through a provider alias (`deepseek-reasoner`) that has since been repointed to a newer model. We document the model the alias resolved to during our runs (DeepSeek-V3.2, thinking mode), but exact re-runs now require third-party hosting of the open weights. Pinned model identifiers should be preferred wherever a provider offers them.

**Scoring corrections and their limits.** The P2 results in this version follow a corrected protocol (Section 3.3); the corrections were made after the original submission and are disclosed there and in the artifact. The extraction rules were validated on selected cases (a 40-case spot-check and

23 previously refusal-labelled responses); a later corpus-wide consistency check found and corrected 25 of the 800 core extractions (Appendix C), and undetected errors may remain. The incompatibility-claim classification was compared with two samples labelled in a separate AI-assisted review, not with human annotation of the corpus. Capacity is scored as a maximum item count although the historical prompts left the unit unstated, and the prompts state one slot’s capacity twice with different values in 23 of 40 scenarios; absolute valid-plan counts depend on how that is read (DeepSeek-V3.2: 24 or 29 of 29), with no rank reversals (Sonnet 4.5 and Gemini 2.5 Pro tie under the alternative reading). P1 answer extraction accepts a number anywhere in the response; a selected inspection corrected seven cases, but corpus-wide extraction accuracy was not estimated. The P1 model order is stable for relative tolerances between 0.1% and 5%, apart from a tie for first place at 5% (Appendix C).

**Pairwise cross-model correlations.** The negative P1–P3 association ( $\rho = -0.56$ ) observed in this case study is descriptive only. It cannot be treated as a reliable finding at  $n = 5$  and would require confirmation with additional models.

## 8. Conclusion

We have introduced ContextFidelity-Bench, a three-paradigm evaluation framework for studying hallucination-related behavior in multi-turn conversations. The framework separates numeric accuracy over supplied data, constraint satisfaction under revision with verified feasibility labels, and source fidelity with a typed gap taxonomy, and provides scenario generators, a documented scoring protocol, prompt templates, and a dataset of 550 conversations to support replication and extension.

An initial case study of five API-accessible models illustrates the framework’s diagnostic use. Numeric accuracy is high for all models (0.880–0.995 on phase-end questions) and feasible-plan success follows a similar order, from 24 of 29 complete valid plans to 3 of 29. Responses to unsatisfiable requests differ in kind, and source-fidelity abstention ranks the models differently from the other two paradigms. These observations come from a single run of five models, and the P2 results follow a scoring protocol corrected during review; they are offered as hypotheses for the framework’s future users rather than as findings about the models. We release the full experimental framework, generators, scored results with their corrections, and replication scripts to support extension with additional models, scenarios, and evaluation dimensions.

## Ethical Statement

No ethical approval was required for this study, as it did not involve human or animal subjects.

This study evaluates computational systems on synthetic financial and scheduling scenarios and manually constructed document sets. It does not involve human participants, human-derived materials, animal subjects, or personally identifying data. The evaluated AI systems and AI-assisted judging tools are described in Section 3.4 and Section 3.5.

## Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## Declaration of competing interests

The author is employed by Google. One of the five models evaluated in the case study (Gemini 2.5 Pro) is a Google product. The research was conducted independently and does not represent the views or endorsement of Google. The author’s employer had no role in the design, execution, analysis, or reporting of this study.

## Declaration of generative AI use

During this work the author used AI coding assistants (Anthropic Claude and OpenAI Codex) to help write the analysis and scoring code. The author takes full responsibility for the content of this publication.

## Data Availability Statements

The scenario generators, prompt templates, data-collection and scoring scripts, model outputs, scored results, figure-generation scripts, and replication instructions are publicly available at <https://github.com/ilpenzo/hallucination-fragility> and archived at <https://doi.org/10.5281/zenodo.23074039> under the MIT License. The archive contains all generated data needed to reproduce the results reported in the paper, except for access credentials required to re-query commercial model APIs.

## CRediT authorship contribution statement

The author was responsible for Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Investigation, Validation, Visualization, Project Administration, Writing - Original Draft, and Writing - Review & Editing.

## A. Inter-Rater Reliability

To validate the P3 LLM judge, all 100 entries (5 models  $\times$  20 scenarios) were independently re-judged by GPT-5.2 (OpenAI, reasoning effort: medium) using the identical rubric and scoring schema applied by the primary judge (Claude Opus 4.6).

Scores were binned into ordinal categories (low: [0, 0.25), mid: [0.25, 0.75), high: [0.75, 1.0]). Because three of five dimensions exhibit extreme marginal skew (nearly all entries scored “high” by both judges), Cohen’s  $\kappa$  is subject to the Feinstein–Cicchetti prevalence paradox [22], producing near-zero or negative values despite  $\geq 95\%$  agreement. We therefore report Gwet’s AC1 [23] as the primary agreement statistic, which remains well-behaved under prevalence skew, alongside raw agreement and  $\kappa$  where the latter is interpretable.

**Table 8**

Inter-rater reliability: Claude Opus 4.6 vs. GPT-5.2 on all 100 P3 entries. AC1 is Gwet’s prevalence-robust agreement coefficient.  $\kappa_{\text{lin}}$ : Cohen’s  $\kappa$  with linear weighting (reported where marginal variance permits interpretation). The overall agreement percentage covers all 500 pairs; the overall AC1 and  $\kappa_{\text{lin}}$  cover the 400 ordinal pairs.

| Dimension         | Agree        | AC1          | $\kappa_{\text{lin}}$ | Note           |
|-------------------|--------------|--------------|-----------------------|----------------|
| Contr. identified | 100%         | 1.000        | undef.                | Prevalence     |
| Contr. resolved   | 95%          | 0.949        | −0.013                | Prev. paradox  |
| Reasoning quality | 98%          | 0.980        | −0.010                | Prev. paradox  |
| Gap abstention    | 96%          | 0.955        | 0.877                 | Almost perfect |
| Fabrication       | 97%          | 0.967        | 0.712                 | Substantial    |
| <b>Overall</b>    | <b>97.2%</b> | <b>0.971</b> | <b>0.718</b>          |                |

Table 8 presents the full results. Overall binned agreement is 97.2% across 500 paired dimension scores (5 dimensions  $\times$  100 entries); the overall AC1 (0.971) and  $\kappa_{\text{lin}}$  (0.718) are computed over the 400 ordinal pairs, and fabrication is a separate binary comparison. Gap abstention and fabrication, the two dimensions with sufficient marginal variance for standard  $\kappa$ , yield  $\kappa_{\text{lin}} = 0.877$  (almost perfect) and  $\kappa = 0.712$  (substantial), respectively. Bootstrap 95% confidence intervals (2,000 replicates) for gap abstention are AC1: [0.907, 0.989],  $\kappa_{\text{lin}}$ : [0.735, 0.973].

The negative  $\kappa$  values for contradiction resolution and reasoning quality reflect the prevalence paradox, not disagreement: both judges assign nearly all entries to the “high” bin. PABAK (prevalence-adjusted  $\kappa$ ; [22]) for these dimensions is 0.925 and 0.970 respectively, consistent with the AC1 values.

**Table 9**

Raw-score (unbinned) agreement between the two judges on the same 100 entries. “Within 0.1” is the share of entries whose scores differ by at most 0.1; the Pearson correlation is undefined where one dimension has no variance.

| Dimension            | Exact | $\leq 0.1$ | MAD   | Max  | $r$    |
|----------------------|-------|------------|-------|------|--------|
| Contr. identified    | 100%  | 100%       | 0.000 | 0.00 | undef. |
| Contr. resolved      | 93%   | 94%        | 0.024 | 0.80 | 0.47   |
| Reasoning quality    | 96%   | 98%        | 0.009 | 0.50 | 0.25   |
| Gap abstention       | 94%   | 94%        | 0.020 | 0.50 | 0.95   |
| Fabrication (binary) | 97%   | —          | —     | —    | —      |

Table 9 reports agreement on the raw scores before binning. The two dimensions with near-zero variance (contradiction resolution and reasoning quality) have low raw correlations for the same prevalence reason that depresses  $\kappa$ , but exact agreement of 93–96% and mean absolute differences below 0.025; gap abstention, where scores vary, correlates at 0.95.

Only three fabrication judgments (across two scenarios) disagreed, and per-model binned agreement ranged from 95% to 100% on every dimension.

## B. Targeted Replication

To assess sensitivity to the single-run design, we replicated 20 scenarios selected for high cross-model variance or high stakes (5 P1, 10 P2, 5 P3; 100 conversations, 5 models, temperature 0, collected in late February 2026). Variation at temperature 0 reflects implementation nondeterminism (for example floating-point and batching differences) rather than sampling; we did not test its cause.

For P1, scored with the same final extraction rule as the main results, the four phase-end answers agreed between runs on 95 of 100 comparisons; per-model phase-end accuracy on the five replicated scenarios changed by at most 0.10 (MiniMax M2.5, 0.80 to 0.90).

For P2, scored with protocol v3.1, the ten replicated scenarios were chosen as the hardest post-conflict cases and all but one are infeasible after turn 15, so the replication speaks mainly to the feasible checkpoints at turns 6 and 11 (21 matched feasible

checkpoints per model). Mean satisfaction of the effective requirements on those checkpoints changed between runs by +0.18 for Sonnet 4.5 (0.74 to 0.92), +0.08 for GPT-4o, +0.06 for Gemini 2.5 Pro, and by less than 0.01 for MiniMax M2.5 and DeepSeek-V3.2. Sonnet 4.5’s change comes from checkpoints where it declined to plan in the original run (asking which capacity value applies, or asserting a conflict) and supplied a plan in the replication, a full-scale change for that checkpoint. Complete valid plans on the 21 checkpoints changed from 14 to 15 (Sonnet 4.5), 20 to 17 (DeepSeek-V3.2), 16 to 18 (Gemini 2.5 Pro), 0 to 1 (GPT-4o), and 15 to 12 (MiniMax M2.5). DeepSeek-V3.2 had the highest satisfaction in both runs; on valid plans it was first in the original run and second to Gemini 2.5 Pro in the replication. We report this variation without attributing it to a cause, and Table 2’s scenario-bootstrap intervals do not include it.

For P3, fabrication judgments agreed on 23 of 25 scenario-model pairs (92%). Of the five original fabrication-labeled pairs in the replicated subset, three replicated on the same scenario-model pairs and two shifted from fabrication to partial abstention. All 20 originally non-fabricating pairs remained clean.

### C. Sensitivity Analyses

Plan extraction.

Table 10 classifies the 160 checkpoint responses per model by the plan the model explicitly proposed, under the block-scoped extraction rules of Section 3.3. The originally submitted extractor read assignments from outside the proposed plan (in 118 of GPT-4o’s and 53 of Gemini 2.5 Pro’s checkpoints it assigned items the model had listed as unassigned), which inflated those models’ satisfaction; the rigid parser of the first scoring pass counted unparsed plans as failures. Neither is used in this version. A corpus-wide comparison of the block-scoped extractor against a permissive scan then exposed five kinds of false negative, listed in the artifact (for example code fences counted as blank lines, and the abbreviation “HR”). The corrected rules changed 25 of 800 core and 9 of 200 replication extractions, mostly for MiniMax M2.5; each change is listed with its source text in the artifact and was checked in AI-assisted review, not by human annotation. The phrase patterns behind the “asks user” counts were corrected in the same pass (question forms added, generic closing offers removed).

**Table 10**

Plan output types over all 160 checkpoint responses per model (block-scoped extraction). “No usable plan”: a placeholder template or an all-unassigned list. Cov.: mean share of the eight items assigned.

| Model          | Complete | Partial | No usable | No plan | Cov. |
|----------------|----------|---------|-----------|---------|------|
| DeepSeek-V3.2  | 159      | 1       | 0         | 0       | 1.00 |
| Sonnet 4.5     | 121      | 18      | 1         | 20      | 0.80 |
| MiniMax M2.5   | 140      | 15      | 3         | 2       | 0.93 |
| Gemini 2.5 Pro | 90       | 47      | 6         | 17      | 0.68 |
| GPT-4o         | 34       | 126     | 0         | 0       | 0.56 |

Capacity reading.

Table 11 compares the primary capacity rule (both capacity statements must hold) with the alternative reading (the later sentence replaces the table value) on the feasible checkpoints. The verified feasibility labels are identical under both. The

Spearman correlation between P1 and the P2 valid-plan rate is +0.90 under the primary rule and +0.87 under the alternative.

**Table 11**

Complete valid plans and capacity violations under the two capacity readings (primary / alternative). Valid plans: of 40 checkpoints at turns 6 and 11 and of 29 feasible post-revision checkpoints. Violations: responses placing more items in a slot than allowed, among responses that assign at least one item ( $n$ ).

| Model          | T6      | T11     | Post-rev. | Violations | $n$ |
|----------------|---------|---------|-----------|------------|-----|
| DeepSeek-V3.2  | 33 / 40 | 36 / 40 | 24 / 29   | 24 / 2     | 160 |
| Sonnet 4.5     | 31 / 33 | 31 / 34 | 16 / 19   | 24 / 11    | 139 |
| Gemini 2.5 Pro | 26 / 32 | 27 / 32 | 15 / 19   | 17 / 2     | 137 |
| MiniMax M2.5   | 30 / 31 | 32 / 34 | 10 / 13   | 39 / 27    | 155 |
| GPT-4o         | 1 / 1   | 1 / 1   | 3 / 4     | 37 / 31    | 160 |

Numeric tolerance.

Table 12 recomputes P1 phase-end accuracy from the raw responses at relative tolerances from 0.1% to 10%, retaining each question’s absolute tolerance and using the final asserted answer for the seven adjudicated responses, so its 1% row reproduces Table 1. The model order at the 1% rule is also obtained at 0.1%, 0.5% and 2%; Sonnet 4.5 ties DeepSeek-V3.2 for first at 5% and 10%, and GPT-4o passes Gemini 2.5 Pro only at 10%.

**Table 12**

P1 phase-end accuracy by relative tolerance (seven adjudicated corrections included).

| Tol.      | Sonnet | DeepSeek | Gemini | GPT-4o | MiniMax |
|-----------|--------|----------|--------|--------|---------|
| 0.1%      | 0.985  | 0.995    | 0.965  | 0.915  | 0.875   |
| 0.5%      | 0.985  | 0.995    | 0.965  | 0.950  | 0.875   |
| 1% (used) | 0.985  | 0.995    | 0.965  | 0.955  | 0.880   |
| 2%        | 0.990  | 0.995    | 0.965  | 0.955  | 0.890   |
| 5%        | 0.995  | 0.995    | 0.965  | 0.960  | 0.915   |
| 10%       | 0.995  | 0.995    | 0.965  | 0.980  | 0.955   |

Incompatibility-claim classification.

The visible answer of each turn-15 and turn-17 response (400 responses) was labelled YES, NO, or UNCLEAR against a frozen written rubric by an LLM classifier (Claude Opus 4.6) with model identity, scenario, and feasibility hidden; each label must carry a verbatim quote. Of the 400 quotes, 396 were verified against the response; the four responses whose quote could not be verified are treated as unclear, together with three that the classifier itself labelled unclear. Two samples (60 selected cases, then 30 further cases) had been labelled beforehand in a separate AI-assisted review, not by human annotators. The classifier agreed with that review on 55 of 60 and on 28 of 29 decided cases (1 labelled unclear); all labels, quotes, disagreements, and the rubric are released.

### References

1. S. Lin, J. Hilton, and O. Evans, “TruthfulQA: Measuring how models mimic human falsehoods,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2022, pp. 3214–3252, doi: <https://doi.org/10.18653/v1/2022.acl-long.229>. [Online]. Available: <https://aclanthology.org/2022.acl-long.229/>
2. J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, “HaluEval: A large-scale hallucination evaluation

- benchmark for large language models,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2023, pp. 6449–6464, doi: <https://doi.org/10.18653/v1/2023.emnlp-main.397>. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.397/>
3. S. Min *et al.*, “FactScore: Fine-grained atomic evaluation of factual precision in long form text generation,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2023, pp. 12 076–12 100, doi: <https://doi.org/10.18653/v1/2023.emnlp-main.741>. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.741/>
  4. P. Liang *et al.*, “Holistic evaluation of language models,” *Transactions on Machine Learning Research*, 2023. [Online]. Available: <https://openreview.net/forum?id=iO4LZibEqW>
  5. A. Srivastava *et al.*, “Beyond the imitation game: Quantifying and extrapolating the capabilities of language models,” *Transactions on Machine Learning Research*, 2023. [Online]. Available: <https://openreview.net/forum?id=uyTL5Bvosj>
  6. J. Xie *et al.*, “TravelPlanner: A benchmark for real-world planning with language agents,” in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 235, 2024, pp. 54 590–54 613. [Online]. Available: <https://proceedings.mlr.press/v235/xie24j.html>
  7. K. Valmeekam, M. Marquez, A. Olmo, S. Sreedharan, and S. Kambhampati, “PlanBench: An extensible benchmark for evaluating large language models on planning and reasoning about change,” in *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023, pp. 38 975–38 987. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2023/hash/7a92bcdede88c7afd108072faf5485c8-Abstract-Datasets\\_and\\_Benchmarks.html](https://proceedings.neurips.cc/paper_files/paper/2023/hash/7a92bcdede88c7afd108072faf5485c8-Abstract-Datasets_and_Benchmarks.html)
  8. Z. Ji *et al.*, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023, doi: <https://doi.org/10.1145/3571730>.
  9. N. F. Liu *et al.*, “Lost in the middle: How language models use long contexts,” *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024, doi: <https://doi.org/10.1162/tacl.a.00638>. [Online]. Available: <https://aclanthology.org/2024.tacl-1.9/>
  10. L. Zheng *et al.*, “Judging LLM-as-a-judge with MT-Bench and Chatbot Arena,” in *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets\\_and\\_Benchmarks.html](https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html)
  11. K. Chen, Q. Chen, J. Zhou, Y. He, and L. He, “DiaHalu: A dialogue-level hallucination evaluation benchmark for large language models,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024, pp. 9057–9079, doi: <https://doi.org/10.18653/v1/2024.findings-emnlp.529>. [Online]. Available: <https://aclanthology.org/2024.findings-emnlp.529/>
  12. V. Rathore, S. Aneesh, and H. Singh, “Temporal graph network: Hallucination detection in multi-turn conversation,” *arXiv preprint*, 2026, arXiv:2601.03051. doi: <https://doi.org/10.48550/arXiv.2601.03051>. [Online]. Available: <https://arxiv.org/abs/2601.03051>
  13. Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, “G-Eval: NLG evaluation using GPT-4 with better human alignment,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2023, pp. 2511–2522, doi: <https://doi.org/10.18653/v1/2023.emnlp-main.153>. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.153/>
  14. W.-L. Chiang *et al.*, “Chatbot Arena: An open platform for evaluating LLMs by human preference,” in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235, 2024, pp. 8359–8388. [Online]. Available: <https://proceedings.mlr.press/v235/chiang24b.html>
  15. M. G. Kendall and B. Babington Smith, “The problem of  $m$  rankings,” *The Annals of Mathematical Statistics*, vol. 10, no. 3, pp. 275–287, 1939, doi: <https://doi.org/10.1214/aoms/1177732186>.
  16. K. Deshpande *et al.*, “MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs,” in *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, 2025, pp. 18 632–18 702, doi: <https://doi.org/10.18653/v1/2025.findings-acl.958>. [Online]. Available: <https://aclanthology.org/2025.findings-acl.958/>
  17. D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, and D. Yu, “LongMemEval: Benchmarking chat assistants on long-term interactive memory,” in *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025. [Online]. Available: <https://arxiv.org/abs/2410.10813>
  18. P. Laban, H. Hayashi, Y. Zhou, and J. Neville, “LLMs get lost in multi-turn conversation,” *arXiv preprint*, 2025, arXiv:2505.06120. [Online]. Available: <https://arxiv.org/abs/2505.06120>
  19. T. Gebru *et al.*, “Datasheds for datasets,” *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021, doi: <https://doi.org/10.1145/3458723>.
  20. A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, “Data and its (dis)contents: A survey of dataset development and use in machine learning research,” *Patterns*, vol. 2, no. 11, p. 100336, 2021, doi: <https://doi.org/10.1016/j.patter.2021.100336>.
  21. D. Kiela *et al.*, “Dynabench: Rethinking benchmarking in NLP,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 4110–4124, doi: <https://doi.org/10.18653/v1/2021.naacl-main.324>. [Online]. Available: <https://aclanthology.org/2021.naacl-main.324/>
  22. T. Byrt, J. Bishop, and J. B. Carlin, “Bias, prevalence and kappa,” *Journal of Clinical Epidemiology*, vol. 46, no. 5, pp. 423–429, 1993, doi: [https://doi.org/10.1016/0895-4356\(93\)90018-V](https://doi.org/10.1016/0895-4356(93)90018-V).
  23. K. L. Gwet, “Computing inter-rater reliability and its variance in the presence of high agreement,” *British Journal of Mathematical and Statistical Psychology*, vol. 61, no. 1, pp. 29–48, 2008, doi: <https://doi.org/10.1348/000711006X126600>.