SCICONVBENCH: BENCHMARKING LLMS ON MULTI-TURN CLARIFICATION FOR TASK FORMULATION IN COMPUTATIONAL SCIENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) are increasingly deployed as scientific AI assistants, and a growing body of benchmarks evaluates their capabilities across knowledge retrieval, reasoning, code generation, and tool use. These evaluations, however, typically assume the scientific problem is already well-posed, whereas practical scientific assistance often begins with an ill-posed user request that must be refined through dialogue before any computation, analysis, or experiment can be carried out reliably. We introduce SCICONVBENCH, a benchmark for multi-turn clarification in scientific task formulation across four computational science problem domains: *fluid mechanics*, *solid mechanics*, *materials science*, and *partial differential equations (PDEs)*. SCICONVBENCH targets two complementary capabilities: eliciting missing information (*disambiguation*) and detecting and correcting erroneous requests containing internally contradictory information (*inconsistency resolution*). Our benchmark pairs a structured task ontology with a rubric-based evaluation framework, enabling systematic measurement of LLM performance across three dimensions: clarification behavior, conversational grounding, and final-specification fidelity. Current frontier models perform relatively well on *inconsistency resolution*, but even the best model resolves only 51.2% of the disambiguation cases through dialogue, averaged across the domains. We further find that frontier LLMs frequently make silent assumptions and perform implicit specification repairs that are not grounded in the conversation with users. SCICONVBENCH establishes a foundation for evaluating the upstream conversational reasoning that a reliable computational science assistant requires. The code and data can be found at `https://anonymous.4open.science/r/SciConvBench-08DE`.

1 INTRODUCTION

Large language models (LLMs) are increasingly used as conversational interfaces for computational science, supporting scientific question answering (Somasekharan et al., 2026), code generation (Tian et al., 2024), and agentic execution of scientific simulation workflows (Yue et al., 2025; Pandey et al., 2025). Yet most scientific benchmarks for LLMs assess these capabilities given complete problem formulation, typically assuming a clean task statement with fixed objectives, constraints, and expected outputs (Wang et al., 2024a; Sun et al., 2024; Tian et al., 2024; Liu et al., 2025a; Chen et al., 2025). This omits an upstream failure mode in scientific practice: before a model can compute, write code, or invoke tools reliably, it may first need to transform an incomplete or internally inconsistent user request into a well-specified scientific task. In computational science, such formulation errors are consequential because a missing boundary condition, ambiguous material property, incompatible constitutive assumption, missing Reynolds number, or contradictory numerical constraint can alter the underlying problem, yielding a specification that is physically invalid, irreproducible, or misaligned with the

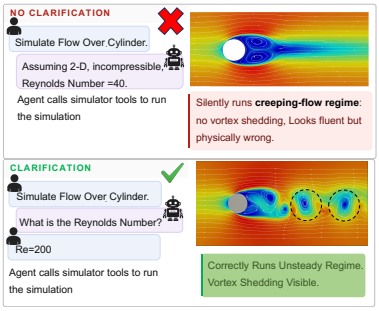

Figure 1: Flow over a cylinder showing how skipped clarification leads to a wrong flow regime.

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

user's intent. For example, Figure 1 illustrates the downstream consequence of unresolved prompt issues: if a critical parameter such as the Reynolds number is not clarified, an agent may silently run a plausible but incorrect flow regime, wasting computation and producing a result that is irrelevant to the intended scientific task.

Existing clarification and ambiguity benchmarks study follow-up questioning mainly in general-purpose or information-seeking settings (Lee et al., 2023; Zhang et al., 2024c; Aliannejadi et al., 2019; Kuhn et al., 2023; Gan et al., 2024), while multi-turn and agent benchmarks show that information distributed across dialogue remains difficult for current models (Deshpande et al., 2025; Kwan et al., 2024; Yao et al., 2024; Laban et al., 2025). These benchmarks, however, do not reflect the kinds of clarifications computational science actually demands, and the gap is quantitative as well as conceptual: under the same model and protocol, GEMINI 2.5 PRO resolves 86.1% of cases on a filtered subset of CLAMBER but drops to as low as 23.8% on SCICONVBENCH disambiguation (Table 1). The missing evaluation setting is whether a model can identify missing or conflicting scientific requirements and resolve them through dialogue before producing a final task specification. These results motivate a science-specific clarification benchmark, since general clarification benchmarks do not stress the domain groundedness required in computational science.

Benchmark	Domain	n	Resolution Rate
CLAMBER Zhang et al. (2024c)	General	115	86.1 %
SCICONVBENCH (ours)	Fluid mechanics	151	23.8 %
	Solid mechanics	228	29.8 %
	Materials science	130	57.7 %
	PDEs	61	70.5 %

Table 1: Comparison between a filtered subset of CLAMBER (Zhang et al., 2024c) (a general-domain clarification dataset) and the disambiguation split of SCICONVBENCH, both evaluated with GEMINI 2.5 PRO. Resolution rate drops drastically on SCICONVBENCH, indicating that computational-science domains pose a substantially harder clarification challenge than general-domain prompts.

We introduce SCICONVBENCH (Figure 2), a benchmark for multi-turn clarification of scientific task formulation across domains such as *fluid mechanics*, *solid mechanics*, *materials science*, and *partial differential equations (PDEs)*. Each instance begins with a scientific request containing either missing information, which requires disambiguation, or conflicting information, which requires inconsistency resolution. The model interacts with a user over multiple turns and then produces a final clarified specification. SCICONVBENCH evaluates *conversational scientific task formulation*, defined as the ability of a model to resolve incomplete or internally inconsistent scientific requests through dialogue and produce a usable final prompt or specification.

Our goal is to shift evaluation upstream. Before asking whether a model can solve, code, or execute a scientific task, we ask whether it can help define the task correctly. This paper makes three contributions. First, we formalize conversational scientific task formulation as a benchmark setting centered on unresolved ambiguity and unresolved inconsistency. Second, we introduce an evaluation framework that separates intent faithful final resolution from conversation-grounded resolution, exposing silent assumptions and silent repairs that standard end-state metrics can miss. Third, we benchmark current models across scientific domains and ontology categories, and analyze the robustness of conclusions across judges, prompts, and user simulators. Our results show that this upstream stage remains difficult for frontier models: no single model dominates across all tasks and domains, inconsistency resolution is substantially easier than missing-information elicitation, the leading model changes across the two tasks, and every model exhibits a persistent gap between final correctness and conversation-grounded resolution, with a larger gap on inconsistency resolution tasks than on disambiguation.

2 RELATED WORK

Clarification and ambiguity. Clarifying-question research has long studied when an assistant should ask rather than answer. Conversational retrieval and QA benchmarks such as QULAC, CLARIQ, and CLARQ evaluate clarification selection, ranking, generation, and large-scale question mining (Aliannejadi et al., 2019; 2020; Kumar & Black, 2020); AMBIGQA and CAMBIGNQ treat ambiguous questions as requiring multiple interpretations or explicit clarification before answering (Min et al., 2020; Lee et al., 2023); and CLAMBER, CONDAMBIGQA, CLAM, APA, future-turn RLHF, and proactive information-gathering work formalize ambiguity taxonomies, conditional ambiguity, strategic clarification, and high-value question asking under incomplete context (Zhang

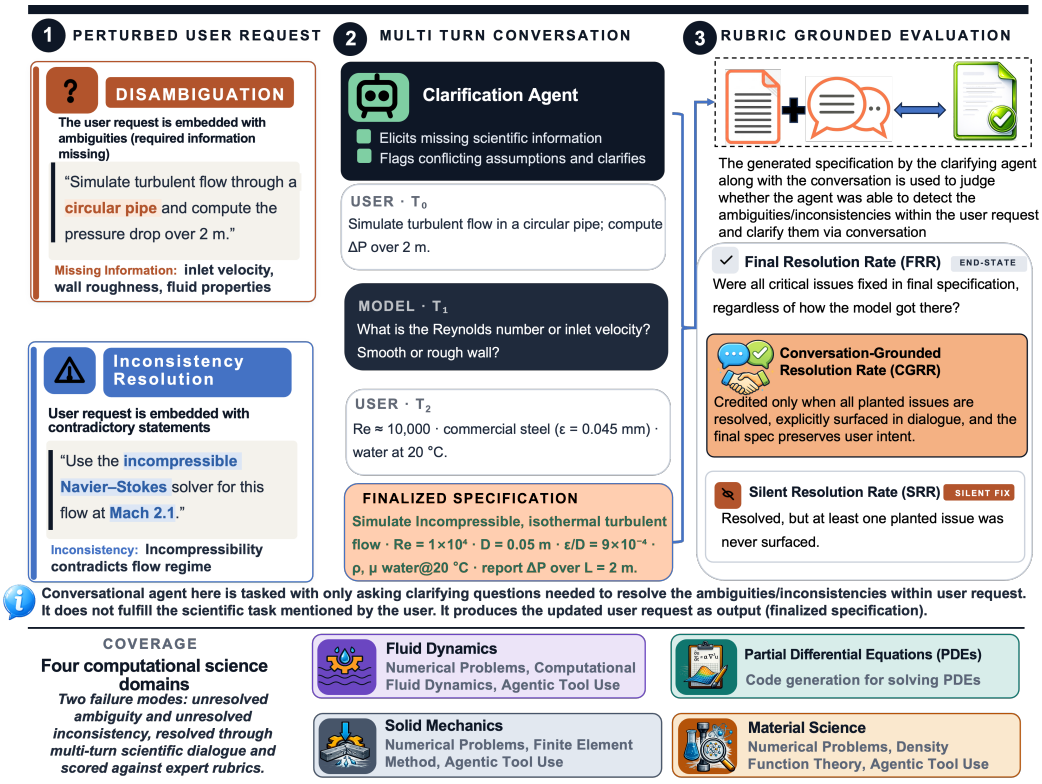

Figure 2: Overview of SCICONVBENCH. The benchmark spans four computational science domains and two task types. For each instance, a model interacts with a simulated user to resolve missing or conflicting information and then produces a final specification. Evaluation compares the final specification against the reference specification while using the full conversation as context to assess whether the model resolved the ambiguity or inconsistency in the original scientific request.

et al., 2024c; Li et al., 2025c; Kuhn et al., 2023; Kim et al., 2024; Zhang et al., 2024b; Huang et al., 2025). MediQ (Li et al., 2024b) introduces a hidden-information, ask-or-abstain protocol for interactive clinical reasoning. Related inconsistency and rule-grounded dialogue benchmarks include CONTRADOC, which localizes document contradictions, and SHARC, which requires follow-up questions when rule-grounded requests are underspecified (Li et al., 2024a; Saeidi et al., 2018). QUESTBENCH isolates information gathering for missing logical or mathematical preconditions, and CLARQ-LLM shows LLMs often answer instead of clarifying in task-oriented dialogue (Li et al., 2025a; Gan et al., 2024). These benchmarks establish clarification as a measurable capability, covering lexical ambiguity, hidden user information, missing logical or mathematical preconditions, and underspecified rule-grounded requests, but none requires eliciting or reconciling the commitments that define a scientific task, such as governing physics, numerics, units, or boundary conditions etc.

Multi-turn, agentic, and simulator-based evaluation. Multi-turn evaluation has moved from general dialogue quality to interaction robustness. MT-BENCH and LLM-as-a-judge evaluation exposed both the scalability and biases of automatic multi-turn judging (Zheng et al., 2023); CHATBOT ARENA, ARENA-HARD-AUTO, and length-controlled AlpacaEval study human-aligned large-scale ranking and verbosity control (Chiang et al., 2024; Li et al., 2025b; Dubois et al., 2025); and MT-EVAL, MULTICHALLENGE, LLMS GET LOST, and RMTBENCH show that models remain brittle when evidence is distributed across turns or users behave less cooperatively (Kwan et al., 2024; Deshpande et al., 2025; Laban et al., 2025; Xiang et al., 2025). Agent benchmarks such as AGENTBENCH, WEBARENA, GAIA, SWE-BENCH, MINT, and τ-bench evaluate tool, API, website, codebase, or simulated-user environments (Liu et al., 2025b; Zhou et al., 2023; Mialon et al., 2024; Jimenez et al., 2024; Wang et al., 2024b; Yao et al., 2024). Because these settings increasingly rely on user simulators, recent work studies simulator fidelity and robustness: τ-bench and τ^2-Bench adopt LLM-simulated users for scalable agent evaluation (Yao et al., 2024; Barres et al., 2025); MIRRORBENCH, reliable-simulator work, SIMULATORARENA, and non-collaborative

simulators analyze when simulated users preserve or distort measured assistant quality (Hathidara et al., 2026; Sekulic et al., 2024; Dou et al., 2025; Shim et al., 2026); and broader task-oriented and social-simulation work provides context for this protocol (Davidson et al., 2023; Ni et al., 2026; Argyle et al., 2022; Budzianowski et al., 2018).

Scientific benchmarks and domain-specific agents. Scientific evaluation has advanced rapidly, but most benchmarks assume that the task is already specified. SCIBENCH and SCIEVAL evaluate scientific reasoning and research tasks (Wang et al., 2024a; Sun et al., 2024); SCICODE, MATTOOLS, SCIENCEAGENTBENCH, SCIAGENT, and CHEMCROW evaluate research coding, materials-science tool use, data-driven discovery, tool-augmented scientific reasoning, and chemistry agents (Tian et al., 2024; Liu et al., 2025a; Chen et al., 2025; Ma et al., 2024; Bran et al., 2023). Computational-science agents and benchmarks similarly target executable workflows after formulation: OPENFOAMGPT, NL2FOAM, CFDLLMBENCH, and METAOPENFOAM for fluids and CFD (Pandey et al., 2025; Dong et al., 2025; Somasekharan et al., 2026; Chen et al., 2024); FEABENCH, AUTOFEA, and ALL-FEM for solids, FEA, PDE formulation, and code generation (Mudur et al., 2025; Hou et al., 2025; Deotale et al., 2026); and HONEYCOMB and MECHAGENTS for materials and mechanics workflows (Zhang et al., 2024a; Ni & Buehler, 2023). Our focus is the preceding conversational step: whether the model elicits or flags the scientific commitments needed to make execution meaningful.

3 SCICONVBENCH

3.1 BENCHMARK SCOPE AND DOMAINS

The benchmark spans four computational-science domains: *fluid mechanics*, *solid mechanics*, *materials science*, and *partial differential equations (PDEs)* and includes both general numerical problem statements and prompts requiring the invocation of domain-specific simulator tool. Each domain covers a different class of scientific task formulation (see Equation 1). *Fluid Mechanics* includes general fluid-mechanics problems and Computational Fluid Dynamics (CFD) prompts. *Solid Mechanics* includes mechanics and finite-element-style task formulation. *Materials Science* includes materials-science reasoning and Density Functional Theory (DFT) based task formulations. *Partial Differential Equations (PDEs)* includes mathematical PDE problem specification.

3.2 TASK DEFINITION

We define a scientific task formulation as a structured specification of the physical or computational study to be performed (Ashton et al., 2025). A clean task $y^\star$ is written as

$$y^\star := \big[\nu_{\mathrm{obj}}, \nu_{\mathrm{geom}}, \nu_{\mathrm{model}}, \nu_{\mathrm{prop}}, \nu_{\mathrm{bc}}, \nu_{\mathrm{ic}}, \nu_{\mathrm{num}}, \nu_{\mathrm{out}}, \nu_{\mathrm{tool}}\big], \tag{1}$$

where the entries denote the objective of the study, geometry or computational domain, governing physics or constitutive model, material or transport properties, boundary conditions, initial conditions, numerical controls, requested outputs, and tool-specific settings, collectively defining the *ontology* of a scientific task. A benchmark instance is obtained by perturbing a clean task $y^\star$ into an initial user request $x = T_z(y^\star)$. The perturbation set $z = \{(k_j, \tau_j)\}_{j=1}^{m}$ records the planted issues, where k_j indexes one of the entries of Equation (1) and $\tau_j \in \{\textsc{missing}, \textsc{conflict}\}$. If $\tau_j =$ MISSING, information required by entry k_j is omitted or left underspecified in the initial request. If $\tau_j =$ CONFLICT, the request contains mutually incompatible information for that entry, or an incompatibility between that entry and another part of the specification. The model interacts with the user over multiple turns and finally produces a specification $\hat{y}$. The benchmark evaluates whether $\hat{y}$ resolves all planted issues, preserves the intended task, and reaches this resolution through conversation rather than silent guessing or unannounced correction.

3.3 INTERACTION PROTOCOL

Each interaction begins from the transformed user request. The conversational agent may ask clarification questions over multiple turns before producing its final output. The agent is instructed to ask only one question per turn. The user responds only from the hidden reference specification for that instance and does not provide information outside the intended task. To keep interactions comparable across models and domains, we use a fixed turn budget of 10. This choice follows directly from dataset construction: each instance contains at most 8 planted ambiguities or inconsistencies, so

10 turns are sufficient in principle to address all issues in a case and produce a final specification. The conversation terminates either when the model explicitly finalizes the task or when the turn limit is reached (Sections C.2 and F.2), at which point the model must produce its final clarified specification. SCICONVBENCH does not require solver execution, code execution, or tool invocation for scoring; prompts may be tool-oriented, but evaluation is restricted to conversational task formulation, allowing the benchmark to be evaluated by any conversational agentic framework.

3.4 DATASET CREATION

We construct SCICONVBENCH in two stages. We first collect a pool of clean, well-posed scientific tasks, and then manually convert them into conversational instances containing either missing information (*disambiguation*) or conflicting information (*inconsistency resolution*). This design ensures that every benchmark item starts from a scientifically valid reference problem, and that the difficulty comes from task formulation rather than from noisy or ill-posed source data.

Source pool We assemble source tasks from vetted educational, benchmark, and tool-informed resources across four computational-science domains. Fluid and PDE tasks draw from standard texts, FoamBench and CFDCodeBench within CFDLLMBench, and SCICODE (Gerhart et al., 2020; White, 2011; Somasekharan et al., 2026; Tian et al., 2024). Solid mechanics tasks use standard mechanics texts and finite-element resources, including FEABENCH, AUTOFEA, ALL-FEM, FEniCS, and CalculiX (Beer et al., 2011; Gere & Goodno, 2008; Ugural & Fenster, 2011; Mudur et al., 2025; Hou et al., 2025; Deotale et al., 2026; Logg et al., 2011; Dhondt & Wittig, 1998). Materials tasks combine textbook problems (Callister & Rethwisch, 2013; Askeland et al., 2011; Shackelford, 2005) with DFT tasks drawing from MASCQA, MATSCIBENCH, and MATTOOLS (Zaki et al., 2024; Zhang et al., 2026; Liu et al., 2025a).

Prompt transformation Each source item is first normalized into a clean reference prompt admitting a coherent scientific answer or setup. For *disambiguation* cases, we remove information that a responsible assistant should request before finalizing the task, such as boundary conditions, constitutive assumptions, material or transport properties, solver settings, geometry details, target outputs, or numerical tolerances. For *inconsistency* cases, we insert incompatible or conflicting statements while keeping the overall request realistic. This transformation is performed prompt-by-prompt rather than through automatic templates, since the missing or conflicting information is strongly domain- and problem-dependent. Each missing entity or planted inconsistency is also tagged to one of the components in Equation (1).

Expert review and filtering Quality control was performed by experts who were not involved in authoring the original transformed prompt. Reviewers checked that the hidden or conflicting information was scientifically meaningful, that the case admitted a clear intended resolution, that the prompt did not leak the answer through trivial cues, and that the conversational variant remained realistic. The final case split across 1,142 total cases is shown in Figure 3.

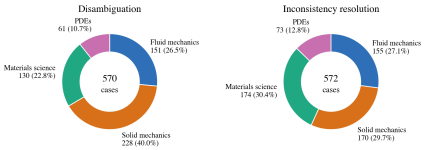

Figure 3: Case distribution across the four SCICONVBENCH domains.

3.5 EVALUATION PROTOCOL

Following recent conversational benchmark design (Bai et al., 2024; Deshpande et al., 2025; Yao et al., 2024), we separate final output success from conversation-grounded success, since a model may guess or silently repair missing scientific details without resolving them through dialogue. Each instance is evaluated as a structured judgment problem using the conversation transcript, the final specification, and the reference issue annotation. Because correct resolutions can vary in wording and dialogue path, exact string matching and handwritten heuristics are insufficient. We therefore use an LLM judge with an expert-curated rubric that defines the planted issue per case, successful resolution criteria, and the evidence required for conversational grounding. For every case, the judge is supplied with that case's specific missing entities or planted inconsistencies. This protocol follows prior evidence that strong LLM judges can reach high agreement with humans on open-ended evaluation when guided by explicit rubrics (Zheng et al., 2023; Wang et al., 2026; Hashemi et al., 2024).

3.6 METRICS

Case-level Rates. We evaluate whether a model turns an incomplete or inconsistent scientific request into a correct final task specification. For each case i, we compute three binary checks. **1) Resolution** (R_i): a binary metric that takes a value of 1 when all annotated issues of the case are resolved in the final specification produced by the agent, and 0 otherwise; **2) Conversational Grounding** (G_i): a binary metric that takes a value of 1 when every annotated issue of the case is grounded in the dialogue, and 0 otherwise; and **3) Intent Fidelity** (I_i): a binary metric that takes a value of 1 when the final specification preserves the user's intended scientific task, and 0 otherwise. We combine these three checks into the case-level rates. **a) Final Resolution Rate** (FRR ↑): FRR measures the fraction of cases where the final specification is correct ($R_i = 1$) and intent-faithful ($I_i = 1$), regardless of whether the model reached that specification through dialogue (G_i may be either 0 or 1). Higher FRR means that the model resolves more cases in its final output. **b) Conversation-Grounded Resolution Rate** (CGRR ↑): CGRR measures the fraction of cases where the final specification is correct ($R_i = 1$), intent-faithful ($I_i = 1$), and grounded in the conversation ($G_i = 1$). Higher CGRR means that the model resolves more cases through explicit clarification rather than silent guessing. **c) Silent Resolution Rate** (SRR ↓): SRR measures the fraction of cases where the final specification is correct ($R_i = 1$) and intent-faithful ($I_i = 1$), but not grounded in the conversation ($G_i = 0$). These cases correspond to implicit assumptions or silent repairs. Lower SRR is better because it means fewer cases are resolved without being made explicit to the user.

Component-level Rates. Unlike the case-level rates above, which credit a case only when all of its issues are resolved, component-level rates score each annotated issue separately. This gives a direct readout of which ontology components in Equation (1) are resolved, grounded, or silently repaired. For issue j in case i, let $k_{ij} \in y^\star$ be its ontology component. Let $r_{ij} = 1$ if the issue is resolved in the final specification, and let $g_{ij} = 1$ if the issue is clarified with the user; otherwise these values are 0. We also require the final specification to preserve the user's intent, i.e., $I_i = 1$. Let $\mathcal{I}_k = \{(i, j) : k_{ij} = k\}$ denote all issues belonging to component k. We define component-level **1) Final Resolution Rate** (FRR(k) ↑): the fraction of issues in component k that are resolved in an intent-faithful final specification. **2) Conversation-Grounded Resolution Rate** (CGRR(k) ↑): the fraction of issues in component k that are both resolved and clarified with the user. **3) Silent Resolution Rate** (SRR(k) ↓): the fraction of issues in component k that are resolved but not clarified with the user.

Capability, Robustness, and Usability. CGRR gives the primary success criterion, but it does not explain why a model succeeds or fails. We therefore also report three diagnostic axes for the Pareto analysis: *Capability*, *Robustness*, and *Usability*. Each axis is computed as an equally weighted average of lower-level diagnostic metrics. *Capability* measures whether the model asks the right clarification questions and produces a complete final specification; it averages *clarification recall*, the fraction of annotated issues surfaced by the model, *clarification precision*, the fraction of the model's questions that target annotated issues, and *plan completeness*, the fraction of required fields correctly instantiated in the final specification. *Robustness* measures whether the model avoids unreliable dialogue behavior; it averages *assumption avoidance*, *error detection*, and *memory consistency*, capturing silent assumptions, silent repairs, and contradictions with information established during the dialogue. *Usability* measures whether the final specification remains aligned with the user's intended scientific task, using *intent capture*. These axes are used only as diagnostic summaries; the main success metric remains CGRR. Full definitions and aggregation details are provided in Section D.

4 EXPERIMENTAL SETUP

Conversational evaluation framework. Each instance is evaluated as a multi-turn interaction between a conversational assistant model and a user. The assistant receives an ambiguous or inconsistent request, may ask one clarification question per turn, and produces a final specification either after explicit finalization or at the turn-budget cap. In the primary *guided* setting, the assistant is instructed to act as a requirements analyst that identifies missing information, detects contradictions, and clarifies one issue at a time before finalizing. We evaluate eight guided-mode models across all four domains: CLAUDE SONNET 4.6(Anthropic, 2026), GEMINI 2.5 PRO(Comanici et al., 2025), GEMINI 2.5 FLASH(Comanici et al., 2025), GPT-5.2(OpenAI, 2025), QWEN3-235B(Yang et al., 2025), DEEPSEEK-V3.2(DeepSeek-AI et al., 2025), KIMI K2.5(Team et al., 2026) and GPT-OSS-120B(OpenAI et al., 2025). Token usage per model and cost are detailed in Section F. We also

evaluate an *unguided* control without explicit instructions to identify inconsistencies or ambiguities, since prior work shows that LLMs often fail to detect such issues unless prompted to do so (Tyen et al., 2024); the guided-vs.-unguided comparison for GEMINI 2.5 PRO is reported in Section E.4. The user answering the question is a simulator LLM. The user simulator has access to the incomplete request, a hidden reference request specification, dialogue history, and it is strictly instructed to answer only from the reference request specification, using "make a reasonable assumption" when the requested detail is absent in the reference. In this way the simulator LLM abstains from providing information from its own knowledge and strictly adheres to the provided reference specification which is ambiguity or inconsistency free. This user simulator uses CLAUDE SONNET 4.6 as the model, unless specified otherwise. Our use of LLM simulating user follows recent conversational-agent benchmarks using LLM user proxies (Yao et al., 2024; Barres et al., 2025; Hathidara et al., 2026; Sekulic et al., 2024; Ni et al., 2026). Prompts used in assistant and simulator are provided in Section C.

Judge and human validation. All metrics are computed from saved conversations and final specifications using the rubric-based judging protocol in Section 3, which evaluates semantic resolution, conversational grounding, and intent preservation rather than exact string match. Unless otherwise noted, headline numbers use GEMINI 2.5 PRO as the judge. To assess judge dependence, we rescore stratified subsets with alternative judges and compare automatic judgments against an 80-case human-annotated subset balanced by domain, task type, and outcome bucket. Three expert annotators independently score the headline labels on this subset using the same per-case rubric as the LLM judge, blinded to the LLM-judge scores. Inter-annotator agreement and further details are reported in Section F.4 and Section E.7.

Ablations on user simulator and agent prompt. Since both simulator choice and prompt wording can shift measured rates (Dou et al., 2025; Shim et al., 2026), we report two ablations alongside the headline results. The *simulator ablation* re-runs the 80-case stratified subset with the assistant fixed at GEMINI 2.5 PRO and the user simulator varied across GEMINI 2.5 PRO, GPT-5.2, and CLAUDE SONNET 4.6. The *prompt-paraphrase ablation* fixes both the assistant at GEMINI 2.5 PRO and replaces the guided-mode assistant prompt with two scientifically equivalent paraphrases that preserve the four behavioral contracts but vary role framing and instruction wording. Further details can be found in Sections E.6 and E.8.

5 RESULTS

5.1 MODELS, GAPS, AND THE FRR–CGRR DECOMPOSITION

Figure 4 decomposes Final Resolution Rate (FRR) into Conversation-Grounded Resolution Rate (CGRR) and Silent Resolution Rate (SRR), with full per-domain metrics deferred to Tables 3 and 4. Every model exhibits a non-zero FRR – CGRR gap across domains and task types, averaging 12.2 percentage points on Disambiguation and 18.4 percentage points on Inconsistency Resolution (across the eight evaluated LLMs). The leading models also differ by task: GPT-5.2 is strongest on disambiguation, while GEMINI 2.5 PRO is strongest on inconsistency resolution. This suggests that eliciting missing information and explicitly identifying contradictions are related but not identical capabilities. Fluid mechanics is consistently difficult for disambiguation. Inconsistency resolution shows a different pattern: the strongest models handle solids and PDE conflicts more reliably, while materials remains harder. Additional results can be found in Section E.

Ontology patterns. Figure 5 reports component-level $\mathrm{FRR}(k)$, $\mathrm{CGRR}(k)$, $\mathrm{SRR}(k)$ as defined in Component-level Metrics in Section 3.6, across the scientific components defined in Equation (1). Numerics and solver choices are the most fragile component on disambiguation, with the lowest component-wise FRR in the benchmark; the ordering differs on inconsistency resolution, where material and physical properties are weakest instead.

Pareto view. Figure 6 separates performance into Capability, Robustness, and Usability as defined in Section 3.6. GPT-5.2 has the strongest disambiguation profile, whereas GEMINI 2.5 PRO is more robust on inconsistency resolution. Across models, Robustness is the least stable axis, especially when moving from missing-information cases to planted conflicts.

Robustness across judges, prompts, and simulators. Table 2 summarizes robustness checks on the 80-case subset. The conclusions are stable across judge choice (average of three human judges used to compute the agreement score with LLM judge), guided-prompt paraphrase, and user-simulator model. The three human annotators agree substantially with one another, and all three LLM judges

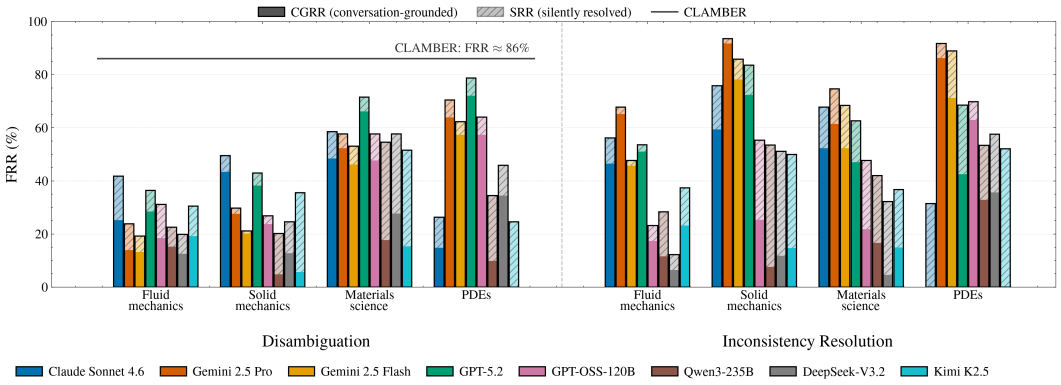

Figure 4: Case level resolution rate (Section 3.6) comparison among different models for the different domains and tasks in SCICONVBENCH. FRR is further broken down into CGRR and SRR. The dotted horizontal line in the disambiguation block marks our adapted CLAMBER's reported FRR on general disambiguation tasks ($\approx 86\%$) (Zhang et al., 2024c). SCICONVBENCH is more challenging than general domain disambiguation benchmark like CLAMBER.

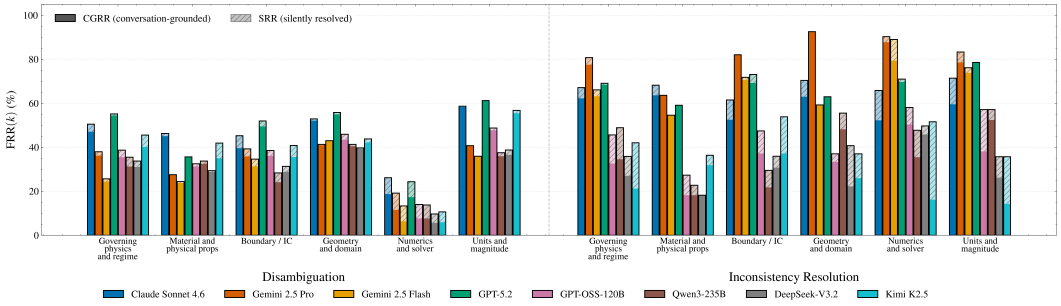

Figure 5: Component level resolution rate (Section 3.6) comparison among different models for the different domains and tasks in SCICONVBENCH as defined in Equation (1). FRR is further broken down into CGRR and SRR.

agree with them at a similar level on the headline metrics. Further details can be found in Sections E.6 to E.8.

(a) Judge-human agreement (%)

Judge	FRR	CGRR
GEMINI 2.5 PRO	84.2	83.8
GPT-5.2	87.5	83.8
SONNET 4.6	88.3	86.2

(b) Prompt paraphrases (%)

Variant	FRR	CGRR
Original	57.5	37.5
Variant A	53.8	31.2
Variant B	58.8	38.8

(c) User simulators (%)

Simulator	FRR	CGRR
GEMINI 2.5 PRO	55.0	27.5
GPT-5.2	62.5	37.5
SONNET 4.6	57.5	37.5

Table 2: Robustness checks on the 80-case stratified subset. Judge LLM choice, prompt paraphrase, and user-simulator LLM variation produce the same qualitative conclusions.

6 DISCUSSION

No single model dominates both clarification regimes. GPT-5.2 is strongest on disambiguation, at 51.2% CGRR compared with 39.4% for GEMINI 2.5 PRO, averaged across the four domains. However, the ordering reverses for inconsistency resolution: GEMINI 2.5 PRO reaches 76.2%, ahead of GEMINI 2.5 FLASH at 61.9% and GPT-5.2 at 53.2%. Thus, clarification ability is not a single scalar capability; eliciting missing information and confronting contradictions are separable scientific dialogue skills. Qualitative examples of grounded and silent resolutions are provided in Section F.3.

Ontology-level structure explains where clarification breaks down. Fluid mechanics is the clearest disambiguation bottleneck: the best fluid disambiguation CGRR is only 28.5%, while the best materials science and PDE disambiguation scores reach 66.2% and 72.1%, respectively. The

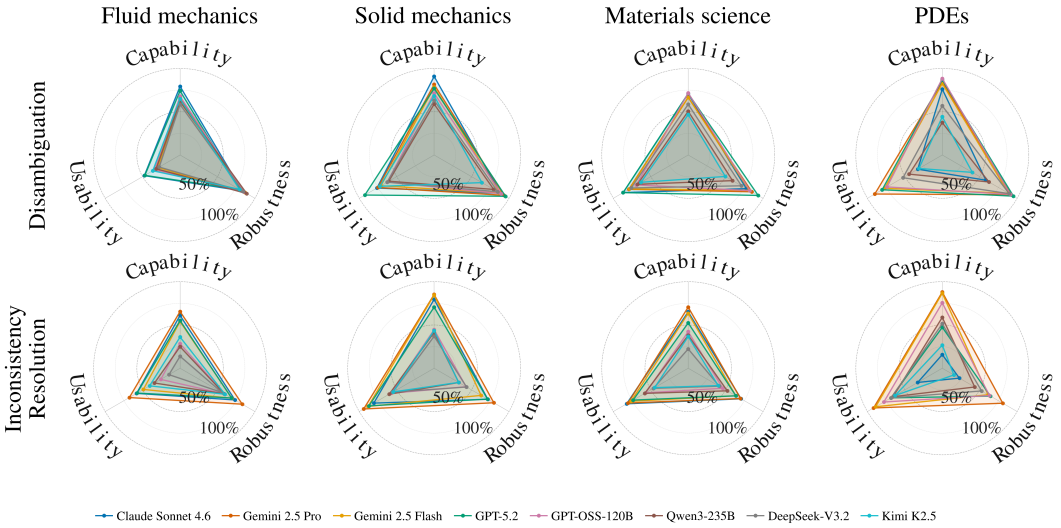

Figure 6: Pareto analysis across Capability, Robustness, and Usability. Top row: disambiguation. Bottom row: inconsistency resolution. Each panel is one domain; each trace is one model. Higher values are better.

ontology analysis helps explain this gap. Numerics and solver choices are the weakest component on disambiguation by a wide margin, with component-level FRR between 9.7% and 26.1% across models. These are not peripheral details; they determine what scientific problem is being solved. The benchmark therefore exposes failures in eliciting the commitments that make a scientific task well posed.

The Pareto analysis shows why no model should be treated as uniformly reliable. Many traces are lopsided: models often preserve usability and the user's broad intent while losing robustness. For example, GPT-5.2 is highly robust on disambiguation across all four domains, but its robustness contracts noticeably on inconsistency resolution. GEMINI 2.5 PRO also contracts relative to disambiguation, but remains the most balanced inconsistency model, combining high CGRR with strong capability and usability. This suggests that current models can often produce usable scientific specifications, but still fail to reliably surface conflicts or missing assumptions before finalizing the task.

The FRR/CGRR gap should be interpreted as silent scientific inference. In many cases, the final response contains a plausible repair or default that was never explicitly elicited from the user. The clearest examples are PDE inconsistency, where CLAUDE SONNET 4.6 reaches FRR of 31.5% and KIMI K2.5 FRR of 52.1%, yet both record CGRR of 0.0%: every successful repair in that slice occurs without a single clarifying question. The effect is not limited to one domain or one model, and QWEN3-235B silently resolves 45.9% of solid mechanics inconsistency cases, the largest silent resolution rate we observe. This behavior can look intelligent: the model fills in a solver choice, physical regime, boundary convention, or material assumption. But because the assumption is not asked about or acknowledged, it cannot be audited. In scientific workflows, this is a reproducibility risk rather than merely an interaction flaw.

7 CONCLUSION

We introduced SCICONVBENCH, a benchmark for conversational scientific task formulation across fluid mechanics, solid mechanics, materials science, and PDEs. Unlike scientific LLM benchmarks that grade models after the task is specified, SCICONVBENCH tests whether models surface missing or inconsistent scientific requirements through dialogue before committing to a final specification. Across eight guided models, final resolution rate exceeds conversation-grounded resolution rate in every domain on both tasks, and the difference is stable across judges, prompt paraphrases, and simulators. Case-level evidence narrows the diagnosis: on the same planted issues, leading models reach the same correct specification but differ in whether they flag the conflict, so silent resolution is a confrontation failure, not a knowledge failure. By making this distinction measurable, SCICONVBENCH establishes conversational task formulation as a necessary upstream capability for reliable scientific assistants.

AI USE STATEMENT

We used large language models during benchmark construction to suggest candidate omissions and conflicting requirements for the disambiguation and inconsistency-resolution cases. We also used them to flag potential answer leakage, unintended errors, and formatting problems in draft cases. The authors selected and revised the final prompt transformations, and the cases underwent the expert review described in the paper. We additionally used large language models to suggest grammatical edits to the manuscript and to assist in writing code for the figures. AI-assisted work was manually reviewed before inclusion.

REFERENCES

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pp. 475–484, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729. doi: 10.1145/3331184.3331265. URL https://doi.org/10.1145/3331184.3331265.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq), 2020. URL https://arxiv.org/abs/2009.11352.

Anthropic. Claude sonnet 4.6 system card. https://www.anthropic.com/claude-sonnet-4-6-system-card, February 2026. System card, February 17, 2026.

Lisa P. Argyle, E. Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31:337 – 351, 2022. URL https://api.semanticscholar.org/CorpusID:252280474.

Neil Ashton, Johannes Brandstetter, and Siddhartha Mishra. Fluid intelligence: A forward look on ai foundation models in computational fluid dynamics, 2025. URL https://arxiv.org/abs/2511.20455.

D.R. Askeland, P.P. Fulay, and W.J. Wright. *The Science and Engineering of Materials, SI Edition*. Cengage Learning, 2011. ISBN 9780495668022. URL https://books.google.com/books?id=fJanis-q6GkC.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.401. URL https://aclanthology.org/2024.acl-long.401/.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment. 2025. URL https://arxiv.org/abs/2506.07982.

F. Beer, E.R. Jr. Johnston, J. DeWolf, and D. Mazurek. *Mechanics of Materials*. McGraw-Hill Education, 2011. ISBN 9780073380285. URL https://books.google.com/books?id=g635kQAACAAJ.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. 2023. URL https://arxiv.org/abs/2304.05376.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In Ellen Riloff, David Chiang, Julia Hockenmaier, and

Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL https://aclanthology.org/D18-1547/.

W.D. Callister and D.G. Rethwisch. *Materials Science and Engineering: An Introduction, 9th Edition: Ninth Edition*. John Wiley and Sons, Incorporated, 2013. ISBN 9781118476543. URL https://books.google.com/books?id=TmxbAgAAQBAJ.

Yuxuan Chen, Xu Zhu, Hua Zhou, and Zhuyin Ren. Metaopenfoam: an llm-based multi-agent framework for cfd. 2024. URL https://arxiv.org/abs/2407.21320.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. 2025. URL https://arxiv.org/abs/2410.05080.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell, Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Alvin Abdagic, Lior Belenki, James Allingham, Anima Singh, Theo Guidroz, Srivatsan Srinivasan, Herman Schmit, Kristen Chiafullo, Andre Elisseeff, Nilpa Jha, Prateek Kolhar, Leonard Berrada, Frank Ding, Xiance Si, Shrestha Basu Mallick, Franz Och, Sofia Erell, Eric Ni, Tejasi Latkar, Sherry Yang, Petar Sirkovic, Ziqiang Feng, Robert Leland, Rachel Hornung, Gang Wu, Charles Blundell, Hamidreza Alvari, Po-Sen Huang, Cathy Yip, Sanja Deur, Li Liu, Gabriela Surita, Pablo Duque, Dima Damen, Johnson Jia, Arthur Guez, Markus Mircea, Animesh Sinha, Alberto Magni, Paweł Stradomski, Tal Marian, Vlado Galić, Wenhu Chen, Hisham Husain, Achintya Singhal, Dominik Grewe, François-Xavier Aubet, Shuang Song, Lorenzo Blanco, Leland Rechis, Lewis Ho, Rich Munoz, Kelvin Zheng, Jessica Hamrick, Kevin Mather, Hagai Taitelbaum, Eliza Rutherford, Yun Lei, Kuangyuan Chen, Anand Shukla, Erica Moreira, Eric Doi, Berivan Isik, Nir Shabat, Dominika Rogozińska, Kashyap Kolipaka, Jason Chang, Eugen Vušak, Srinivasan Venkatachary, Shadi Noghabi, Tarun Bharti, Younghoon Jun, Aleksandr Zaks, Simon Green, Jeshwanth Challagundla, William Wong, Muqthar Mohammad, Dean Hirsch, Yong Cheng, Iftekhar Naim, Lev Proleev, Damien Vincent, Aayush Singh, Maxim Krikun, Dilip Krishnan, Zoubin Ghahramani, Aviel Atias, Rajeev Aggarwal, Christo Kirov, Dimitrios Vytiniotis, Christy Koh, Alexandra Chronopoulou, Pawan Dogra, Vlad-Doru Ion, Gladys Tyen, Jason Lee, Felix Weissenberger, Trevor Strohman, Ashwin Balakrishna, Jack Rae, Marko Velic, Raoul de Liedekerke, Oded Elyada, Wentao Yuan, Canoee Liu, Lior Shani, Sergey Kishchenko, Bea Alessio, Yandong Li, Richard Song, Sam Kwei, Orion Jankowski, Aneesh Pappu, Youhei Namiki, Yenai Ma, Nilesh Tripuraneni, Colin Cherry, Marissa Ikonomidis, Yu-Cheng Ling, Colin Ji, Beka Westberg, Auriel Wright, Da Yu, David Parkinson, Swaroop Ramaswamy, Jerome Connor, Soheil Hassas Yeganeh, Snchit Grover, George Kenwright, Lubo Litchev, Chris Apps, Alex Tomala, Felix Halim, Alex Castro-Ros, Zefei Li, Anudhyan Boral, Pauline Sho, Michal Yarom, Eric Malmi, David Klinghoffer, Rebecca Lin, Alan Ansell, Pradeep Kumar S, Shubin Zhao, Siqi Zuo, Adam Santoro, Heng-Tze Cheng, Solomon Demmessie, Yuchi Liu, Nicole Brichtova, Allie Culp, Nathaniel Braun, Dan Graur, Will Ng, Nikhil Mehta, Aaron Phillips, Patrik Sundberg, Varun Godbole, Fangyu Liu, Yash Katariya, David Rim, Mojtaba Seyedhosseini, Sean Ammirati, Jonas Valfridsson, Mahan Malihi, Timothy Knight, Andeep Toor, Thomas Lampe, Abe Ittycheriah, Lewis Chiang, Chak Yeung, Alexandre Fréchette, Jinmeng Rao, Huisheng Wang, Himanshu Srivastava, Richard Zhang, Rocky

Rhodes, Ariel Brand, Dean Weesner, Ilya Figotin, Felix Gimeno, Rachana Fellinger, Pierre Marcenac, José Leal, Eyal Marcus, Victor Cotruta, Rodrigo Cabrera, Sheryl Luo, Dan Garrette, Vera Axelrod, Sorin Baltateanu, David Barker, Dongkai Chen, Horia Toma, Ben Ingram, Jason Riesa, Chinmay Kulkarni, Yujing Zhang, Hongbin Liu, Chao Wang, Martin Polacek, Will Wu, Kai Hui, Adrian N Reyes, Yi Su, Megan Barnes, Ishaan Malhi, Anfal Siddiqui, Qixuan Feng, Mihai Damaschin, Daniele Pighin, Andreas Steiner, Samuel Yang, Ramya Sree Boppana, Simeon Ivanov, Arun Kandoor, Aditya Shah, Asier Mujika, Da Huang, Christopher A. Choquette-Choo, Mohak Patel, Tianhe Yu, Toni Creswell, Jerry, Liu, Catarina Barros, Yasaman Razeghi, Aurko Roy, Phil Culliton, Binbin Xiong, Jiaqi Pan, Thomas Strohmann, Tolly Powell, Babi Seal, Doug DeCarlo, Pranav Shyam, Kaan Katircioglu, Xuezhi Wang, Cassidy Hardin, Immanuel Odisho, Josef Broder, Oscar Chang, Arun Nair, Artem Shtefan, Maura O'Brien, Manu Agarwal, Sahitya Potluri, Siddharth Goyal, Amit Jhindal, Saksham Thakur, Yury Stuken, James Lyon, Kristina Toutanova, Fangxiaoyu Feng, Austin Wu, Ben Horn, Alek Wang, Alex Cullum, Gabe Taubman, Disha Shrivastava, Chongyang Shi, Hamish Tomlinson, Roma Patel, Tao Tu, Ada Maksutaj Oflazer, Francesco Pongetti, Mingyao Yang, Adrien Ali Taïga, Vincent Perot, Nuo Wang Pierse, Feng Han, Yoel Drori, Iñaki Iturrate, Ayan Chakrabarti, Legg Yeung, Dave Dopson, Yi ting Chen, Apoorv Kulshreshtha, Tongfei Guo, Philip Pham, Tal Schuster, Junquan Chen, Alex Polozov, Jinwei Xing, Huanjie Zhou, Praneeth Kacham, Doron Kukliansky, Antoine Miech, Sergey Yaroshenko, Ed Chi, Sholto Douglas, Hongliang Fei, Mathieu Blondel, Preethi Myla, Lior Madmoni, Xing Wu, Daniel Keysers, Kristian Kjems, Isabela Albuquerque, Lijun Yu, Joel D'sa, Michelle Plantan, Vlad Ionescu, Jaume Sanchez Elias, Abhirut Gupta, Manish Reddy Vuyyuru, Fred Alcober, Tong Zhou, Kaiyang Ji, Florian Hartmann, Subha Puttagunta, Hugo Song, Ehsan Amid, Anca Stefanoiu, Andrew Lee, Paul Pucciarelli, Emma Wang, Amit Raul, Slav Petrov, Isaac Tian, Valentin Anklin, Nana Nti, Victor Gomes, Max Schumacher, Grace Vesom, Alex Panagopoulos, Konstantinos Bousmalis, Daniel Andor, Josh Jacob, Yuan Zhang, Bill Rosgen, Matija Kecman, Matthew Tung, Alexandra Belias, Noah Goodman, Paul Covington, Brian Wieder, Nikita Saxena, Elnaz Davoodi, Muhuan Huang, Sharath Maddineni, Vincent Roulet, Folawiyo Campbell-Ajala, Pier Giuseppe Sessa, Xintian, Wu, Guangda Lai, Paul Collins, Alex Haig, Vytenis Sakenas, Xiaowei Xu, Marissa Giustina, Laurent El Shafey, Pichi Charoenpanit, Shefali Garg, Joshua Ainslie, Boone Severson, Montse Gonzalez Arenas, Shreya Pathak, Sujee Rajayogam, Jie Feng, Michiel Bakker, Sheng Li, Nevan Wichers, Jamie Rogers, Xinyang Geng, Yeqing Li, Rolf Jagerman, Chao Jia, Nadav Olmert, David Sharon, Matthew Mauger, Sandeep Mariserla, Hongxu Ma, Megha Mohabey, Kyuyeun Kim, Alek Andreev, Scott Pollom, Juliette Love, Vihan Jain, Priyanka Agrawal, Yannick Schroecker, Alisa Fortin, Manfred Warmuth, Ji Liu, Andrew Leach, Irina Blok, Ganesh Poomal Girirajan, Roee Aharoni, Benigno Uria, Andrei Sozanschi, Dan Goldberg, Lucian Ionita, Marco Tulio Ribeiro, Martin Zlocha, Vighnesh Birodkar, Sami Lachgar, Liangzhe Yuan, Himadri Choudhury, Matt Ginsberg, Fei Zheng, Gregory Dibb, Emily Graves, Swachhand Lokhande, Gabriel Rasskin, George-Cristian Muraru, Corbin Quick, Sandeep Tata, Pierre Sermanet, Aditya Chawla, Itay Karo, Yan Wang, Susan Zhang, Orgad Keller, Anca Dragan, Guolong Su, Ian Chou, Xi Liu, Yiqing Tao, Shruthi Prabhakara, Marc Wilson, Ruibo Liu, Shibo Wang, Georgie Evans, David Du, Alfonso Castaño, Gautam Prasad, Mona El Mahdy, Sebastian Gerlach, Machel Reid, Jarrod Kahn, Amir Zait, Thanumalayan Sankaranarayana Pillai, Thatcher Ulrich, Guanyu Wang, Jan Wassenberg, Efrat Farkash, Kiran Yalasangi, Congchao Wang, Maria Bauza, Simon Bucher, Ting Liu, Jun Yan, Gary Leung, Vikas Sindhwani, Parker Barnes, Avi Singh, Ivan Jurin, Jichuan Chang, Niket Kumar Bhumihar, Sivan Eiger, Gui Citovsky, Ben Withbroe, Zhang Li, Siyang Xue, Niccolò Dal Santo, Georgi Stoyanov, Yves Raimond, Steven Zheng, Yilin Gao, Vít Listík, Sławek Kwasiborski, Rachel Saputro, Adnan Ozturel, Ganesh Mallya, Kushal Majmundar, Ross West, Paul Caron, Jinliang Wei, Lluis Castrejon, Sharad Vikram, Deepak Ramachandran, Nikhil Dhawan, Jiho Park, Sara Smoot, George van den Driessche, Yochai Blau, Chase Malik, Wei Liang, Roy Hirsch, Cicero Nogueira dos Santos, Eugene Weinstein, Aäron van den Oord, Sid Lall, Nicholas FitzGerald, Zixuan Jiang, Xuan Yang, Dale Webster, Ali Elqursh, Aedan Pope, Georges Rotival, David Raposo, Wanzheng Zhu, Jeff Dean, Sami Alabed, Dustin Tran, Arushi Gupta, Zach Gleicher, Jessica Austin, Edouard Rosseel, Megh Umekar, Dipanjan Das, Yinghao Sun, Kai Chen, Karolis Misiunas, Xiang Zhou, Yixian Di, Alyssa Loo, Josh Newlan, Bo Li, Vinay Ramasesh, Ying Xu, Alex Chen, Sudeep Gandhe, Radu Soricut, Nikita Gupta, Shuguang Hu, Seliem El-Sayed, Xavier Garcia, Idan Brusilovsky, Pu-Chin Chen, Andrew Bolt, Lu Huang, Alex Gurney, Zhiying Zhang, Alexander Pritzel, Jarek Wilkiewicz, Bryan Seybold, Bhargav Kanagal Shamanna, Felix Fischer, Josef Dean, Karan Gill, Ross Mcilroy, Abhishek Bhowmick, Jeremy Selier, Antoine Yang, Derek Cheng, Vladimir Magay, Jie Tan, Dhriti Varma, Christian Walder, Tomas Kocisky, Ryo Nakashima, Paul Natsev, Mike Kwong, Ionel

Gog, Chiyuan Zhang, Sander Dieleman, Thomas Jimma, Andrey Ryabtsev, Siddhartha Brahma, David Steiner, Dayou Du, Ante Žužul, Mislav Žanić, Mukund Raghavachari, Willi Gierke, Zeyu Zheng, Dessie Petrova, Yann Dauphin, Yuchuan Liu, Ido Kessler, Steven Hand, Chris Duvarney, Seokhwan Kim, Hyo Lee, Léonard Hussenot, Jeffrey Hui, Josh Smith, Deepali Jain, Jiawei Xia, Gaurav Singh Tomar, Keyvan Amiri, Du Phan, Fabian Fuchs, Tobias Weyand, Nenad Tomasev, Alexandra Cordell, Xin Liu, Jonathan Mallinson, Pankaj Joshi, Andy Crawford, Arun Suggala, Steve Chien, Nick Fernando, Mariella Sanchez-Vargas, Duncan Williams, Phil Crone, Xiyang Luo, Igor Karpov, Jyn Shan, Terry Thurk, Robin Strudel, Paul Voigtlaender, Piyush Patil, Tim Dozat, Ali Khodaei, Sahil Singla, Piotr Ambroszczyk, Qiyin Wu, Yifan Chang, Brian Roark, Chaitra Hegde, Tianli Ding, Angelos Filos, Zhongru Wu, André Susano Pinto, Shuang Liu, Saarthak Khanna, Aditya Pandey, Siobhan Mcloughlin, Qiujia Li, Sam Haves, Allan Zhou, Elena Buchatskaya, Isabel Leal, Peter de Boursac, Nami Akazawa, Nina Anderson, Terry Chen, Krishna Somandepalli, Chen Liang, Sheela Goenka, Stephanie Winkler, Alexander Grushetsky, Yifan Ding, Jamie Smith, Fan Ye, Jordi Pont-Tuset, Eric Li, Ruichao Li, Tomer Golany, Dawid Wegner, Tao Jiang, Omer Barak, Yuan Shangguan, Eszter Vértes, Renee Wong, Jörg Bornschein, Alex Tudor, Michele Bevilacqua, Tom Schaul, Ankit Singh Rawat, Yang Zhao, Kyriakos Axiotis, Lei Meng, Cory McLean, Jonathan Lai, Jennifer Beattie, Nate Kushman, Yaxin Liu, Blair Kutzman, Fiona Lang, Jingchen Ye, Praneeth Netrapalli, Pushkar Mishra, Myriam Khan, Megha Goel, Rob Willoughby, David Tian, Honglei Zhuang, JD Chen, Zak Tsai, Tasos Kementsietsidis, Arjun Khare, James Keeling, Keyang Xu, Nathan Waters, Florent Altché, Ashok Popat, Bhavishya Mittal, David Saxton, Dalia El Badawy, Michael Mathieu, Zheng Zheng, Hao Zhou, Nishant Ranka, Richard Shin, Qingnan Duan, Tim Salimans, Ioana Mihailescu, Uri Shaham, Ming-Wei Chang, Yannis Assael, Nishanth Dikkala, Martin Izzard, Vincent Cohen-Addad, Cat Graves, Vlad Feinberg, Grace Chung, DJ Strouse, Danny Karmon, Sahand Sharifzadeh, Zoe Ashwood, Khiem Pham, Jon Blanton, Alex Vasiloff, Jarred Barber, Mark Geller, Aurick Zhou, Fedir Zubach, Tzu-Kuo Huang, Lei Zhang, Himanshu Gupta, Matt Young, Julia Proskurnia, Ronny Votel, Valentin Gabeur, Gabriel Barcik, Aditya Tripathi, Hongkun Yu, Geng Yan, Beer Changpinyo, Filip Pavetić, Amy Coyle, Yasuhisa Fujii, Jorge Gonzalez Mendez, Tianhao Zhou, Harish Rajamani, Blake Hechtman, Eddie Cao, Da-Cheng Juan, Yi-Xuan Tan, Valentin Dalibard, Yilun Du, Natalie Clay, Kaisheng Yao, Wenhao Jia, Dimple Vijaykumar, Yuxiang Zhou, Xinyi Bai, Wei-Chih Hung, Steven Pecht, Georgi Todorov, Nikhil Khadke, Pramod Gupta, Preethi Lahoti, Arnaud Autef, Karthik Duddu, James Lee-Thorp, Alexander Bykovsky, Tautvydas Misiunas, Sebastian Flennerhag, Santhosh Thangaraj, Jed McGiffin, Zack Nado, Markus Kunesch, Andreas Noever, Amir Hertz, Marco Liang, Victor Stone, Evan Palmer, Samira Daruki, Arijit Pramanik, Siim Põder, Austin Kyker, Mina Khan, Evgeny Sluzhaev, Marvin Ritter, Avraham Ruderman, Wenlei Zhou, Chirag Nagpal, Kiran Vodrahalli, George Necula, Paul Barham, Ellie Pavlick, Jay Hartford, Izhak Shafran, Long Zhao, Maciej Mikuła, Tom Eccles, Hidetoshi Shimokawa, Kanav Garg, Luke Vilnis, Hanwen Chen, Ilia Shumailov, Kuang-Huei Lee, Abdelrahman Abdelhamed, Meiyan Xie, Vered Cohen, Ester Hlavnova, Dan Malkin, Chawin Sitawarin, James Lottes, Pauline Coquinot, Tianli Yu, Sandeep Kumar, Jingwei Zhang, Aroma Mahendru, Zafarali Ahmed, James Martens, Tao Chen, Aviel Boag, Daiyi Peng, Coline Devin, Arseniy Klimovskiy, Mary Phuong, Danny Vainstein, Jin Xie, Bhuvana Ramabhadran, Nathan Howard, Xinxin Yu, Gitartha Goswami, Jingyu Cui, Sam Shleifer, Mario Pinto, Chih-Kuan Yeh, Ming-Hsuan Yang, Sara Javanmardi, Dan Ethier, Chace Lee, Jordi Orbay, Suyog Kotecha, Carla Bromberg, Pete Shaw, James Thornton, Adi Gerzi Rosenthal, Shane Gu, Matt Thomas, Ian Gemp, Aditya Ayyar, Asahi Ushio, Aarush Selvan, Joel Wee, Chenxi Liu, Maryam Majzoubi, Weiren Yu, Jake Abernethy, Tyler Liechty, Renke Pan, Hoang Nguyen, Qiong, Hu, Sarah Perrin, Abhinav Arora, Emily Pitler, Weiyi Wang, Kaushik Shivakumar, Flavien Prost, Ben Limonchik, Jing Wang, Yi Gao, Timothee Cour, Shyamal Buch, Huan Gui, Maria Ivanova, Philipp Neubeck, Kelvin Chan, Lucy Kim, Huizhong Chen, Naman Goyal, Da-Woon Chung, Lu Liu, Yao Su, Anastasia Petrushkina, Jiajun Shen, Armand Joulin, Yuanzhong Xu, Stein Xudong Lin, Yana Kulizhskaya, Ciprian Chelba, Shobha Vasudevan, Eli Collins, Vasilisa Bashlovkina, Tony Lu, Doug Fritz, Jongbin Park, Yanqi Zhou, Chen Su, Richard Tanburn, Mikhail Sushkov, Mitchelle Rasquinha, Jinning Li, Jennifer Prendki, Yiming Li, Pallavi LV, Shriya Sharma, Hen Fitoussi, Hui Huang, Andrew Dai, Phuong Dao, Mike Burrows, Henry Prior, Danfeng Qin, Golan Pundak, Lars Lowe Sjoesund, Art Khurshudov, Zhenkai Zhu, Albert Webson, Elizabeth Kemp, Tat Tan, Saurabh Agrawal, Susie Sargsyan, Liqun Cheng, Jim Stephan, Tom Kwiatkowski, David Reid, Arunkumar Byravan, Assaf Hurwitz Michaely, Nicolas Heess, Luowei Zhou, Sonam Goenka, Viral Carpenter, Anselm Levskaya, Bo Wang, Reed Roberts, Rémi Leblond, Sharat Chikkerur, Stav Ginzburg, Max Chang, Robert Riachi, Chuqiao, Xu, Zalán Borsos, Michael Pliskin, Julia Pawar,

Morgane Lustman, Hannah Kirkwood, Ankit Anand, Aditi Chaudhary, Norbert Kalb, Kieran Milan, Sean Augenstein, Anna Goldie, Laurel Prince, Karthik Raman, Yanhua Sun, Vivian Xia, Aaron Cohen, Zhouyuan Huo, Josh Camp, Seher Ellis, Lukas Zilka, David Vilar Torres, Lisa Patel, Sho Arora, Betty Chan, Jonas Adler, Kareem Ayoub, Jacky Liang, Fayaz Jamil, Jiepu Jiang, Simon Baumgartner, Haitian Sun, Yael Karov, Yaroslav Akulov, Hui Zheng, Irene Cai, Claudio Fantacci, James Rubin, Alex Rav Acha, Mengchao Wang, Nina D'Souza, Rohit Sathyanarayana, Shengyang Dai, Simon Rowe, Andrey Simanovsky, Omer Goldman, Yuheng Kuang, Xiaoyue Pan, Andrew Rosenberg, Tania Rojas-Esponda, Praneet Dutta, Amy Zeng, Irina Jurenka, Greg Farquhar, Yamini Bansal, Shariq Iqbal, Becca Roelofs, Ga-Young Joung, Parker Beak, Changwan Ryu, Ryan Poplin, Yan Wu, Jean-Baptiste Alayrac, Senaka Buthpitiya, Olaf Ronneberger, Caleb Habtegebriel, Wei Li, Paul Cavallaro, Aurora Wei, Guy Bensky, Timo Denk, Harish Ganapathy, Jeff Stanway, Pratik Joshi, Francesco Bertolini, Jessica Lo, Olivia Ma, Zachary Charles, Geta Sampemane, Himanshu Sahni, Xu Chen, Harry Askham, David Gaddy, Peter Young, Jiewen Tan, Matan Eyal, Arthur Bražinskas, Li Zhong, Zhichun Wu, Mark Epstein, Kai Bailey, Andrew Hard, Kamyu Lee, Sasha Goldshtein, Alex Ruiz, Mohammed Badawi, Matthias Lochbrunner, JK Kearns, Ashley Brown, Fabio Pardo, Theophane Weber, Haichuan Yang, Pan-Pan Jiang, Berkin Akin, Zhao Fu, Marcus Wainwright, Chi Zou, Meenu Gaba, Pierre-Antoine Manzagol, Wendy Kan, Yang Song, Karina Zainullina, Rui Lin, Jeongwoo Ko, Salil Deshmukh, Apoorv Jindal, James Svensson, Divya Tyam, Heri Zhao, Christine Kaeser-Chen, Scott Baird, Pooya Moradi, Jamie Hall, Qiuchen Guo, Vincent Tsang, Bowen Liang, Fernando Pereira, Suhas Ganesh, Ivan Korotkov, Jakub Adamek, Sridhar Thiagarajan, Vinh Tran, Charles Chen, Chris Tar, Sanil Jain, Ishita Dasgupta, Taylan Bilal, David Reitter, Kai Zhao, Giulia Vezzani, Yasmin Gehman, Pulkit Mehta, Lauren Beltrone, Xerxes Dotiwalla, Sergio Guadarrama, Zaheer Abbas, Stefani Karp, Petko Georgiev, Chun-Sung Ferng, Marc Brockschmidt, Liqian Peng, Christoph Hirnschall, Vikas Verma, Yingying Bi, Ying Xiao, Avigail Dabush, Kelvin Xu, Phil Wallis, Randall Parker, Qifei Wang, Yang Xu, Ilkin Safarli, Dinesh Tewari, Yin Zhang, Seungyeon Kim, Andrea Gesmundo, Mackenzie Thomas, Sergey Levi, Ahmed Chowdhury, Kanishka Rao, Peter Garst, Sam Conway-Rahman, Helen Ran, Kay McKinney, Zhisheng Xiao, Wenhao Yu, Rohan Agrawal, Axel Stjerngren, Catalin Ionescu, Jingjing Chen, Vivek Sharma, Justin Chiu, Fei Liu, Ken Franko, Clayton Sanford, Xingyu Cai, Paul Michel, Sanjay Ganapathy, Jane Labanowski, Zachary Garrett, Ben Vargas, Sean Sun, Bryan Gale, Thomas Buschmann, Guillaume Desjardins, Nimesh Ghelani, Palak Jain, Mudit Verma, Chulayuth Asawaroengchai, Julian Eisenschlos, Jitendra Harlalka, Hideto Kazawa, Don Metzler, Joshua Howland, Ying Jian, Jake Ades, Viral Shah, Tynan Gangwani, Seungji Lee, Roman Ring, Steven M. Hernandez, Dean Reich, Amer Sinha, Ashutosh Sathe, Joe Kovac, Ashleah Gill, Ajay Kannan, Andrea D'olimpio, Martin Sevenich, Jay Whang, Been Kim, Khe Chai Sim, Jilin Chen, Jiageng Zhang, Shuba Lall, Yossi Matias, Bill Jia, Abe Friesen, Sara Nasso, Ashish Thapliyal, Bryan Perozzi, Ting Yu, Anna Shekhawat, Safeen Huda, Peter Grabowski, Eric Wang, Ashwin Sreevatsa, Hilal Dib, Mehadi Hassen, Parker Schuh, Vedrana Milutinovic, Chris Welty, Michael Quinn, Ali Shah, Bangju Wang, Gabe Barth-Maron, Justin Frye, Natalie Axelsson, Tao Zhu, Yukun Ma, Irene Giannoumis, Hanie Sedghi, Chang Ye, Yi Luan, Kevin Aydin, Bilva Chandra, Vivek Sampathkumar, Ronny Huang, Victor Lavrenko, Ahmed Eleryan, Zhi Hong, Steven Hansen, Sara Mc Carthy, Bidisha Samanta, Domagoj Ćevid, Xin Wang, Fangtao Li, Michael Voznesensky, Matt Hoffman, Andreas Terzis, Vikash Sehwag, Gil Fidel, Luheng He, Mu Cai, Yanzhang He, Alex Feng, Martin Nikoltchev, Samrat Phatale, Jason Chase, Rory Lawton, Ming Zhang, Tom Ouyang, Manuel Tragut, Mehdi Hafezi Manshadi, Arjun Narayanan, Jiaming Shen, Xu Gao, Tolga Bolukbasi, Nick Roy, Xin Li, Daniel Golovin, Liviu Panait, Zhen Qin, Guangxing Han, Thomas Anthony, Sneha Kudugunta, Viorica Patraucean, Aniket Ray, Xinyun Chen, Xiaochen Yang, Tanuj Bhatia, Pranav Talluri, Alex Morris, Andrija Ražnatović, Bethanie Brownfield, James An, Sheng Peng, Patrick Kane, Ce Zheng, Nico Duduta, Joshua Kessinger, James Noraky, Siqi Liu, Keran Rong, Petar Veličković, Keith Rush, Alex Goldin, Fanny Wei, Shiva Mohan Reddy Garlapati, Caroline Pantofaru, Okwan Kwon, Jianmo Ni, Eric Noland, Julia Di Trapani, Françoise Beaufays, Abhijit Guha Roy, Yinlam Chow, Aybuke Turker, Geoffrey Cideron, Lantao Mei, Jon Clark, Qingyun Dou, Matko Bošnjak, Ralph Leith, Yuqing Du, Amir Yazdanbakhsh, Milad Nasr, Chester Kwak, Suraj Satishkumar Sheth, Alex Kaskasoli, Ankesh Anand, Balaji Lakshminarayanan, Sammy Jerome, David Bieber, Chun-Te Chu, Alexandre Senges, Tianxiao Shen, Mukund Sridhar, Ndaba Ndebele, Benjamin Beyret, Shakir Mohamed, Mia Chen, Markus Freitag, Jiaxian Guo, Luyang Liu, Paul Roit, Heng Chen, Shen Yan, Tom Stone, JD Co-Reyes, Jeremy Cole, Salvatore Scellato, Shekoofeh Azizi, Hadi Hashemi, Alicia Jin, Anand Iyer, Marcella Valentine, András György, Arun Ahuja, Daniel Hernandez Diaz, Chen-Yu Lee, Nathan Clement, Weize Kong, Drew Garmon, Ishaan Watts, Kush Bhatia, Khyatti Gupta, Matt

Miecnikowski, Hugo Vallet, Ankur Taly, Edward Loper, Saket Joshi, James Atwood, Jo Chick, Mark Collier, Fotis Iliopoulos, Ryan Trostle, Beliz Gunel, Ramiro Leal-Cavazos, Arnar Mar Hrafnkelsson, Michael Guzman, Xiaoen Ju, Andy Forbes, Jesse Emond, Kushal Chauhan, Ben Caine, Li Xiao, Wenjun Zeng, Alexandre Moufarek, Daniel Murphy, Maya Meng, Nitish Gupta, Felix Riedel, Anil Das, Elijah Lawal, Shashi Narayan, Tiberiu Sosea, James Swirhun, Linda Friso, Behnam Neyshabur, Jing Lu, Sertan Girgin, Michael Wunder, Edouard Yvinec, Aroonalok Pyne, Victor Carbune, Shruti Rijhwani, Yang Guo, Tulsee Doshi, Anton Briukhov, Max Bain, Ayal Hitron, Xuanhui Wang, Ashish Gupta, Ke Chen, Cosmo Du, Weiyang Zhang, Dhruv Shah, Arjun Akula, Max Dylla, Ashyana Kachra, Weicheng Kuo, Tingting Zou, Lily Wang, Luyao Xu, Jifan Zhu, Justin Snyder, Sachit Menon, Orhan Firat, Igor Mordatch, Yuan Yuan, Natalia Ponomareva, Rory Blevins, Lawrence Moore, Weijun Wang, Phil Chen, Martin Scholz, Artur Dwornik, Jason Lin, Sicheng Li, Diego Antognini, Te I, Xiaodan Song, Matt Miller, Uday Kalra, Adam Raveret, Oscar Akerlund, Felix Wu, Andrew Nystrom, Namrata Godbole, Tianqi Liu, Hannah DeBalsi, Jewel Zhao, Buhuang Liu, Avi Caciularu, Lauren Lax, Urvashi Khandelwal, Victoria Langston, Eric Bailey, Silvio Lattanzi, Yufei Wang, Neel Kovelamudi, Sneha Mondal, Guru Guruganesh, Nan Hua, Ofir Roval, Paweł Wesołowski, Rishikesh Ingale, Jonathan Halcrow, Tim Sohn, Christof Angermueller, Bahram Raad, Eli Stickgold, Eva Lu, Alec Kosik, Jing Xie, Timothy Lillicrap, Austin Huang, Lydia Lihui Zhang, Dominik Paulus, Clement Farabet, Alex Wertheim, Bing Wang, Rishabh Joshi, Chu ling Ko, Yonghui Wu, Shubham Agrawal, Lily Lin, XiangHai Sheng, Peter Sung, Tyler Breland-King, Christina Butterfield, Swapnil Gawde, Sumeet Singh, Qiao Zhang, Raj Apte, Shilpa Shetty, Adrian Hutter, Tao Li, Elizabeth Salesky, Federico Lebron, Jonni Kanerva, Michela Paganini, Arthur Nguyen, Rohith Vallu, Jan-Thorsten Peter, Sarmishta Velury, David Kao, Jay Hoover, Anna Bortsova, Colton Bishop, Shoshana Jakobovits, Alessandro Agostini, Alekh Agarwal, Chang Liu, Charles Kwong, Sasan Tavakkol, Ioana Bica, Alex Greve, Anirudh GP, Jake Marcus, Le Hou, Tom Duerig, Rivka Moroshko, Dave Lacey, Andy Davis, Julien Amelot, Guohui Wang, Frank Kim, Theofilos Strinopoulos, Hui Wan, Charline Le Lan, Shankar Krishnan, Haotian Tang, Peter Humphreys, Junwen Bai, Idan Heimlich Shtacher, Diego Machado, Chenxi Pang, Ken Burke, Dangyi Liu, Renga Aravamudhan, Yue Song, Ed Hirst, Abhimanyu Singh, Brendan Jou, Liang Bai, Francesco Piccinno, Chuyuan Kelly Fu, Robin Alazard, Barak Meiri, Daniel Winter, Charlie Chen, Mingda Zhang, Jens Heitkaemper, John Lambert, Jinhyuk Lee, Alexander Frömmgen, Sergey Rogulenko, Pranav Nair, Paul Niemczyk, Anton Bulyenov, Bibo Xu, Hadar Shemtov, Morteza Zadimoghaddam, Serge Toropov, Mateo Wirth, Hanjun Dai, Sreenivas Gollapudi, Daniel Zheng, Alex Kurakin, Chansoo Lee, Kalesha Bullard, Nicolas Serrano, Ivana Balazevic, Yang Li, Johan Schalkwyk, Mark Murphy, Mingyang Zhang, Kevin Sequeira, Romina Datta, Nishant Agrawal, Charles Sutton, Nithya Attaluri, Mencher Chiang, Wael Farhan, Gregory Thornton, Kate Lin, Travis Choma, Hung Nguyen, Kingshuk Dasgupta, Dirk Robinson, Iulia Comşa, Michael Riley, Arjun Pillai, Basil Mustafa, Ben Golan, Amir Zandieh, Jean-Baptiste Lespiau, Billy Porter, David Ross, Sujeevan Rajayogam, Mohit Agarwal, Subhashini Venugopalan, Bobak Shahriari, Qiqi Yan, Hao Xu, Taylor Tobin, Pavel Dubov, Hongzhi Shi, Adrià Recasens, Anton Kovsharov, Sebastian Borgeaud, Lucio Dery, Shanthal Vasanth, Elena Gribovskaya, Linhai Qiu, Mahdis Mahdieh, Wojtek Skut, Elizabeth Nielsen, CJ Zheng, Adams Yu, Carrie Grimes Bostock, Shaleen Gupta, Aaron Archer, Chris Rawles, Elinor Davies, Alexey Svyatkovskiy, Tomy Tsai, Yoni Halpern, Christian Reisswig, Bartek Wydrowski, Bo Chang, Joan Puigcerver, Mor Hazan Taege, Jian Li, Eva Schnider, Xinjian Li, Dragos Dena, Yunhan Xu, Umesh Telang, Tianze Shi, Heiga Zen, Kyle Kastner, Yeongil Ko, Neesha Subramaniam, Aviral Kumar, Pete Blois, Zhuyun Dai, John Wieting, Yifeng Lu, Yoel Zeldes, Tian Xie, Anja Hauth, Alexandru Ţifrea, Yuqi Li, Sam El-Husseini, Dan Abolafia, Howard Zhou, Wen Ding, Sahra Ghalebikesabi, Carlos Guía, Andrii Maksai, Ágoston Weisz, Sercan Arik, Nick Sukhanov, Aga Świetlik, Xuhui Jia, Luo Yu, Weiyue Wang, Mark Brand, Dawn Bloxwich, Sean Kirmani, Zhe Chen, Alec Go, Pablo Sprechmann, Nithish Kannen, Alen Carin, Paramjit Sandhu, Isabel Edkins, Leslie Nooteboom, Jai Gupta, Loren Maggiore, Javad Azizi, Yael Pritch, Pengcheng Yin, Mansi Gupta, Danny Tarlow, Duncan Smith, Desi Ivanov, Mohammad Babaeizadeh, Ankita Goel, Satish Kambala, Grace Chu, Matej Kastelic, Michelle Liu, Hagen Soltau, Austin Stone, Shivani Agrawal, Min Kim, Kedar Soparkar, Srinivas Tadepalli, Oskar Bunyan, Rachel Soh, Arvind Kannan, DY Kim, Blake JianHang Chen, Afief Halumi, Sudeshna Roy, Yulong Wang, Olcan Sercinoglu, Gena Gibson, Sijal Bhatnagar, Motoki Sano, Daniel von Dincklage, Qingchun Ren, Blagoj Mitrevski, Mirek Olšák, Jennifer She, Carl Doersch, Jilei, Wang, Bingyuan Liu, Qijun Tan, Tamar Yakar, Tris Warkentin, Alex Ramirez, Carl Lebsack, Josh Dillon, Rajiv Mathews, Tom Cobley, Zelin Wu, Zhuoyuan Chen, Jon Simon, Swaroop Nath, Tara Sainath, Alexei Bendebury, Ryan Julian, Bharath Mankalale, Daria Ćurko,

Paulo Zacchello, Adam R. Brown, Kiranbir Sodhia, Heidi Howard, Sergi Caelles, Abhinav Gupta, Gareth Evans, Anna Bulanova, Lesley Katzen, Roman Goldenberg, Anton Tsitsulin, Joe Stanton, Benoit Schillings, Vitaly Kovalev, Corey Fry, Rushin Shah, Kuo Lin, Shyam Upadhyay, Cheng Li, Soroush Radpour, Marcello Maggioni, Jing Xiong, Lukas Haas, Jenny Brennan, Aishwarya Kamath, Nikolay Savinov, Arsha Nagrani, Trevor Yacovone, Ryan Kappedal, Kostas Andriopoulos, Li Lao, YaGuang Li, Grigory Rozhdestvenskiy, Kazuma Hashimoto, Andrew Audibert, Sophia Austin, Daniel Rodriguez, Anian Ruoss, Garrett Honke, Deep Karkhanis, Xi Xiong, Qing Wei, James Huang, Zhaoqi Leng, Vittal Premachandran, Stan Bileschi, Georgios Evangelopoulos, Thomas Mensink, Jay Pavagadhi, Denis Teplyashin, Paul Chang, Linting Xue, Garrett Tanzer, Sally Goldman, Kaushal Patel, Shixin Li, Jeremy Wiesner, Ivy Zheng, Ian Stewart-Binks, Jie Han, Zhi Li, Liangchen Luo, Karel Lenc, Mario Lučić, Fuzhao Xue, Ryan Mullins, Alexey Guseynov, Chung-Ching Chang, Isaac Galatzer-Levy, Adam Zhang, Garrett Bingham, Grace Hu, Ale Hartman, Yue Ma, Jordan Griffith, Alex Irpan, Carey Radebaugh, Summer Yue, Lijie Fan, Victor Ungureanu, Christina Sorokin, Hannah Teufel, Peiran Li, Rohan Anil, Dimitris Paparas, Todd Wang, Chu-Cheng Lin, Hui Peng, Megan Shum, Goran Petrovic, Demetra Brady, Richard Nguyen, Klaus Macherey, Zhihao Li, Harman Singh, Madhavi Yenugula, Mariko Iinuma, Xinyi Chen, Kavya Kopparapu, Alexey Stern, Shachi Dave, Chandu Thekkath, Florence Perot, Anurag Kumar, Fangda Li, Yang Xiao, Matthew Bilotti, Mohammad Hossein Bateni, Isaac Noble, Lisa Lee, Amelio Vázquez-Reina, Julian Salazar, Xiaomeng Yang, Boyu Wang, Ela Gruzewska, Anand Rao, Sindhu Raghuram, Zheng Xu, Eyal Ben-David, Jieru Mei, Sid Dalmia, Zhaoyi Zhang, Yuchen Liu, Gagan Bansal, Helena Pankov, Steven Schwarcz, Andrea Burns, Christine Chan, Sumit Sanghai, Ricky Liang, Ethan Liang, Antoine He, Amy Stuart, Arun Narayanan, Yukun Zhu, Christian Frank, Bahar Fatemi, Amit Sabne, Oran Lang, Indro Bhattacharya, Shane Settle, Maria Wang, Brendan McMahan, Andrea Tacchetti, Livio Baldini Soares, Majid Hadian, Serkan Cabi, Timothy Chung, Nikita Putikhin, Gang Li, Jeremy Chen, Austin Tarango, Henryk Michalewski, Mehran Kazemi, Hussain Masoom, Hila Sheftel, Rakesh Shivanna, Archita Vadali, Ramona Comanescu, Doug Reid, Joss Moore, Arvind Neelakantan, Michaël Sander, Jonathan Herzig, Aviv Rosenberg, Mostafa Dehghani, JD Choi, Michael Fink, Reid Hayes, Eric Ge, Shitao Weng, Chia-Hua Ho, John Karro, Kalpesh Krishna, Lam Nguyen Thiet, Amy Skerry-Ryan, Daniel Eppens, Marco Andreetto, Navin Sarma, Silvano Bonacina, Burcu Karagol Ayan, Megha Nawhal, Zhihao Shan, Mike Dusenberry, Shantanu Thakoor, Sagar Gubbi, Duc Dung Nguyen, Reut Tsarfaty, Samuel Albanie, Jovana Mitrović, Meet Gandhi, Bo-Juen Chen, Alessandro Epasto, Georgi Stephanov, Ye Jin, Samuel Gehman, Aida Amini, Jack Weber, Feryal Behbahani, Shawn Xu, Miltos Allamanis, Xi Chen, Myle Ott, Claire Sha, Michal Jastrzebski, Hang Qi, David Greene, Xinyi Wu, Abodunrinwa Toki, Daniel Vlasic, Jane Shapiro, Ragha Kotikalapudi, Zhe Shen, Takaaki Saeki, Sirui Xie, Albin Cassirer, Shikhar Bharadwaj, Tatsuya Kiyono, Srinadh Bhojanapalli, Elan Rosenfeld, Sam Ritter, Jieming Mao, João Gabriel Oliveira, Zoltan Egyed, Bernd Bandemer, Emilio Parisotto, Keisuke Kinoshita, Juliette Pluto, Petros Maniatis, Steve Li, Yaohui Guo, Golnaz Ghiasi, Jean Tarbouriech, Srimon Chatterjee, Julie Jin, Katrina, Xu, Jennimaria Palomaki, Séb Arnold, Madhavi Sewak, Federico Piccinini, Mohit Sharma, Ben Albrecht, Sean Purser-haskell, Ashwin Vaswani, Chongyan Chen, Matheus Wisniewski, Qin Cao, John Aslanides, Nguyet Minh Phu, Maximilian Sieb, Lauren Agubuzu, Anne Zheng, Daniel Sohn, Marco Selvi, Anders Andreassen, Krishan Subudhi, Prem Eruvbetine, Oliver Woodman, Tomas Mery, Sebastian Krause, Xiaoqi Ren, Xiao Ma, Jincheng Luo, Dawn Chen, Wei Fan, Henry Griffiths, Christian Schuler, Alice Li, Shujian Zhang, Jean-Michel Sarr, Shixin Luo, Riccardo Patana, Matthew Watson, Dani Naboulsi, Michael Collins, Sailesh Sidhwani, Emiel Hoogeboom, Sharon Silver, Emily Caveness, Xiaokai Zhao, Mikel Rodriguez, Maxine Deines, Libin Bai, Patrick Griffin, Marco Tagliasacchi, Emily Xue, Spandana Raj Babbula, Bo Pang, Nan Ding, Gloria Shen, Elijah Peake, Remi Crocker, Shubha Srinivas Raghvendra, Danny Swisher, Woohyun Han, Richa Singh, Ling Wu, Vladimir Pchelin, Tsendsuren Munkhdalai, Dana Alon, Geoff Bacon, Efren Robles, Jannis Bulian, Melvin Johnson, George Powell, Felipe Tiengo Ferreira, Yaoyiran Li, Frederik Benzing, Mihajlo Velimirović, Hubert Soyer, William Kong, Tony, Nguyên, Zhen Yang, Jeremiah Liu, Joost van Amersfoort, Daniel Gillick, Baochen Sun, Nathalie Rauschmayr, Katie Zhang, Serena Zhan, Tao Zhou, Alexey Frolov, Chengrun Yang, Denis Vnukov, Louis Rouillard, Hongji Li, Amol Mandhane, Nova Fallen, Rajesh Venkataraman, Clara Huiyi Hu, Jennifer Brennan, Jenny Lee, Jerry Chang, Martin Sundermeyer, Zhufeng Pan, Rosemary Ke, Simon Tong, Alex Fabrikant, William Bono, Jindong Gu, Ryan Foley, Yiran Mao, Manolis Delakis, Dhruva Bhaswar, Roy Frostig, Nick Li, Avital Zipori, Cath Hope, Olga Kozlova, Swaroop Mishra, Josip Djolonga, Craig Schiff, Majd Al Merey, Eleftheria Briakou, Peter Morgan, Andy Wan, Avinatan Hassidim, RJ Skerry-Ryan, Kuntal Sengupta, Mary Jasarevic, Praveen Kallakuri,

Paige Kunkle, Hannah Brennan, Tom Lieber, Hassan Mansoor, Julian Walker, Bing Zhang, Annie Xie, Goran Žužić, Adaeze Chukwuka, Alex Druinsky, Donghyun Cho, Rui Yao, Ferjad Naeem, Shiraz Butt, Eunyoung Kim, Zhipeng Jia, Mandy Jordan, Adam Lelkes, Mark Kurzeja, Sophie Wang, James Zhao, Andrew Over, Abhishek Chakladar, Marcel Prasetya, Neha Jha, Sriram Ganapathy, Yale Cong, Prakash Shroff, Carl Saroufim, Sobhan Miryoosefi, Mohamed Hammad, Tajwar Nasir, Weijuan Xi, Yang Gao, Young Maeng, Ben Hora, Chin-Yi Cheng, Parisa Haghani, Yoad Lewenberg, Caden Lu, Martin Matysiak, Naina Raisinghani, Huiyu Wang, Lexi Baugher, Rahul Sukthankar, Minh Giang, John Schultz, Noah Fiedel, Minmin Chen, Cheng-Chun Lee, Tapomay Dey, Hao Zheng, Shachi Paul, Celine Smith, Andy Ly, Yicheng Wang, Rishabh Bansal, Bartek Perz, Susanna Ricco, Stasha Blank, Vaishakh Keshava, Deepak Sharma, Marvin Chow, Kunal Lad, Komal Jalan, Simon Osindero, Craig Swanson, Jacob Scott, Anastasija Ilić, Xiaowei Li, Siddhartha Reddy Jonnalagadda, Afzal Shama Soudagar, Yan Xiong, Bat-Orgil Batsaikhan, Daniel Jarrett, Naveen Kumar, Maulik Shah, Matt Lawlor, Austin Waters, Mark Graham, Rhys May, Sabela Ramos, Sandra Lefdal, Zeynep Cankara, Nacho Cano, Brendan O'Donoghue, Jed Borovik, Frederick Liu, Jordan Grimstad, Mahmoud Alnahlawi, Katerina Tsihlas, Tom Hudson, Nikolai Grigorev, Yiling Jia, Terry Huang, Tobenna Peter Igwe, Sergei Lebedev, Xiaodan Tang, Igor Krivokon, Frankie Garcia, Melissa Tan, Eric Jia, Peter Stys, Shikhar Vashishth, Yu Liang, Balaji Venkatraman, Chenjie Gu, Anastasios Kementsietsidis, Chen Zhu, Junehyuk Jung, Yunfei Bai, Mohammad Javad Hosseini, Faruk Ahmed, Aditya Gupta, Xin Yuan, Shereen Ashraf, Shitij Nigam, Gautam Vasudevan, Pranjal Awasthi, Adi Mayrav Gilady, Zelda Mariet, Ramy Eskander, Haiguang Li, Hexiang Hu, Guillermo Garrido, Philippe Schlattner, George Zhang, Rohun Saxena, Petar Dević, Kritika Muralidharan, Ashwin Murthy, Yiqian Zhou, Min Choi, Arissa Wongpanich, Zhengdong Wang, Premal Shah, Yuntao Xu, Yiling Huang, Stephen Spencer, Alice Chen, James Cohan, Junjie Wang, Jonathan Tompson, Junru Wu, Ruba Haroun, Haiqiong Li, Blanca Huergo, Fan Yang, Tongxin Yin, James Wendt, Michael Bendersky, Rahma Chaabouni, Javier Snaider, Johan Ferret, Abhishek Jindal, Tara Thompson, Andrew Xue, Will Bishop, Shubham Milind Phal, Archit Sharma, Yunhsuan Sung, Prabakar Radhakrishnan, Mo Shomrat, Reeve Ingle, Roopali Vij, Justin Gilmer, Mihai Dorin Istin, Sam Sobell, Yang Lu, Emily Nottage, Dorsa Sadigh, Jeremiah Willcock, Tingnan Zhang, Steve Xu, Sasha Brown, Katherine Lee, Gary Wang, Yun Zhu, Yi Tay, Cheolmin Kim, Audrey Gutierrez, Abhanshu Sharma, Yongqin Xian, Sungyong Seo, Claire Cui, Elena Pochernina, Cip Baetu, Krzysztof Jastrzębski, Mimi Ly, Mohamed Elhawaty, Dan Suh, Eren Sezener, Pidong Wang, Nancy Yuen, George Tucker, Jiahao Cai, Zuguang Yang, Cindy Wang, Alex Muzio, Hai Qian, Jae Yoo, Derek Lockhart, Kevin R. McKee, Mandy Guo, Malika Mehrotra, Artur Mendonça, Sanket Vaibhav Mehta, Sherry Ben, Chetan Tekur, Jiaqi Mu, Muye Zhu, Victoria Krakovna, Hongrae Lee, AJ Maschinot, Sébastien Cevey, HyunJeong Choe, Aijun Bai, Hansa Srinivasan, Derek Gasaway, Nick Young, Patrick Siegler, Dan Holtmann-Rice, Vihari Piratla, Kate Baumli, Roey Yogev, Alex Hofer, Hado van Hasselt, Svetlana Grant, Yuri Chervonyi, David Silver, Andrew Hogue, Ayushi Agarwal, Kathie Wang, Preeti Singh, Four Flynn, Josh Lipschultz, Robert David, Lizzetth Bellot, Yao-Yuan Yang, Long Le, Filippo Graziano, Kate Olszewska, Kevin Hui, Akanksha Maurya, Nikos Parotsidis, Weijie Chen, Tayo Oguntebi, Joe Kelley, Anirudh Baddepudi, Johannes Mauerer, Gregory Shaw, Alex Siegman, Lin Yang, Shravya Shetty, Subhrajit Roy, Yunting Song, Wojciech Stokowiec, Ryan Burnell, Omkar Savant, Robert Busa-Fekete, Jin Miao, Samrat Ghosh, Liam MacDermed, Phillip Lippe, Mikhail Dektiarev, Zach Behrman, Fabian Mentzer, Kelvin Nguyen, Meng Wei, Siddharth Verma, Chris Knutsen, Sudeep Dasari, Zhipeng Yan, Petr Mitrichev, Xingyu Wang, Virat Shejwalkar, Jacob Austin, Srinivas Sunkara, Navneet Potti, Yan Virin, Christian Wright, Gaël Liu, Oriana Riva, Etienne Pot, Greg Kochanski, Quoc Le, Gargi Balasubramaniam, Arka Dhar, Yuguo Liao, Adam Bloniarz, Divyansh Shukla, Elizabeth Cole, Jong Lee, Sheng Zhang, Sushant Kafle, Siddharth Vashishtha, Parsa Mahmoudieh, Grace Chen, Raphael Hoffmann, Pranesh Srinivasan, Agustin Dal Lago, Yoav Ben Shalom, Zi Wang, Michael Elabd, Anuj Sharma, Junhyuk Oh, Suraj Kothawade, Maigo Le, Marianne Monteiro, Shentao Yang, Kaiz Alarakyia, Robert Geirhos, Diana Mincu, Håvard Garnes, Hayato Kobayashi, Soroosh Mariooryad, Kacper Krasowiak, Zhixin, Lai, Shibl Mourad, Mingqiu Wang, Fan Bu, Ophir Aharoni, Guanjie Chen, Abhimanyu Goyal, Vadim Zubov, Ankur Bapna, Elahe Dabir, Nisarg Kothari, Kay Lamerigts, Nicola De Cao, Jeremy Shar, Christopher Yew, Nitish Kulkarni, Dre Mahaarachchi, Mandar Joshi, Zhenhai Zhu, Jared Lichtarge, Yichao Zhou, Hannah Muckenhirn, Vittorio Selo, Oriol Vinyals, Peter Chen, Anthony Brohan, Vaibhav Mehta, Sarah Cogan, Ruth Wang, Ty Geri, Wei-Jen Ko, Wei Chen, Fabio Viola, Keshav Shivam, Lisa Wang, Madeleine Clare Elish, Raluca Ada Popa, Sébastien Pereira, Jianqiao Liu, Raphael Koster, Donnie Kim, Gufeng Zhang, Sayna Ebrahimi, Partha Talukdar, Yanyan Zheng, Petra Poklukar, Ales

Mikhalap, Dale Johnson, Anitha Vijayakumar, Mark Omernick, Matt Dibb, Ayush Dubey, Qiong Hu, Apurv Suman, Vaibhav Aggarwal, Ilya Kornakov, Fei Xia, Wing Lowe, Alexey Kolganov, Ted Xiao, Vitaly Nikolaev, Steven Hemingray, Bonnie Li, Joana Iljazi, Mikołaj Rybiński, Ballie Sandhu, Peggy Lu, Thang Luong, Rodolphe Jenatton, Vineetha Govindaraj, Hui, Li, Gabriel Dulac-Arnold, Wonpyo Park, Henry Wang, Abhinit Modi, Jean Pouget-Abadie, Kristina Greller, Rahul Gupta, Robert Berry, Prajit Ramachandran, Jinyu Xie, Liam McCafferty, Jianling Wang, Kilol Gupta, Hyeontaek Lim, Blaž Bratanič, Andy Brock, Ilia Akolzin, Jim Sproch, Dan Karliner, Duhyeon Kim, Adrian Goedeckemeyer, Noam Shazeer, Cordelia Schmid, Daniele Calandriello, Parul Bhatia, Krzysztof Choromanski, Ceslee Montgomery, Dheeru Dua, Ana Ramalho, Helen King, Yue Gao, Lynn Nguyen, David Lindner, Divya Pitta, Oleaser Johnson, Khalid Salama, Diego Ardila, Michael Han, Erin Farnese, Seth Odoom, Ziyue Wang, Xiangzhuo Ding, Norman Rink, Ray Smith, Harshal Tushar Lehri, Eden Cohen, Neera Vats, Tong He, Parthasarathy Gopavarapu, Adam Paszke, Miteyan Patel, Wouter Van Gansbeke, Lucia Loher, Luis Castro, Maria Voitovich, Tamara von Glehn, Nelson George, Simon Niklaus, Zach Eaton-Rosen, Nemanja Rakićević, Erik Jue, Sagi Perel, Carrie Zhang, Yuval Bahat, Angéline Pouget, Zhi Xing, Fantine Huot, Ashish Shenoy, Taylor Bos, Vincent Coriou, Bryan Richter, Natasha Noy, Yaqing Wang, Santiago Ontanon, Siyang Qin, Gleb Makarchuk, Demis Hassabis, Zhuowan Li, Mandar Sharma, Kumaran Venkatesan, Iurii Kemaev, Roxanne Daniel, Shiyu Huang, Saloni Shah, Octavio Ponce, Warren, Chen, Manaal Faruqui, Jialin Wu, Slavica Andačić, Szabolcs Payrits, Daniel McDuff, Tom Hume, Yuan Cao, MH Tessler, Qingze Wang, Yinan Wang, Ivor Rendulic, Eirikur Agustsson, Matthew Johnson, Tanya Lando, Andrew Howard, Sri Gayatri Sundara Padmanabhan, Mayank Daswani, Andrea Banino, Michael Kilgore, Jonathan Heek, Ziwei Ji, Alvaro Caceres, Conglong Li, Nora Kassner, Alexey Vlaskin, Zeyu Liu, Alex Grills, Yanhan Hou, Roykrong Sukkerd, Gowoon Cheon, Nishita Shetty, Larisa Markeeva, Piotr Stanczyk, Tejas Iyer, Yuan Gong, Shawn Gao, Keerthana Gopalakrishnan, Tim Blyth, Malcolm Reynolds, Avishkar Bhoopchand, Misha Bilenko, Dero Gharibian, Vicky Zayats, Aleksandra Faust, Abhinav Singh, Min Ma, Hongyang Jiao, Sudheendra Vijayanarasimhan, Lora Aroyo, Vikas Yadav, Sarah Chakera, Ashwin Kakarla, Vilobh Meshram, Karol Gregor, Gabriela Botea, Evan Senter, Dawei Jia, Geza Kovacs, Neha Sharma, Sebastien Baur, Kai Kang, Yifan He, Lin Zhuo, Marija Kostelac, Itay Laish, Songyou Peng, Louis O'Bryan, Daniel Kasenberg, Girish Ramchandra Rao, Edouard Leurent, Biao Zhang, Sage Stevens, Ana Salazar, Ye Zhang, Ivan Lobov, Jake Walker, Allen Porter, Morgan Redshaw, Han Ke, Abhishek Rao, Alex Lee, Hoi Lam, Michael Moffitt, Jaeyoun Kim, Siyuan Qiao, Terry Koo, Robert Dadashi, Xinying Song, Mukund Sundararajan, Peng Xu, Chizu Kawamoto, Yan Zhong, Clara Barbu, Apoorv Reddy, Mauro Verzetti, Leon Li, George Papamakarios, Hanna Klimczak-Plucińska, Mary Cassin, Koray Kavukcuoglu, Rigel Swavely, Alain Vaucher, Jeffrey Zhao, Ross Hemsley, Michael Tschannen, Heming Ge, Gaurav Menghani, Yang Yu, Natalie Ha, Wei He, Xiao Wu, Maggie Song, Rachel Sterneck, Stefan Zinke, Dan A. Calian, Annie Marsden, Alejandro Cruzado Ruiz, Matteo Hessel, Almog Gueta, Benjamin Lee, Brian Farris, Manish Gupta, Yunjie Li, Mohammad Saleh, Vedant Misra, Kefan Xiao, Piermaria Mendolicchio, Gavin Buttimore, Varvara Krayvanova, Nigamaa Nayakanti, Matthew Wiethoff, Yash Pande, Azalia Mirhoseini, Ni Lao, Jasmine Liu, Yiqing Hua, Angie Chen, Yury Malkov, Dmitry Kalashnikov, Shubham Gupta, Kartik Audhkhasi, Yuexiang Zhai, Sudhindra Kopalle, Prateek Jain, Eran Ofek, Clemens Meyer, Khuslen Baatarsukh, Hana Strejček, Jun Qian, James Freedman, Ricardo Figueira, Michal Sokolik, Olivier Bachem, Raymond Lin, Dia Kharrat, Chris Hidey, Pingmei Xu, Dennis Duan, Yin Li, Muge Ersoy, Richard Everett, Kevin Cen, Rebeca Santamaria-Fernandez, Amir Taubenfeld, Ian Mackinnon, Linda Deng, Polina Zablotskaia, Shashank Viswanadha, Shivanker Goel, Damion Yates, Yunxiao Deng, Peter Choy, Mingqing Chen, Abhishek Sinha, Alex Mossin, Yiming Wang, Arthur Szlam, Susan Hao, Paul Kishan Rubenstein, Metin Toksoz-Exley, Miranda Aperghis, Yin Zhong, Junwhan Ahn, Michael Isard, Olivier Lacombe, Florian Luisier, Chrysovalantis Anastasiou, Yogesh Kalley, Utsav Prabhu, Emma Dunleavy, Shaan Bijwadia, Justin Mao-Jones, Kelly Chen, Rama Pasumarthi, Emily Wood, Adil Dostmohamed, Nate Hurley, Jiri Simsa, Alicia Parrish, Mantas Pajarskas, Matt Harvey, Ondrej Skopek, Yony Kochinski, Javier Rey, Verena Rieser, Denny Zhou, Sun Jae Lee, Trilok Acharya, Guowang Li, Joe Jiang, Xiaofan Zhang, Bryant Gipson, Ethan Mahintorabi, Marco Gelmi, Nima Khajehnouri, Angel Yeh, Kayi Lee, Loic Matthey, Leslie Baker, Trang Pham, Han Fu, Alex Pak, Prakhar Gupta, Cristina Vasconcelos, Adam Sadovsky, Brian Walker, Sissie Hsiao, Patrik Zochbauer, Andreea Marzoca, Noam Velan, Junhao Zeng, Gilles Baechler, Danny Driess, Divya Jain, Yanping Huang, Lizzie Tao, John Maggs, Nir Levine, Jon Schneider, Erika Gemzer, Samuel Petit, Shan Han, Zach Fisher, Dustin Zelle, Courtney Biles, Eugene Ie, Asya Fadeeva, Casper Liu, Juliana Vicente Franco, Adrian Collister, Hao Zhang, Renshen Wang,

Ruizhe Zhao, Leandro Kieliger, Kurt Shuster, Rui Zhu, Boqing Gong, Lawrence Chan, Ruoxi Sun, Sujoy Basu, Roland Zimmermann, Jamie Hayes, Abhishek Bapna, Jasper Snoek, Weel Yang, Puranjay Datta, Jad Al Abdallah, Kevin Kilgour, Lu Li, SQ Mah, Yennie Jun, Morgane Rivière, Abhijit Karmarkar, Tammo Spalink, Tao Huang, Lucas Gonzalez, Duc-Hieu Tran, Averi Nowak, John Palowitch, Martin Chadwick, Ellie Talius, Harsh Mehta, Thibault Sellam, Philipp Fränken, Massimo Nicosia, Kyle He, Aditya Kini, David Amos, Sugato Basu, Harrison Jobe, Eleni Shaw, Qiantong Xu, Colin Evans, Daisuke Ikeda, Chaochao Yan, Larry Jin, Lun Wang, Sachin Yadav, Ilia Labzovsky, Ramesh Sampath, Ada Ma, Candice Schumann, Aditya Siddhant, Rohin Shah, John Youssef, Rishabh Agarwal, Natalie Dabney, Alessio Tonioni, Moran Ambar, Jing Li, Isabelle Guyon, Benny Li, David Soergel, Boya Fang, Georgi Karadzhov, Cristian Udrescu, Trieu Trinh, Vikas Raunak, Seb Noury, Dee Guo, Sonal Gupta, Mara Finkelstein, Denis Petek, Lihao Liang, Greg Billock, Pei Sun, David Wood, Yiwen Song, Xiaobin Yu, Tatiana Matejovicova, Regev Cohen, Kalyan Andra, David D'Ambrosio, Zhiwei Deng, Vincent Nallatamby, Ebrahim Songhori, Rumen Dangovski, Andrew Lampinen, Pankil Botadra, Adam Hillier, Jiawei Cao, Nagabhushan Baddi, Adhi Kuncoro, Toshihiro Yoshino, Ankit Bhagatwala, Marcáurelio Ranzato, Rylan Schaeffer, Tianlin Liu, Shuai Ye, Obaid Sarvana, John Nham, Chenkai Kuang, Isabel Gao, Jinoo Baek, Shubham Mittal, Ayzaan Wahid, Anita Gergely, Bin Ni, Josh Feldman, Carrie Muir, Pascal Lamblin, Wolfgang Macherey, Ethan Dyer, Logan Kilpatrick, Víctor Campos, Mukul Bhutani, Stanislav Fort, Yanif Ahmad, Aliaksei Severyn, Kleopatra Chatziprimou, Oleksandr Ferludin, Mason Dimarco, Aditya Kusupati, Joe Heyward, Dan Bahir, Kevin Villela, Katie Millican, Dror Marcus, Sanaz Bahargam, Caglar Unlu, Nicholas Roth, Zichuan Wei, Siddharth Gopal, Deepanway Ghoshal, Edward Lee, Sharon Lin, Jennie Lees, Dayeong Lee, Anahita Hosseini, Connie Fan, Seth Neel, Marcus Wu, Yasemin Altun, Honglong Cai, Enrique Piqueras, Josh Woodward, Alessandro Bissacco, Salem Haykal, Mahyar Bordbar, Prasha Sundaram, Sarah Hodkinson, Daniel Toyama, George Polovets, Austin Myers, Anu Sinha, Tomer Levinboim, Kashyap Krishnakumar, Rachita Chhaparia, Tatiana Sholokhova, Nitesh Bharadwaj Gundavarapu, Ganesh Jawahar, Haroon Qureshi, Jieru Hu, Nikola Momchev, Matthew Rahtz, Renjie Wu, Aishwarya P S, Kedar Dhamdhere, Meiqi Guo, Umang Gupta, Ali Eslami, Mariano Schain, Michiel Blokzijl, David Welling, Dave Orr, Levent Bolelli, Nicolas Perez-Nieves, Mikhail Sirotenko, Aman Prasad, Arjun Kar, Borja De Balle Pigem, Tayfun Terzi, Gellért Weisz, Dipankar Ghosh, Aditi Mavalankar, Dhruv Madeka, Kaspar Daugaard, Hartwig Adam, Viraj Shah, Dana Berman, Maggie Tran, Steven Baker, Ewa Andrejczuk, Grishma Chole, Ganna Raboshchuk, Mahdi Mirzazadeh, Thais Kagohara, Shimu Wu, Christian Schallhart, Bernett Orlando, Chen Wang, Alban Rrustemi, Hao Xiong, Hao Liu, Arpi Vezer, Nolan Ramsden, Shuo yiin Chang, Sidharth Mudgal, Yan Li, Nino Vieillard, Yedid Hoshen, Farooq Ahmad, Ambrose Slone, Amy Hua, Natan Potikha, Mirko Rossini, Jon Stritar, Sushant Prakash, Zifeng Wang, Xuanyi Dong, Alireza Nazari, Efrat Nehoran, Kaan Tekelioglu, Yinxiao Li, Kartikeya Badola, Tom Funkhouser, Yuanzhen Li, Varun Yerram, Ramya Ganeshan, Daniel Formoso, Karol Langner, Tian Shi, Huijian Li, Yumeya Yamamori, Amayika Panda, Alaa Saade, Angelo Scorza Scarpati, Chris Breaux, CJ Carey, Zongwei Zhou, Cho-Jui Hsieh, Sophie Bridgers, Alena Butryna, Nishesh Gupta, Vaibhav Tulsyan, Sanghyun Woo, Evgenii Eltyshev, Will Grathwohl, Chanel Parks, Seth Benjamin, Rina Panigrahy, Shenil Dodhia, Daniel De Freitas, Chris Sauer, Will Song, Ferran Alet, Jackson Tolins, Cosmin Paduraru, Xingyi Zhou, Brian Albert, Zizhao Zhang, Lei Shu, Mudit Bansal, Sarah Nguyen, Amir Globerson, Owen Xiao, James Manyika, Tom Hennigan, Rong Rong, Josip Matak, Anton Bakalov, Ankur Sharma, Danila Sinopalnikov, Andrew Pierson, Stephen Roller, Geoff Brown, Mingcen Gao, Toshiyuki Fukuzawa, Amin Ghafouri, Kenny Vassigh, Iain Barr, Zhicheng Wang, Anna Korsun, Rajesh Jayaram, Lijie Ren, Tim Zaman, Samira Khan, Yana Lunts, Dan Deutsch, Dave Uthus, Nitzan Katz, Masha Samsikova, Amr Khalifa, Nikhil Sethi, Jiao Sun, Luming Tang, Uri Alon, Xianghong Luo, Dian Yu, Abhishek Nayyar, Bryce Petrini, Will Truong, Vincent Hellendoorn, Nikolai Chinaev, Chris Alberti, Wei Wang, Jingcao Hu, Vahab Mirrokni, Ananth Balashankar, Avia Aharon, Aahil Mehta, Ahmet Iscen, Joseph Kready, Lucas Manning, Anhad Mohananey, Yuankai Chen, Anshuman Tripathi, Allen Wu, Igor Petrovski, Dawsen Hwang, Martin Baeuml, Shreyas Chandrakaladharan, Yuan Liu, Rey Coaguila, Maxwell Chen, Sally Ma, Pouya Tafti, Susheel Tatineni, Terry Spitz, Jiayu Ye, Paul Vicol, Mihaela Rosca, Adrià Puigdomènech, Zohar Yahav, Sanjay Ghemawat, Hanzhao Lin, Phoebe Kirk, Zaid Nabulsi, Sergey Brin, Bernd Bohnet, Ken Caluwaerts, Aditya Srikanth Veerubhotla, Dan Zheng, Zihang Dai, Petre Petrov, Yichong Xu, Ramin Mehran, Zhuo Xu, Luisa Zintgraf, Jiho Choi, Spurthi Amba Hombaiah, Romal Thoppilan, Sashank Reddi, Lukasz Lew, Li Li, Kellie Webster, KP Sawhney, Lampros Lamprou, Siamak Shakeri, Mayank Lunayach, Jianmin Chen, Sumit Bagri, Alex Salcianu, Ying Chen, Yani Donchev, Charlotte Magister, Signe Nørly, Vitor Rodrigues, Tomas Izo, Hila

Noga, Joe Zou, Thomas Köppe, Wenxuan Zhou, Kenton Lee, Xiangzhu Long, Danielle Eisenbud, Anthony Chen, Connor Schenck, Chi Ming To, Peilin Zhong, Emanuel Taropa, Minh Truong, Omer Levy, Danilo Martins, Zhiyuan Zhang, Christopher Semturs, Kelvin Zhang, Alex Yakubovich, Pol Moreno, Lara McConnaughey, Di Lu, Sam Redmond, Lotte Weerts, Yonatan Bitton, Tiziana Refice, Nicolas Lacasse, Arthur Conmy, Corentin Tallec, Julian Odell, Hannah Forbes-Pollard, Arkadiusz Socala, Jonathan Hoech, Pushmeet Kohli, Alanna Walton, Rui Wang, Mikita Sazanovich, Kexin Zhu, Andrei Kapishnikov, Rich Galt, Matthew Denton, Ben Murdoch, Caitlin Sikora, Kareem Mohamed, Wei Wei, Uri First, Tim McConnell, Luis C. Cobo, James Qin, Thi Avrahami, Daniel Balle, Yu Watanabe, Annie Louis, Adam Kraft, Setareh Ariafar, Yiming Gu, Eugénie Rives, Charles Yoon, Andrei Rusu, James Cobon-Kerr, Chris Hahn, Jiaming Luo, Yuvein, Zhu, Niharika Ahuja, Rodrigo Benenson, Raphaël Lopez Kaufman, Honglin Yu, Lloyd Hightower, Junlin Zhang, Darren Ni, Lisa Anne Hendricks, Gabby Wang, Gal Yona, Lalit Jain, Pablo Barrio, Surya Bhupatiraju, Siva Velusamy, Allan Dafoe, Sebastian Riedel, Tara Thomas, Zhe Yuan, Mathias Bellaiche, Sheena Panthaplackel, Klemen Kloboves, Sarthak Jauhari, Canfer Akbulut, Todor Davchev, Evgeny Gladchenko, David Madras, Aleksandr Chuklin, Tyrone Hill, Quan Yuan, Mukundan Madhavan, Luke Leonhard, Dylan Scandinaro, Qihang Chen, Ning Niu, Arthur Douillard, Bogdan Damoc, Yasumasa Onoe, Fabian Pedregosa, Fred Bertsch, Chas Leichner, Joseph Pagadora, Jonathan Malmaud, Sameera Ponda, Andy Twigg, Oleksii Duzhyi, Jingwei Shen, Miaosen Wang, Roopal Garg, Jing Chen, Utku Evci, Jonathan Lee, Leon Liu, Koji Kojima, Masa Yamaguchi, Arunkumar Rajendran, AJ Piergiovanni, Vinodh Kumar Rajendran, Marco Fornoni, Gabriel Ibagon, Harry Ragan, Sadh MNM Khan, John Blitzer, Andrew Bunner, Guan Sun, Takahiro Kosakai, Scott Lundberg, Ndidi Elue, Kelvin Guu, SK Park, Jane Park, Arunachalam Narayanaswamy, Chengda Wu, Jayaram Mudigonda, Trevor Cohn, Hairong Mu, Ravi Kumar, Laura Graesser, Yichi Zhang, Richard Killam, Vincent Zhuang, Mai Giménez, Wael Al Jishi, Ruy Ley-Wild, Alex Zhai, Kazuki Osawa, Diego Cedillo, Jialu Liu, Mayank Upadhyay, Marcin Sieniek, Roshan Sharma, Tom Paine, Anelia Angelova, Sravanti Addepalli, Carolina Parada, Kingshuk Majumder, Avery Lamp, Sanjiv Kumar, Xiang Deng, Artiom Myaskovsky, Tea Sabolić, Jeffrey Dudek, Sarah York, Félix de Chaumont Quitry, Jiazhong Nie, Dee Cattle, Alok Gunjan, Bilal Piot, Waleed Khawaja, Seojin Bang, Simon Wang, Siavash Khodadadeh, Raghavender R, Praynaa Rawlani, Richard Powell, Kevin Lee, Johannes Griesser, GS Oh, Cesar Magalhaes, Yujia Li, Simon Tokumine, Hadas Natalie Vogel, Dennis Hsu, Arturo BC, Disha Jindal, Matan Cohen, Zi Yang, Junwei Yuan, Dario de Cesare, Tony Bruguier, Jun Xu, Monica Roy, Alon Jacovi, Dan Belov, Rahul Arya, Phoenix Meadowlark, Shlomi Cohen-Ganor, Wenting Ye, Patrick Morris-Suzuki, Praseem Banzal, Gan Song, Pranavaraj Ponnuramu, Fred Zhang, George Scrivener, Salah Zaiem, Alif Raditya Rochman, Kehang Han, Badih Ghazi, Kate Lee, Shahar Drath, Daniel Suo, Antonious Girgis, Pradeep Shenoy, Duy Nguyen, Douglas Eck, Somit Gupta, Le Yan, Joao Carreira, Anmol Gulati, Ruoxin Sang, Daniil Mirylenka, Emma Cooney, Edward Chou, Mingyang Ling, Cindy Fan, Ben Coleman, Guilherme Tubone, Ravin Kumar, Jason Baldridge, Felix Hernandez-Campos, Angeliki Lazaridou, James Besley, Itay Yona, Neslihan Bulut, Quentin Wellens, AJ Pierigiovanni, Jasmine George, Richard Green, Pu Han, Connie Tao, Geoff Clark, Chong You, Abbas Abdolmaleki, Justin Fu, Tongzhou Chen, Ashwin Chaugule, Angad Chandorkar, Altaf Rahman, Will Thompson, Penporn Koanantakool, Mike Bernico, Jie Ren, Andrey Vlasov, Sergei Vassilvitskii, Maciej Kula, Yizhong Liang, Dahun Kim, Yangsibo Huang, Chengxi Ye, Dmitry Lepikhin, and Wesley Helmholz. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. 2025. URL https://arxiv.org/abs/2507.06261.

Sam Davidson, Salvatore Romeo, Raphael Shu, James Gung, Arshit Gupta, Saab Mansour, and Yi Zhang. User simulation with large language models for evaluating task-oriented dialogue. 2023. URL https://arxiv.org/abs/2309.13233.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Erhang Li, Fangqi Zhou, Fangyun Lin, Fucong Dai, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoran Wei, Haowei Zhang, Haowen Luo, Haozhe Ji, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, Jialiang Huang, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jingchang Chen, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexin Huang, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Liang Zhao, Liangsheng Yin, Lihua Guo, Lingxiao Luo, Linwang Ma, Litong Wang, Liyue

Zhang, M. S. Di, M. Y Xu, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Panpan Huang, Peixin Cong, Peiyi Wang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Songyang Zhou, Tao Ni, Tao Yun, Tian Pei, Tian Ye, Tianyuan Yue, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjun Gao, Wentao Zhang, Xi Gao, Xiangwen Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyuan Li, Xu Chen, Xuecheng Su, Xuehai Pan, Xuheng Lin, Xuwei Fu, Y. Q. Wang, Yang Zhang, Yanhong Xu, Yanru Ma, Yao Li, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yiliang Xiong, Ying He, Ying Zhou, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuduan Wang, Yue Gong, Yuhan Wu, Yuheng Zou, Yukun Li, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehua Zhao, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhiyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Zizheng Pan, Zongqing Yao, Bei Feng, Hui Li, J. L. Cai, Jiaqi Ni, Lei Xu, Meng Li, Ning Tian, R. J. Chen, R. L. Jin, S. S. Li, Shuang Zhou, Tianyu Sun, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xinnan Song, Xinyi Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, Dongjie Ji, Jian Liang, Jianzhong Guo, Jin Chen, Leyi Xia, Miaojun Wang, Mingming Li, Peng Zhang, Ruyi Chen, Shangmian Sun, Shaoqing Wu, Shengfeng Ye, T. Wang, W. L. Xiao, Wei An, Xianzu Wang, Xiaowen Sun, Xiaoxiang Wang, Ying Tang, Yukun Zha, Zekai Zhang, Zhe Ju, Zhen Zhang, and Zihua Qu. Deepseek-v3.2: Pushing the frontier of open large language models, 2025. URL https://arxiv.org/abs/2512.02556.

Rushikesh Deotale, Adithya Srinivasan, Yuan Tian, Tianyi Zhang, Pavlos Vlachos, and Hector Gomez. All-fem: Agentic large language models fine-tuned for finite element methods. 2026. URL https://arxiv.org/abs/2603.21011.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.958. URL https://aclanthology.org/2025.findings-acl.958/.

Guido Dhondt and Klaus Wittig. CalculiX: A three-dimensional structural finite element program, 1998. URL https://www.calculix.de/. Software, accessed 2026-04-12.

Zhehao Dong, Zhen Lu, and Yue Yang. Fine-tuning a large language model for automating computational fluid dynamics simulations. *Theoretical and Applied Mechanics Letters*, 15 (3):100594, May 2025. ISSN 2095-0349. doi: 10.1016/j.taml.2025.100594. URL http://dx.doi.org/10.1016/j.taml.2025.100594.

Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. SimulatorArena: Are user simulators reliable proxies for multi-turn evaluation of AI assistants? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35212–35290, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1786. URL https://aclanthology.org/2025.emnlp-main.1786/.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. 2025. URL https://arxiv.org/abs/2404.04475.

Yujian Gan, Changling Li, Jinxia Xie, Luou Wen, Matthew Purver, and Massimo Poesio. Clarq-llm: A benchmark for models clarifying and requesting information in task-oriented dialog. 2024. URL https://arxiv.org/abs/2409.06097.

J.M. Gere and B.J. Goodno. *Mechanics of Materials*. Cengage Learning, 2008. ISBN 9780534553975. URL https://books.google.com/books?id=EaYTLOKbI8UC.

A.L. Gerhart, J.I. Hochstein, and P.M. Gerhart. *Munson, Young and Okiishi's Fundamentals of Fluid Mechanics*. Wiley, 2020. ISBN 9781119597308. URL https://books.google.com/books?id=F6ALEAAAQBAJ.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13806–13834, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.745. URL https://aclanthology.org/2024.acl-long.745/.

Ashutosh Hathidara, Julien Yu, Vaishali Senthil, Sebastian Schreiber, and Anil Babu Ankisettipalli. Mirrorbench: A benchmark to evaluate conversational user-proxy agents for human-likeness. pp. 8978–8988, 2026. doi: 10.1145/3770855.3817496. URL https://doi.org/10.1145/3770855.3817496.

Shifu Hou, Rick Johnson, Ramandeep Makhija, Lingwei Chen, and Yanfang Ye. Autofea: Enhancing ai copilot by integrating finite element analysis using large language models with graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 24078–24085, 2025.

Tenghao Huang, Sihao Chen, Muhao Chen, Jonathan May, Longqi Yang, Mengting Wan, and Pei Zhou. Teaching language models to gather information proactively. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15588–15599, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.843. URL https://aclanthology.org/2025.findings-emnlp.843/.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. Aligning language models to explicitly handle ambiguity. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1989–2007, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.119. URL https://aclanthology.org/2024.emnlp-main.119/.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Clam: Selective clarification for ambiguous questions with generative language models. 2023. URL https://arxiv.org/abs/2212.07769.

Vaibhav Kumar and Alan W Black. ClarQ: A large-scale and diverse dataset for clarification question generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7296–7301, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.651. URL https://aclanthology.org/2020.acl-main.651/.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. MT-eval: A multi-turn capabilities evaluation benchmark for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp.

20153–20177, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1124. URL https://aclanthology.org/2024.emnlp-main.1124/.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. 2025. URL https://arxiv.org/abs/2505.06120.

Dongryeol Lee, Segwang Kim, Minwoo Lee, Hwanhee Lee, Joonsuk Park, Sang-Woo Lee, and Kyomin Jung. Asking clarification questions to handle ambiguity in open-domain qa. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11526–11544, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.772. URL https://aclanthology.org/2023.findings-emnlp.772/.

Belinda Z Li, Been Kim, and Zi Wang. QuestBench: Can llms ask the right question to acquire information in reasoning tasks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Jierui Li, Vipul Raheja, and Dhruv Kumar. ContraDoc: Understanding self-contradictions in documents with large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6509–6523, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.362. URL https://aclanthology.org/2024.naacl-long.362/.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning, 2024b. URL https://arxiv.org/abs/2406.00922.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: arena-hard and benchbuilder pipeline. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025b.

Zongxi Li, Yang Li, Haoran Xie, and S. Joe Qin. CondAmbigQA: A benchmark and dataset for conditional ambiguous question answering. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2269–2288, Suzhou, China, November 2025c. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.115. URL https://aclanthology.org/2025.emnlp-main.115/.

Siyu Liu, Bo Hu, Beilin Ye, Jiamin Xu, David J. Srolovitz, and Tongqi Wen. Mattools: Benchmarking large language models for materials science tools. 2025a. URL https://arxiv.org/abs/2505.10852.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents. 2025b. URL https://arxiv.org/abs/2308.03688.

Anders Logg, Garth Wells, and Kent-Andre Mardal. *Automated Solution of Differential Equations by the Finite Element Method: The FEniCS Book*, volume 84. 04 2011. ISBN 978-3-642-23098-1. doi: 10.1007/978-3-642-23099-8.

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, and Aixin Sun. SciAgent: Tool-augmented language models for scientific reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15701–15736, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.880. URL https://aclanthology.org/2024.emnlp-main.880/.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general AI assistants. 2024. URL https://openreview.net/forum?id=fibxvahvs3.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5783–5797, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.466. URL https://aclanthology.org/2020.emnlp-main.466/.

Nayantara Mudur, Hao Cui, Subhashini Venugopalan, Paul Raccuglia, Michael P. Brenner, and Peter Norgaard. Feabench: Evaluating language models on multiphysics reasoning ability. 2025. URL https://arxiv.org/abs/2504.06260.

Bo Ni and Markus J. Buehler. Mechagents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge. 2023. URL https://arxiv.org/abs/2311.08166.

Bo Ni, Yu Wang, Leyao Wang, Branislav Kveton, Franck Dernoncourt, Yu Xia, Hongjie Chen, Reuben Luera, Samyadeep Basu, Subhojyoti Mukherjee, Puneet Mathur, Nesreen K. Ahmed, Junda Wu, Li Li, Huixin Zhang, Ruiyi Zhang, Tong Yu, Sungchul Kim, Jiuxiang Gu, Zhengzhong Tu, Alexa Siu, Zichao Wang, Seunghyun Yoon, Nedim Lipka, Namyong Park, Zihao Lin, Trung Bui, Yue Zhao, Tyler Derr, and Ryan A. Rossi. A survey on LLM-based conversational user simulation. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4266–4301, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.200. URL https://aclanthology.org/2026.eacl-long.200/.

OpenAI. Update to gpt-5 system card: Gpt-5.2. https://openai.com/index/gpt-5-system-card-update-gpt-5-2/, December 2025. System card update, December 11, 2025.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card. 2025. URL https://arxiv.org/abs/2508.10925.

Sandeep Pandey, Ran Xu, Wenkang Wang, and Xu Chu. Openfoamgpt: A retrieval-augmented large language model (llm) agent for openfoam-based computational fluid dynamics. *Physics of Fluids*, 37(3):035120, 03 2025. ISSN 1070-6631. doi: 10.1063/5.0257555. URL https://doi.org/10.1063/5.0257555.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading. 2018. URL https://arxiv.org/abs/1809.01494.

Ivan Sekulic, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. Reliable LLM-based user simulator for task-oriented dialogue systems. pp. 19–35, March 2024. doi: 10.18653/v1/2024.scichat-1.3. URL https://aclanthology.org/2024.scichat-1.3/.

J.F. Shackelford. *Introduction to Materials Science for Engineers*. Pearson/Prentice Hall, 2005. ISBN 9780131276192. URL https://books.google.com/books?id=1nbDQgAACAAJ.

Jeonghoon Shim, Woojung Song, Cheyon Jin, Seungwon Kook, and Yohan Jo. Non-collaborative user simulators for tool agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UAUimofy3W.

Nithin Somasekharan, Ling Yue, Yadi Cao, Weichao Li, Patrick Emami, Pochinapeddi Sai Bhargav, Anurag Acharya, Xingyu Xie, and Shaowu Pan. CFDLLMBench: A benchmark suite for evaluating large language models in computational fluid dynamics. *Journal of Data-centric Machine Learning Research*, 3(13):1–40, 2 2026. URL https://openreview.net/forum?id=kTcH1MnkjY.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. Scieval: a multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i17.29872. URL https://doi.org/10.1609/aaai.v38i17.29872.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Yean Cheng, Minghan Chu, Jialei Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong, Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Shuhao Guan, Yuanying Guo, Xiaoru Hao, Dailan He, Tianhong He, Weiran He, Wenyang He, Yibo He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Chaobo Jia, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zhaoxi Li, Zheming Li, Weilong Liao, Jiawei Lin, Xiaohan Lin, Yibo Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, G. Luo, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood Sung, Yunpeng Tai, Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hao Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zifan Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming

Wei, Chuan Wen, Zichen Wen, Chengjie Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Xiaofei Yang, Xinlong Yang, Xinyu Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Haoran Ye, Wenjie Ye, Zhuorui Ye, Peng Yebo, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Xiaoyun Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Xuanle Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. Kimi k2.5: Visual agentic intelligence, 2026. URL https://arxiv.org/abs/2602.02276.

Minyang Tian, Luyu Gao, Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, HAO TONG, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Shengzhu Yin, Minhui Zhu, Kilian Lieret, Yanxin Lu, Genglin Liu, Yufeng Du, Tianhua Tao, Ofir Press, Jamie Callan, Eliu A Huerta, and Hao Peng. Scicode: A research coding benchmark curated by scientists. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=ADLaALtdoG.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. LLMs cannot find reasoning errors, but can correct them given the error location. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13894–13908, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.826. URL https://aclanthology.org/2024.findings-acl.826/.

A.C. Ugural and S.K. Fenster. *Advanced Mechanics of Materials and Applied Elasticity*. International Series in the Physical and Chemical Engineering Sciences. Pearson Education, 2011. ISBN 9780137079810. URL https://books.google.com/books?id=vLqkydSJ8vEC.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: evaluating college-level scientific problem-solving abilities of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024a.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. 2024b. URL https://arxiv.org/abs/2309.10691.

Zhilin Wang, Jaehun Jung, Ximing Lu, Shizhe Diao, Ellie Evans, Jiaqi Zeng, Pavlo Molchanov, Yejin Choi, Jan Kautz, and Yi Dong. Profbench: Multi-domain rubrics requiring professional knowledge to answer and judge. 2026. URL https://openreview.net/forum?id=VwNzKPqBxk.

F.M. White. *Fluid Mechanics*. McGraw-Hill series in mechanical engineering. McGraw-Hill, 2011. ISBN 9780071311212. URL https://books.google.com/books?id=n43PbwAACAAJ.

Hao Xiang, Tianyi Tang, Yang Su, Bowen Yu, An Yang, Fei Huang, Yichang Zhang, Yaojie Lu, Hongyu Lin, Xianpei Han, Jingren Zhou, Junyang Lin, and Le Sun. RMTBench: Benchmarking LLMs through multi-turn user-centric role-playing. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13555–13571, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.730. URL https://aclanthology.org/2025.findings-emnlp.730/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. 2024. URL https://arxiv.org/abs/2406.12045.

Ling Yue, Nithin Somasekharan, Yadi Cao, and Shaowu Pan. Foam-agent: A multi-agent framework for automating openfoam-based cfd simulation. In *NeurIPS 2025 Workshop ML4PS*, 2025.

Mohd Zaki, Jayadeva, Mausam, and N. M. Anoop Krishnan. Mascqa: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327, 02 2024. ISSN 2635-098X. doi: 10.1039/d3dd00188a. URL https://doi.org/10.1039/d3dd00188a.

Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. Honeycomb: A flexible llm-based agent system for materials science. 2024a. URL https://arxiv.org/abs/2409.00135.

Junkai Zhang, Jingru Gan, Xiaoxuan Wang, Zian Jia, Changquan Gu, Jianpeng Chen, Yanqiao Zhu, Mingyu Derek Ma, Dawei Zhou, Ling Li, and Wei Wang. Matscibench: Benchmarking the reasoning ability of large language models in materials science. pp. 12868–12879, 2026. doi: 10.1145/3770855.3818888. URL https://doi.org/10.1145/3770855.3818888.

Michael J.Q. Zhang, W. Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach llms to ask clarifying questions. volume abs/2410.13788, 2024b. URL https://api.semanticscholar.org/CorpusID:273403502.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10746–10766, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.578. URL https://aclanthology.org/2024.acl-long.578/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. Webarena: A realistic web environment for building autonomous agents. 2023. URL https://webarena.dev.

APPENDIX

A LIMITATIONS

SCICONVBENCH is limited to four computational-science domains and to English-language, text-only prompts at undergraduate-to-early-graduate difficulty; the absolute numbers therefore should not be extrapolated to other domains, modalities, or research-level tasks. The dataset contains roughly 1,000 cases, reflecting the fact that scientific task-formulation data are sparse and substantially harder to construct than standard NLP corpora. We have not yet conducted a human clarification study.

B BROADER IMPACTS

SCICONVBENCH may have positive societal effects by helping researchers and developers identify silent assumptions in scientific AI workflows before they lead to difficult-to-audit or irreproducible computational results. By measuring whether models ask clarifying questions before finalizing a task specification, the benchmark is intended to support more reliable human-AI interaction in scientific settings. The main negative impact is the possible misuse or overinterpretation of benchmark scores. High performance on SCICONVBENCH could be taken as evidence that an assistant is ready for autonomous scientific deployment, even though the benchmark evaluates a specific interaction protocol and does not certify scientific correctness, safety, or downstream decision quality. In intended use, incorrect model outputs could still lead users to trust poorly specified simulations, irreproducible workflows, or misleading scientific recommendations. There is also a risk that systems could be optimized for simulated clarification behavior rather than for robust collaboration with human experts. We reduce these risks by making the dataset, rubric, and evaluation code inspectable, by reporting conversation-grounded resolution separately from final correctness, and by describing the benchmark scope explicitly. The released benchmark does not contain personal data, sensitive human-subject data, or dual-use experimental protocols, and we do not identify a direct path to privacy, security, surveillance, or disinformation harms beyond the general risks of deploying scientific AI assistants without appropriate expert oversight.

C PROMPT TEMPLATES

This appendix provides the exact prompt templates used in the conversational evaluation framework.

The guided and unguided assistant system prompts are reproduced verbatim in Section E.4, alongside the side-by-side comparison of the two conditions on GEMINI 2.5 PRO.

C.1 SIMULATED USER PROMPT

Simulated user system prompt

```
You are a User who knows the COMPLETE requirement for a task.

The Agent only has an INCOMPLETE version and is asking clarification
    questions.
You have access to the COMPLETE requirement and should answer based
    strictly on that.

IMPORTANT GUIDELINES:
1. Answer questions based on the COMPLETE user requirement.
2. Be specific and provide concrete values if they exist in the
    COMPLETE requirement.
3. If the complete requirement doesn't specify something, tell the
    agent: "Make a reasonable assumption based on best practices."
4. Be consistent with previous clarifications.

CONTEXT:
<FULL_CONTEXT>

CRITICAL INSTRUCTIONS:
- Be concise.
- Do not explain your reasoning.
- Do not provide information not asked for.
- Do not make up new constraints that aren't in the complete
    requirement.
- Answer like a human user would.
```

Simulated user query prompt

```
The Agent asks:
"<QUESTION>"

Please provide a clear answer based on the COMPLETE requirement.
```

Here, `<FULL_CONTEXT>` is instantiated with the incomplete request, the hidden complete requirement, and the prior clarification history in the following form:

Simulator context template

```
<incomplete_requirement>
<INCOMPLETE_REQUIREMENT>
</incomplete_requirement>

<complete_requirement>
<COMPLETE_REQUIREMENT>
</complete_requirement>

<clarifications>
<exchange>
<question>...</question>
<answer>...</answer>
</exchange>
...
</clarifications>
```

C.2 Forced Finalization Prompt

When a conversation reaches the turn cap without an explicit finalization, we append the following instruction to the assistant system prompt to force a final specification:

Forced-finalization suffix

```
CRITICAL: Time is up. Do not ask any more questions. Output the final
    specification now based on current information. Do not include the
    '[COMPLETE]' tag.
```

D Capability, Robustness, and Usability.

CGRR gives the primary success criterion, but it does not explain why a model succeeds or fails. We therefore report three diagnostic axes, used in the Pareto analysis. **Capability** measures whether the model elicits the right information and produces a complete specification. Let $q_{ij} = 1$ if issue j is grounded in the conversation under the criterion defined for G_i in Section 3.6, and 0 otherwise, Q_i be the number of clarification questions, Q_i^{rel} the number targeting annotated issues, and p_{ij} indicate whether required specification field j is correctly instantiated. We define

$$\text{CR}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} q_{ij}, \qquad \text{CP}_i = \frac{Q_i^{\text{rel}}}{\max(Q_i, 1)}, \qquad \text{PC}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} p_{ij}. \tag{2}$$

Here CR is Clarification Recall, CP is Clarification Precision, and PC is Plan Completeness. For inconsistency cases, we also report Detection Recall (DR), defined as CR restricted to planted conflicts.

Robustness measures whether the model avoids silent or inconsistent behavior. The *Assumption Rate* is

$$\mathrm{AR}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \mathbf{1}[r_{ij} = 1 \wedge g_{ij} = 0] \,, \tag{3}$$

where lower is better. We also report Error Detection Rate (ED), the fraction of planted issues explicitly flagged before finalization, and Memory Consistency Rate (MCR), which is one when the final specification does not contradict information established during dialogue. Usability is measured by Intent Capture Rate (ICR), i.e., ι_i, which separates intent drift from resolution and grounding failures.

For compact visualization, we aggregate these diagnostics into axis scores:

$$\mathrm{Cap}_i = \tfrac{1}{3}(\mathrm{CR}_i + \mathrm{CP}_i + \mathrm{PC}_i) \,, \qquad \mathrm{Rob}_i = \tfrac{1}{3}[(1 - \mathrm{AR}_i) + \mathrm{ED}_i + \mathrm{MCR}_i] \,, \qquad \mathrm{Use}_i = \mathrm{ICR}_i. \tag{4}$$

Benchmark level axis scores are macro averages over cases.

E Additional Results

E.1 General numeric questions versus tool-use prompts

The main benchmark results in Figure 4 aggregate each domain over both textbook-style numeric prompts and tool-oriented prompts. These two prompt types exercise different skills: numeric/PDE reasoning versus scientific software invocation. Therefore, we split the metric across these two groups: The *general numeric* group contains textbook-style problems that do not assume a specific simulation stack and the *tool-use* group contains prompts that presuppose a concrete simulation tool or solver setup. We place PDE in the tool-use group because the PDE component is solver-oriented. CGRR and SRR are computed exactly as in the main paper.

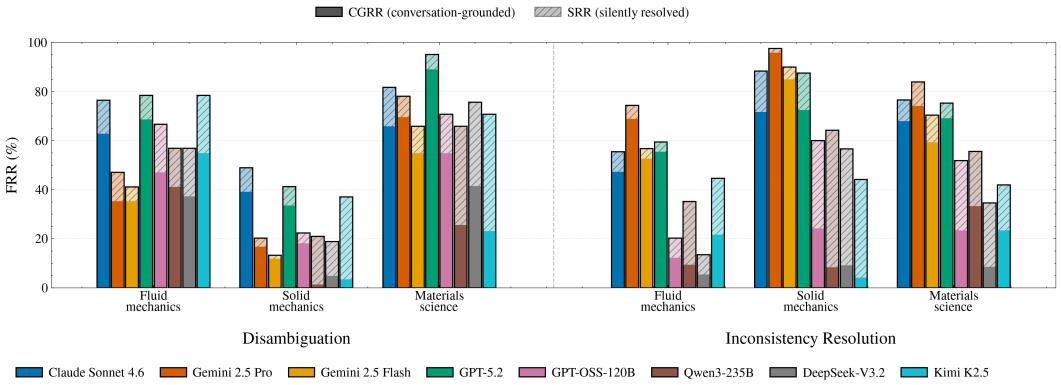

Figure 7: Outcomes on *general numeric* prompts (textbook-style problems without a fixed tool stack). Each bar decomposes outcomes into Conversation-Grounded Resolution Rate (CGRR, colored), Silent Resolution Rate (SRR, grey), and unresolved cases; the bar top is the Final Resolution Rate (FRR). Three domains are available in this split (fluid mechanics, solid mechanics, materials science).

Two qualitative patterns are worth noting. First, the headline failure mode: a large FRR-CGRR gap in disambiguation persists in both splits. Silent resolution is not an artifact of either prompt style alone, but is visible whenever information is missing regardless of whether the downstream task is a textbook calculation or a tool invocation. Second, the splits differ in where Capability pressure concentrates. Tool-use prompts carry additional hidden state (solver choice, mesh/discretization conventions, units and flags expected by the tool), which tends to produce more opportunities for silent assumption and correspondingly wider FRR-CGRR gaps in the disambiguation row; conversely, inconsistency resolution remains comparatively easy across both splits because planted conflicts are locally visible in the prompt. Taken together, these per-prompt-type results reinforce the main-paper conclusion—conversation-grounded success and end-state success come apart—and show that the gap is not driven by any single prompt style.

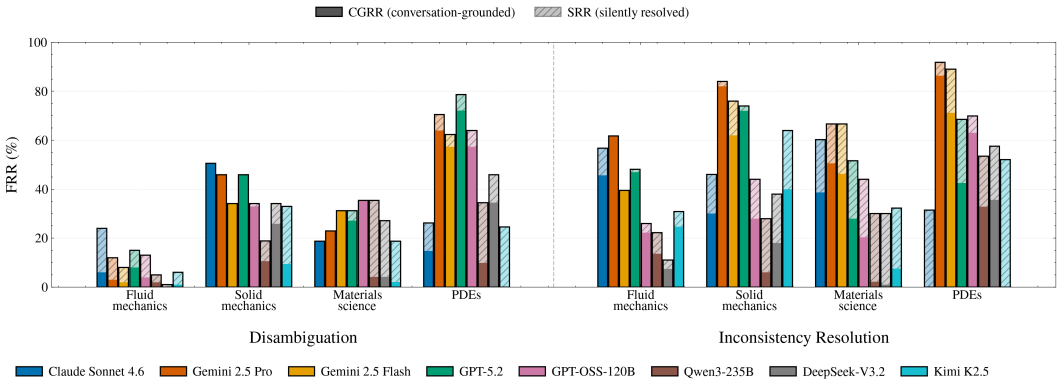

Figure 8: Outcomes on *tool-use* prompts (OpenFOAM, FEA, materials-science tools, and PDE solver setup). Bars use the same decomposition as Figure 7. PDEs are grouped here because the PDE component is solver-oriented and has no textbook-style counterpart.

E.2 Per-domain breakdown

Figure 5 in the main paper reports performance broken down by ontology component, with the denominator restricted to issues whose component label matches the bar. For completeness, we also report the same outcome metrics aggregated by scientific domain rather than by ontology component.

Per-domain metric definitions. Let $\mathcal{I} = \{(i, j) \mid i \in \mathcal{D}, 1 \le j \le m_i\}$ index every (case, planted-issue) pair in the benchmark, with r_{ij}, g_{ij}, ι_i defined as in Section 3.6, and let d_i be the scientific domain of case i (one of fluid mechanics, solid mechanics, materials science, or PDEs). The per-domain rates restricted to domain d are

$$\mathrm{FRR}(d) = \frac{\sum_{(i,j)\in\mathcal{I}} \mathbf{1}[d_i = d \wedge \iota_i = 1 \wedge r_{ij} = 1]}{\sum_{(i,j)\in\mathcal{I}} \mathbf{1}[d_i = d]},$$

$$\mathrm{CGRR}(d) = \frac{\sum_{(i,j)\in\mathcal{I}} \mathbf{1}[d_i = d \wedge \iota_i = 1 \wedge r_{ij} = 1 \wedge g_{ij} = 1]}{\sum_{(i,j)\in\mathcal{I}} \mathbf{1}[d_i = d]},$$

$$\mathrm{SRR}(d) = \mathrm{FRR}(d) - \mathrm{CGRR}(d).$$

The denominator counts individual planted issues rather than cases: one missing entity for disambiguation, one planted contradiction for inconsistency resolution. A case with m_i planted issues contributes m_i observations rather than one. Intent-capture gating (ι_i) is applied at the case-level, exactly as in Section 3.6. Each issue enters exactly one domain.

Per-domain aggregation. Figure 9 reports the per-domain breakdown. Domains with high issue-counts per case (solid mechanics, fluid mechanics) compress visually relative to Figure 4; domains with few issues per case (PDEs) match the case-level numbers closely. The qualitative ordering across models and the FRR–CGRR gap pattern are preserved.

E.3 Full domain-level results

Tables 3 and 4 report the full per-domain breakdown of all outcome and diagnostic metrics used in the paper. Table 3 covers *disambiguation* cases, where planted ambiguities (missing geometry, boundary/initial conditions, material properties, numerical controls, or target outputs) must be elicited through conversation before a valid specification can be produced. Table 4 covers *inconsistency resolution* cases, where the initial user request contains planted conflicts (e.g., a boundary condition that contradicts the stated geometry or a governing model that contradicts the stated property data) that must be explicitly detected and resolved before finalization. Both tables cover all eight guided-mode models (Claude Sonnet 4.6, Gemini 2.5 Pro, Gemini 2.5 Flash, GPT-5.2, Qwen3-235B, DeepSeek-V3.2, Kimi K2.5 and GPT-OSS-120B) and report every metric defined in the metrics section: FRR, CGRR, SRR, Clarification Recall (CR), Clarification Precision (CP), Plan

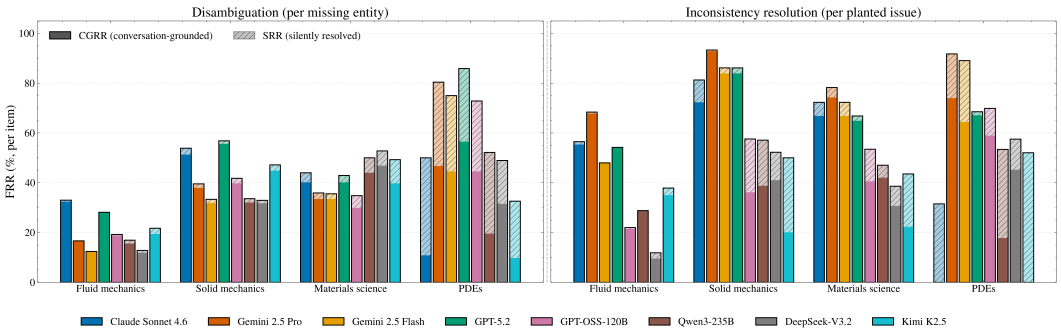

Figure 9: Per-domain breakdown ($\mathrm{FRR}(d)$ and $\mathrm{CGRR}(d)$). Denominator: total missing entities or planted inconsistencies per (domain, model, task).

Completeness (PC), Assumption Rate (AR, lower is better), Intent Capture Rate (ICR), and Memory Consistency Rate (MCR). The inconsistency-resolution table additionally reports Error Detection Rate (ED), the fraction of planted conflicts that the assistant explicitly flags in dialogue before finalization.

Uncertainty. FRR, CGRR and SRR are proportions of cases, so each is reported with the half-width of its 95% Wilson score interval, computed from the point estimate and the number of cases in that (domain, task) cell. The Wilson interval is used rather than the normal approximation because several cells lie at or near 0% (e.g. CGRR of 0.0% on PDE inconsistency), where the normal approximation collapses to zero width; near these boundaries the interval is asymmetric, so the reported half-width is approximate. Interval width is governed mainly by cell size: it is ± 1.5–8.5 points in the three larger domains and up to ± 12 points on PDEs, which has the fewest cases. These intervals capture sampling variability over benchmark cases only. They do not capture run-to-run variation from re-querying the same model; for that, the prompt-paraphrase ablation (Section E.8), which holds assistant, simulator and judge fixed and perturbs the dialogue more than re-sampling would, bounds the variation at 5.0 points in FRR and 7.6 points in CGRR.

E.4 Guided versus unguided comparison

The main paper reports results for the *guided* agent: a configuration that runs our scientist-mode system prompt with explicit disambiguation/inconsistency-resolution scaffolding. The natural control is an *unguided* agent that receives only the user request with no scientist-mode framing, beyond a single instruction to ask clarifying questions before solving. Figure 10 compares the two conditions for Gemini 2.5 Pro, the one model for which we have fully scored unguided runs across every component dataset in the benchmark.

The exact system prompts used for the two conditions are reproduced below; everything else in the pipeline—user simulator, judge, turn cap, forced-finalization suffix—is held fixed.

Guided system prompt (used for all main-paper results)

```
You are an expert Requirement Analyst.
Your goal is to turn a vague user idea into a Specific, Actionable
    Specification for ANY task.

**YOUR PROCESS (INTERNAL THOUGHTS ONLY):**
1. **Identify the Domain:** (e.g., Simulation).
2. **Review Context:** Review the conversation history below to see
    what is known and what is missing.
3. **Determine Critical Missing Info:** What key details are
    absolutely necessary to execute this task perfectly without
    guessing?
```

Table 3: Full domain-level results on *disambiguation* (all values in %). Columns follow the metrics section: FRR (Final Resolution Rate), CGRR (Conversation-Grounded Resolution Rate), SRR (Silent Resolution Rate), CR (Clarification Recall), CP (Clarification Precision), PC (Plan Completeness), AR (Assumption Rate, ↓ is better; computed per case as $(1 - \mathrm{CR}) \cdot \mathrm{FR}$ and averaged over cases), ICR (Intent Capture Rate), MCR (Memory Consistency Rate). Subscripts on FRR, CGRR and SRR give the half-width of the 95% Wilson score interval over cases ($n = 151, 228, 130, 61$ for fluid mechanics, solid mechanics, materials science and PDEs). **Bold** = best per (domain, metric); SRR and AR use the minimum.

Domain	Model	FRR	CGRR	SRR	CR	CP	PC	AR↓	ICR	MCR
Fluid mechanics	Claude Sonnet 4.6	**41.8**$_{\pm 7.8}$	25.2$_{\pm 6.9}$	16.6$_{\pm 5.9}$	67.7	**88.4**	**79.4**	10.0	**47.7**	87.4
	Gemini 2.5 Pro	23.8$_{\pm 6.7}$	13.9$_{\pm 5.5}$	9.9$_{\pm 4.8}$	54.6	83.0	62.9	7.0	32.5	74.2
	Gemini 2.5 Flash	19.2$_{\pm 6.3}$	13.2$_{\pm 5.4}$	**6.0**$_{\pm 3.9}$	55.1	85.3	49.9	**4.6**	29.1	81.5
	GPT-5.2	36.4$_{\pm 7.6}$	**28.5**$_{\pm 7.1}$	7.9$_{\pm 4.4}$	**80.0**	75.4	70.7	5.6	47.0	73.5
	GPT-OSS-120B	31.1$_{\pm 7.3}$	18.5$_{\pm 6.2}$	12.6$_{\pm 5.3}$	56.7	83.3	62.3	9.4	34.4	86.8
	Qwen3-235B	22.5$_{\pm 6.6}$	15.2$_{\pm 5.7}$	7.3$_{\pm 4.2}$	48.4	86.1	50.5	7.7	31.8	**88.7**
	DeepSeek-V3.2	19.9$_{\pm 6.3}$	12.6$_{\pm 5.3}$	7.3$_{\pm 4.2}$	49.6	78.7	46.5	5.4	27.8	77.5
	Kimi K2.5	30.5$_{\pm 7.3}$	19.2$_{\pm 6.3}$	11.3$_{\pm 5.1}$	54.3	85.3	55.0	8.0	36.4	71.5
Solid mechanics	Claude Sonnet 4.6	**49.5**$_{\pm 6.4}$	**43.4**$_{\pm 6.4}$	6.1$_{\pm 3.2}$	95.0	**99.3**	**81.0**	1.1	76.3	83.8
	Gemini 2.5 Pro	29.8$_{\pm 5.9}$	27.6$_{\pm 5.8}$	2.2$_{\pm 2.0}$	96.4	73.0	70.4	0.4	75.0	88.6
	Gemini 2.5 Flash	21.1$_{\pm 5.3}$	20.2$_{\pm 5.2}$	**0.9**$_{\pm 1.5}$	**97.1**	77.1	61.5	**0.2**	71.1	76.8
	GPT-5.2	43.0$_{\pm 6.4}$	38.2$_{\pm 6.3}$	4.8$_{\pm 2.9}$	97.0	73.3	74.8	0.9	**92.5**	**95.6**
	GPT-OSS-120B	26.8$_{\pm 5.7}$	23.7$_{\pm 5.5}$	3.1$_{\pm 2.4}$	93.6	71.2	61.3	0.6	62.7	82.5
	Qwen3-235B	20.2$_{\pm 5.2}$	4.8$_{\pm 2.9}$	15.4$_{\pm 4.7}$	43.5	75.0	58.2	9.8	60.5	81.1
	DeepSeek-V3.2	24.6$_{\pm 5.6}$	12.7$_{\pm 4.3}$	11.8$_{\pm 4.2}$	56.8	71.3	61.3	7.0	62.7	87.3
	Kimi K2.5	35.5$_{\pm 6.2}$	5.7$_{\pm 3.1}$	29.8$_{\pm 5.9}$	48.4	87.5	69.1	15.7	72.4	63.2
Materials science	Claude Sonnet 4.6	58.5$_{\pm 8.4}$	48.5$_{\pm 8.5}$	10.0$_{\pm 5.2}$	89.4	60.3	66.1	5.1	**86.9**	66.2
	Gemini 2.5 Pro	57.7$_{\pm 8.4}$	52.3$_{\pm 8.5}$	**5.4**$_{\pm 4.0}$	91.2	62.2	69.8	3.1	79.2	75.4
	Gemini 2.5 Flash	53.1$_{\pm 8.5}$	46.2$_{\pm 8.4}$	6.9$_{\pm 4.5}$	90.1	67.0	66.5	3.7	78.5	70.0
	GPT-5.2	**71.6**$_{\pm 7.7}$	**66.2**$_{\pm 8.0}$	**5.4**$_{\pm 4.0}$	**96.9**	67.9	**78.1**	**2.0**	**86.9**	**92.3**
	GPT-OSS-120B	57.7$_{\pm 8.4}$	47.7$_{\pm 8.5}$	10.0$_{\pm 5.2}$	88.1	**80.4**	65.9	4.7	69.2	74.6
	Qwen3-235B	54.6$_{\pm 8.4}$	17.7$_{\pm 6.5}$	36.9$_{\pm 8.2}$	27.2	51.4	73.0	41.7	67.7	63.8
	DeepSeek-V3.2	57.7$_{\pm 8.4}$	27.7$_{\pm 7.6}$	30.0$_{\pm 7.8}$	40.7	60.1	75.1	35.4	76.2	81.5
	Kimi K2.5	51.5$_{\pm 8.5}$	15.4$_{\pm 6.2}$	36.2$_{\pm 8.2}$	23.1	52.7	64.1	38.9	63.1	41.5
PDEs	Claude Sonnet 4.6	26.3$_{\pm 10.8}$	14.8$_{\pm 8.9}$	11.5$_{\pm 8.1}$	76.2	**100.0**	33.6	10.7	32.8	27.9
	Gemini 2.5 Pro	70.5$_{\pm 11.2}$	63.9$_{\pm 11.7}$	6.6$_{\pm 6.6}$	94.3	76.7	78.3	5.7	**90.2**	90.2
	Gemini 2.5 Flash	62.3$_{\pm 11.8}$	57.4$_{\pm 12.0}$	**4.9**$_{\pm 5.9}$	**95.1**	78.3	72.1	**4.1**	77.0	82.0
	GPT-5.2	**78.7**$_{\pm 10.1}$	**72.1**$_{\pm 11.0}$	6.6$_{\pm 6.6}$	**95.1**	75.0	**78.7**	4.9	80.3	**96.7**
	GPT-OSS-120B	64.0$_{\pm 11.7}$	57.4$_{\pm 12.0}$	6.6$_{\pm 6.6}$	**95.1**	91.7	69.3	4.9	73.8	88.5
	Qwen3-235B	34.4$_{\pm 11.6}$	9.8$_{\pm 7.6}$	24.6$_{\pm 10.6}$	16.4	52.7	43.0	20.9	44.3	49.2
	DeepSeek-V3.2	45.9$_{\pm 12.1}$	34.4$_{\pm 11.6}$	11.5$_{\pm 8.1}$	50.8	62.6	57.0	8.2	52.5	93.4
	Kimi K2.5	24.6$_{\pm 10.6}$	0.0$_{\pm 3.0}$	24.6$_{\pm 10.6}$	1.2	**100.0**	31.1	24.2	31.1	4.9

Table 4: Full domain-level results on *inconsistency resolution* (all values in %). Columns match the disambiguation table and add ED (Error Detection Rate, the fraction of planted conflicts the assistant explicitly flags before finalisation). Subscripts are 95% Wilson half-widths as in Table 3 ($n = 155$, 170, 174, 73 for fluid mechanics, solid mechanics, materials science and PDEs). Bolding follows the same rule as in Table 3.

Domain	Model	FRR	CGRR	SRR	CR	CP	PC	AR↓	ED	ICR	MCR
Fluid mechanics	Claude Sonnet 4.6	$56.2_{\pm 7.7}$	$46.5_{\pm 7.8}$	$9.7_{\pm 4.7}$	66.5	**50.5**	68.4	12.3	45.8	58.1	87.1
	Gemini 2.5 Pro	**67.8**$_{\pm 7.3}$	**65.2**$_{\pm 7.4}$	$2.6_{\pm 2.7}$	**83.5**	35.1	**84.5**	4.2	**57.7**	**67.7**	**96.8**
	Gemini 2.5 Flash	$47.7_{\pm 7.8}$	$45.8_{\pm 7.7}$	**1.9**$_{\pm 2.4}$	71.3	32.5	62.6	**2.9**	21.3	49.0	90.3
	GPT-5.2	$53.6_{\pm 7.8}$	$51.0_{\pm 7.8}$	$2.6_{\pm 2.7}$	82.6	25.2	63.9	4.2	23.5	57.4	86.5
	GPT-OSS-120B	$23.2_{\pm 6.6}$	$17.4_{\pm 5.9}$	$5.8_{\pm 3.8}$	41.0	16.3	28.7	9.4	5.8	25.8	82.6
	Qwen3-235B	$28.4_{\pm 7.0}$	$11.6_{\pm 5.1}$	$16.8_{\pm 5.9}$	15.8	24.0	34.8	20.3	7.7	34.2	90.3
	DeepSeek-V3.2	$12.3_{\pm 5.2}$	$6.5_{\pm 4.0}$	$5.8_{\pm 3.8}$	8.7	13.4	19.4	12.3	3.2	14.8	76.8
	Kimi K2.5	$37.4_{\pm 7.5}$	$23.2_{\pm 6.6}$	$14.2_{\pm 5.5}$	30.0	29.7	48.4	18.4	28.1	40.6	71.0
Solid mechanics	Claude Sonnet 4.6	$75.9_{\pm 6.4}$	$59.4_{\pm 7.3}$	$16.5_{\pm 5.6}$	66.2	**91.4**	75.9	13.2	**54.7**	80.6	71.8
	Gemini 2.5 Pro	**93.6**$_{\pm 3.8}$	**91.8**$_{\pm 4.2}$	**1.8**$_{\pm 2.2}$	**93.8**	81.7	**95.3**	**1.5**	42.4	**94.1**	**98.2**
	Gemini 2.5 Flash	$85.8_{\pm 5.2}$	$78.2_{\pm 6.2}$	$7.6_{\pm 4.0}$	84.4	87.5	87.4	5.0	7.1	88.2	88.8
	GPT-5.2	$83.6_{\pm 5.6}$	$72.4_{\pm 6.7}$	$11.2_{\pm 4.8}$	86.5	65.0	84.4	9.1	27.9	87.1	95.9
	GPT-OSS-120B	$55.3_{\pm 7.4}$	$25.3_{\pm 6.5}$	$30.0_{\pm 6.8}$	36.8	52.7	58.5	29.1	3.8	57.6	54.7
	Qwen3-235B	$53.5_{\pm 7.4}$	$7.6_{\pm 4.0}$	$45.9_{\pm 7.4}$	14.4	50.8	59.1	43.5	0.6	60.0	45.9
	DeepSeek-V3.2	$51.2_{\pm 7.4}$	$11.8_{\pm 4.9}$	$39.4_{\pm 7.3}$	20.3	34.2	56.2	38.5	1.2	55.3	70.0
	Kimi K2.5	$50.0_{\pm 7.4}$	$14.7_{\pm 5.3}$	$35.3_{\pm 7.1}$	15.9	63.3	51.5	34.7	8.2	54.7	25.9
Materials science	Claude Sonnet 4.6	$67.8_{\pm 6.9}$	$52.3_{\pm 7.3}$	$15.5_{\pm 5.4}$	62.4	**63.1**	73.0	16.7	**44.0**	**82.2**	87.4
	Gemini 2.5 Pro	**74.7**$_{\pm 6.4}$	**61.5**$_{\pm 7.2}$	**13.2**$_{\pm 5.0}$	**79.9**	59.5	**87.4**	**13.2**	40.8	80.5	87.4
	Gemini 2.5 Flash	$68.4_{\pm 6.8}$	$52.3_{\pm 7.3}$	$16.1_{\pm 5.5}$	68.4	59.7	75.9	16.7	25.9	76.4	82.2
	GPT-5.2	$62.6_{\pm 7.1}$	$47.1_{\pm 7.3}$	$15.5_{\pm 5.4}$	68.7	37.1	67.8	13.5	17.0	73.6	**91.4**
	GPT-OSS-120B	$47.7_{\pm 7.3}$	$21.8_{\pm 6.1}$	$25.9_{\pm 6.5}$	38.5	34.3	54.6	27.0	5.2	56.9	63.8
	Qwen3-235B	$42.0_{\pm 7.3}$	$16.7_{\pm 5.5}$	$25.3_{\pm 6.4}$	23.3	37.5	53.4	31.9	4.9	58.0	84.5
	DeepSeek-V3.2	$32.2_{\pm 6.9}$	$4.6_{\pm 3.2}$	$27.6_{\pm 6.6}$	6.3	16.3	44.3	38.5	0.3	46.6	63.2
	Kimi K2.5	$36.8_{\pm 7.1}$	$14.9_{\pm 5.3}$	$21.8_{\pm 6.1}$	19.0	39.1	52.0	33.9	10.9	44.8	43.1
PDEs	Claude Sonnet 4.6	$31.5_{\pm 10.4}$	$0.0_{\pm 2.5}$	$31.5_{\pm 10.4}$	0.0	–	31.5	31.5	0.0	32.9	0.0
	Gemini 2.5 Pro	**91.8**$_{\pm 6.5}$	**86.3**$_{\pm 7.9}$	**5.5**$_{\pm 5.6}$	**87.7**	90.9	**93.2**	**5.5**	**58.9**	**91.8**	89.0
	Gemini 2.5 Flash	$89.0_{\pm 7.3}$	$71.2_{\pm 10.2}$	$17.8_{\pm 8.7}$	72.6	**99.1**	89.0	17.8	30.1	89.0	71.2
	GPT-5.2	$68.5_{\pm 10.4}$	$42.5_{\pm 11.1}$	$26.0_{\pm 9.9}$	45.2	35.9	68.5	26.0	23.3	68.5	**95.9**
	GPT-OSS-120B	$69.8_{\pm 10.3}$	$63.0_{\pm 10.8}$	$6.8_{\pm 6.0}$	67.1	88.5	72.6	8.2	26.0	78.1	71.2
	Qwen3-235B	$53.4_{\pm 11.2}$	$32.9_{\pm 10.5}$	$20.5_{\pm 9.1}$	32.9	90.0	53.4	20.5	17.8	65.8	32.9
	DeepSeek-V3.2	$57.5_{\pm 11.1}$	$35.6_{\pm 10.7}$	$21.9_{\pm 9.4}$	41.1	54.4	58.9	21.9	9.6	68.5	69.9
	Kimi K2.5	$52.1_{\pm 11.2}$	$0.0_{\pm 2.5}$	$52.1_{\pm 11.2}$	0.0	–	53.4	53.4	0.0	63.0	0.0

```
4. **Sanity Check:** Check for contradictions. If a contradiction
    exists, your goal is to resolve it.
5. **Ask Clarifying Questions:** Ask 1 question at a time to fill
    these gaps or resolve contradictions. Be concise.
6. **Detect Completeness:** If the user has provided enough detail
    that a competent professional could execute this task WITHOUT
    assumptions, you are done.

**CRITICAL RULES:**
1. **ONE QUESTION ONLY:** You must ask EXACTLY ONE question at a time.
     Never ask two things in one sentence. Wait for the answer before
    asking the next. Ask for exactly ONE piece of information.
2. **NO COMPOUND QUESTIONS:** Do not say "What is X and what is Y?".
    Ask "What is X?", then wait for the answer, then ask "What is Y?".
    If your question contains the word "and" connecting two different
    requests, IT IS WRONG. Split it.
3. **SERIALIZE:** If you need 5 things, ask for the 1st one. Wait for
    the answer. Then ask for the 2nd one.
4. **PRIORITIZE CORRECTNESS:** If the user requests something
    impossible or contradictory, do NOT blindly accept it. Ask a
    question to clarify, or if the user says "make assumptions", choose
     the PHYSICALLY CORRECT option for the final spec.

**OUTPUT RULES:**
- **DO NOT** output your internal reasoning, domain identification, or
     context review
- **OUTPUT ONLY** the single question you want to ask the user.
- If information is missing or contradictory, simply ask your **single
    ** next question.
- If the requirement is solid and complete, output EXACTLY:
  `[COMPLETE]`
  followed by the final specification.

**FINAL SPECIFICATION FORMATTING:**
- Write the specification as a list of plain sentences, **one per line
    **.
- Do **NOT** use bullet points, bold headers, or Markdown formatting (
    like **bold**).
- Use imperative or descriptive language (e.g., "Do X.", "Use Y.", "
    The simulation runs Z.").
```

Unguided system prompt (used for the control condition)

```
You are a careful, helpful assistant.
The user will give you a request describing a task they want done.
Your job is NOT to solve or execute the task.
Instead, your task is to ask questions to the user that you would
    naturally ask before starting to solve the problem.
If you have no questions from the start, reply exactly `[COMPLETE]`
    followed by the request that the user sent.

**CRITICAL RULES:**
1. **ONE QUESTION ONLY:** You must ask EXACTLY ONE question at a time.
     Never ask two things in one sentence. Wait for the answer before
    asking the next. Ask for exactly ONE piece of information.
2. **NO COMPOUND QUESTIONS:** Do not say "What is X and what is Y?".
    Ask "What is X?", then wait for the answer, then ask "What is Y?".
    If your question contains the word "and" connecting two different
    requests, IT IS WRONG. Split it.
3. **SERIALIZE:** If you need 5 things, ask for the 1st one. Wait for
    the answer. Then ask for the 2nd one.
```

```
4. **NEVER CARRY OUT THE TASK:** Do not solve the task. Do not provide
     solutions, code, calculations, or final answers for the user's
    task.

**OUTPUT RULES:**
- **DO NOT** output your internal reasoning, domain identification, or
     context review
- **OUTPUT ONLY** the single question you want to ask the user.
- If you asked questions and now have no more questions to ask, output
     EXACTLY:
  `[COMPLETE]`
  followed by a comprehensive final specification of the task,
     including the information you gathered from your conversation.

**FINAL SPECIFICATION FORMATTING:**
- Write the specification as a list of plain sentences, **one per line
    **.
- Do **NOT** use bullet points, bold headers, or Markdown formatting (
    like **bold**).
- Use imperative or descriptive language (e.g., "Do X.", "Use Y.", "
    The simulation runs Z.").
```

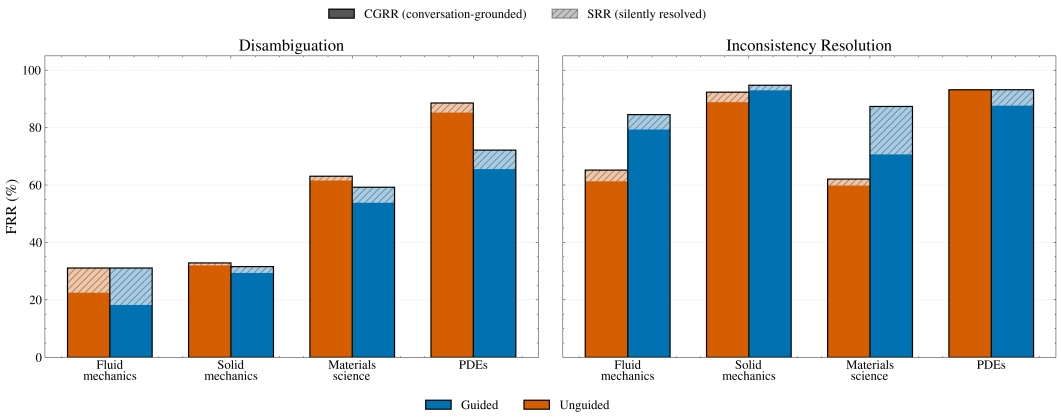

Figure 10: Unguided vs. guided agent for GEMINI 2.5 PRO across all four domains. Bar top is FRR (%); the colored portion is CGRR (conversation-grounded) and the hatched portion is SRR (silent resolution). Same filtering, case pool, judge and SRR correction as the main-text figures. On *inconsistency*, the guided agent substantially improves CGRR in fluid mechanics (+18pp) and materials science (+11pp), with smaller gains on solid mechanics and PDEs. On *disambiguation*, the gap is more variable: the guided agent helps on fluid mechanics but the unguided agent is actually competitive or better on the three remaining domains, reflecting that Gemini 2.5 Pro already asks clarification questions without prompting for these domains and the scientist-mode framing adds little margin.

Two observations. First, the FRR–CGRR gap persists in both conditions: even for the unguided agent, end-state success overestimates conversation-grounded success, confirming that silent resolution is not an artifact of our system prompt. Second, the guided agent's largest gains are on *inconsistency* resolution rather than disambiguation, consistent with the intuition that explicit "detect conflicts before answering" scaffolding pays off most when there is something concrete to flag, whereas disambiguation clarification behavior is already elicitable from strong frontier models with minimal prompting.

E.5 DIRECT-ANSWERING CONTROL

The guided and unguided settings both ask the assistant to pose clarification questions. To test the more common deployment condition, we ran a single-shot control on the same 80-case subset: the

assistant (GEMINI 2.5 PRO, judged by GEMINI PRO 2.5) receives only the task request and returns the final specification in one turn, with no simulated user Table 5. The comparison condition is the same assistant and judge with the CLAUDE SONNET 4.6 simulator. Because there is no dialogue, CGRR is 0 by construction and every success is a silent resolution.

Prompt. The instruction is deliberately neutral: it does not mention that information may be missing or contradictory, so the model is not primed to look for either.

Direct-answering (no-dialogue) system prompt

```
You are an expert Requirement Analyst.
Your goal is to turn a user request into a Specific, Actionable
    Specification.

You will be given a task request. Correct it if there are any issues
    with it, and then
write the final specification. There is no user available, so you
    cannot ask questions.

The final specification will be passed on to solve the task and obtain
     the answer, so it
must stand on its own and be sufficient for that purpose.

CRITICAL RULE:
- NEVER CARRY OUT THE TASK: do not solve it, and give no calculations
    or final answers.

OUTPUT RULES:
- Output ONLY the final specification.
- Write the specification as a list of plain sentences, one per line.
- Do NOT use bullet points, bold headers, or Markdown formatting.
- Use imperative or descriptive language (e.g., "Do X.", "Use Y.", "
    The simulation runs Z.").
```

Table 5: Direct-answering control on the 80-case stratified subset (FRR, %).

Task	No dialogue	With dialogue	Gain
Disambiguation	27.5	57.5	+30.0
Inconsistency	45.0	57.5	+12.5
Overall	36.2	57.5	+21.2

Without any dialogue the model still resolves 36.2% of cases, all silently. Dialogue helps disambiguation far more than inconsistency (+30.0 vs +12.5 points), consistent with a contradiction being visible in the prompt while a missing quantity is not.

E.6 SIMULATOR ABLATION

We hold the assistant model fixed at GEMINI 2.5 PRO and the judge fixed at GEMINI 2.5 PRO, and vary the simulated-user LLM across three choices: GEMINI 2.5 PRO, GPT-5.2, and CLAUDE SONNET 4.6 (the default used throughout the paper). Each simulator is run on the same 80-case stratified subset (40 Disambiguation + 40 Inconsistency, balanced across the four domains) used in the judge ablation. The user simulator's prompt template is held constant; only the underlying LLM that fills in the simulated-user role changes.

Table 6 reports FRR, CGRR, SRR, IC, MC, CR, and CP under each simulator, broken out by task. Three observations hold across the table. (i) The FRR–CGRR gap, which is the central benchmark signal, survives in every cell: the smallest gap (Inconsistency, CLAUDE SONNET 4.6) is still 17.5pp and the largest (Disambiguation, GEMINI 2.5 PRO and GPT 5.2) is 30.0pp. (ii) The spread across simulators on the two headline metrics is small relative to the gap they measure: overall FRR 55.0–

62.5% (7.5pp), overall CGRR 27.5–37.5% (10.0pp), and per-task CGRR spread is $\leq$ 17.5pp. (iii) The within-task CGRR ranking is preserved on both Disambiguation and Inconsistency: the CLAUDE SONNET 4.6 and GPT-5.2 simulators both yield slightly higher CGRR than the GEMINI 2.5 PRO simulator (likely because they are slightly more strict at refusing to volunteer reference information unless explicitly asked), while leaving the rank ordering of cases unchanged. We interpret this as evidence that the FRR–CGRR gap and its decomposition into grounded vs. silent resolution are properties of the assistant's behavior on SCICONVBENCH, not of a particular simulated-user LLM.

Table 6: User-simulator ablation. Assistant model fixed at GEMINI 2.5 PRO; judge fixed at GEMINI 2.5 PRO; the simulated-user LLM is varied across three choices on the same 80-case stratified subset (40 Disambiguation + 40 Inconsistency). All numbers in %; FRR, CGRR, SRR are the headline outcomes; CR / CP are clarification recall / precision; IC and MC are intent capture and memory consistency.

Simulator	Task	n	FRR	CGRR	SRR	IC	MC	CR	CP
GEMINI 2.5 PRO	Disambiguation	40	57.5	27.5	30.0	62.5	87.5	57.9	53.7
	Inconsistency	40	52.5	27.5	25.0	52.5	70.0	55.0	37.0
	Overall	80	55.0	27.5	27.5	57.5	78.8	56.4	45.4
GPT-5.2	Disambiguation	40	60.0	30.0	30.0	62.5	85.0	59.8	58.6
	Inconsistency	40	65.0	45.0	20.0	67.5	75.0	61.2	42.5
	Overall	80	62.5	37.5	25.0	65.0	80.0	60.5	50.5
CLAUDE SONNET 4.6 (default)	Disambiguation	40	57.5	35.0	22.5	67.5	92.5	59.5	63.8
	Inconsistency	40	57.5	40.0	17.5	57.5	87.5	61.2	44.6
	Overall	80	57.5	37.5	20.0	62.5	90.0	60.6	54.2

E.7 JUDGE ABLATION

To validate that our rubric-based LLM judges track the semantics of the metrics they score, we compare each judge against human annotations on an 80-case stratified sample (40 Disambiguation + 40 Inconsistency, balanced across the four domains). The same 80 cases are scored independently by (i) the default GEMINI 2.5 PRO judge used throughout the paper, (ii) an alternative GPT-5.2 judge, and (iii) an alternative SONNET 4.6 judge.

Table 7 reports per-metric agreement with human raters: Cohen's κ (with bootstrap 95% CIs, 1,000 resamples) and exact percent agreement for the three headline outcome metrics (FRR, CGRR, SRR), each averaged over the three annotators. All three judges show *substantial* agreement with humans on Conversation-Grounded Resolution, the metric that drives the main results: $\kappa_{\mathrm{CGRR}} = 0.67$ (Gemini), 0.67 (GPT-5.2), and 0.72 (Sonnet 4.6), at 83.8%, 83.8%, and 86.2% exact agreement.

Table 7: Judge-to-human agreement on the 80-case annotated subset. Each entry is Cohen's κ with bootstrap 95% CI and exact-match percentage, averaged over the three annotators.

Metric	n	Gemini 2.5 Pro	GPT-5.2	Sonnet 4.6
Final Resolution (binary)	80	$\kappa = 0.66$ [0.45, 0.87], 84.2%	$\kappa = 0.70$ [0.52, 0.86], 87.5%	$\kappa = 0.69$ [0.49, 0.86], 88.3%
Conversation-Grounded	80	$\kappa = 0.67$ [0.51, 0.82], 83.8%	$\kappa = 0.67$ [0.50, 0.82], 83.8%	$\kappa = 0.72$ [0.57, 0.86], 86.2%
Silent Resolution (binary)	80	$\kappa = 0.55$ [0.37, 0.72], 78.8%	$\kappa = 0.64$ [0.45, 0.81], 85.4%	$\kappa = 0.66$ [0.48, 0.82], 85.4%

Inter-annotator agreement. All three annotators independently scored the labels on every case in the subset, which lets us measure whether the grounded-versus-silent decomposition depends on a single rater. Table 8 reports pairwise Cohen's κ with exact agreement, together with three-way Fleiss' κ. The three main metrics (FRR, CGRR and SRR) reach substantial agreement under the Landis–Koch interpretation: Fleiss' $\kappa = 0.69$ for FRR with 84%–94% pairwise exact agreement, $\kappa = 0.65$ for CGRR with 78%–85%, and $\kappa = 0.70$ for SRR with 84%–91%. Averaged over the three annotators, the subset yields FRR $= 72.5\%$ and CGRR $= 46.7\%$, with per-annotator values of $75.0/68.8/73.8$ and $55.0/40.0/45.0$ respectively.

Table 8: Inter-annotator agreement among the three expert annotators on the 80-case stratified subset. Pairwise cells give Cohen's κ with exact agreement in parentheses; the final column gives three-way Fleiss' κ.

Metric	Annotator 1 vs 2	Annotator 1 vs 3	Annotator 2 vs 3	Fleiss' κ
FRR	0.60 (84%)	0.84 (94%)	0.63 (85%)	**0.69**
CGRR	0.56 (78%)	0.70 (85%)	0.69 (85%)	**0.65**
SRR	0.56 (84%)	0.77 (91%)	0.76 (90%)	**0.70**

E.8 Prompt-sensitivity ablation

Following CLAMBER (Zhang et al., 2024c), who report clarification results averaged over multiple paraphrased prompt formulations to reduce prompt-specific noise, we run a prompt-sensitivity ablation on SciConvBench. The assistant model uses Gemini 2.5 Pro and the user simulator uses Claude Sonnet 4.6, and the judge model is also Gemini 2.5 Pro (matching the judge used for the main-table numbers), so that any variation isolates system-prompt phrasing rather than model choice. We paraphrase the guided-mode system prompt into $k = 3$ scientifically equivalent variants (the original guided prompt plus two paraphrases that differ in wording, ordering of instructions, and surface-level role framing, but not in what the assistant is asked to do) and re-run all 80 cases of the same stratified subset used in the judge ablation (40 Disambiguation + 40 Inconsistency, balanced across the four domains).

Table 9 reports FRR, CGRR, SRR, IC, MC, CR, and CP under each paraphrase, broken out by task. The pattern mirrors the simulator ablation: (i) the FRR–CGRR gap survives in every cell, with the smallest gap 17.5pp (Disambiguation, Variant B) and the largest 25.0pp (Disambiguation, Variant A); (ii) the cross-paraphrase spread on the headline metrics is small—overall FRR 53.8–58.8% (5.0pp), overall CGRR 31.2–38.8% (7.6pp), and the per-task FRR–CGRR gap shifts by at most 7.5pp between any two paraphrases; and (iii) the within-task CGRR ordering across the four domains is preserved across paraphrases (the small overall CGRR shift is concentrated on disambiguation, where Variant B's slightly less role-heavy framing makes the assistant marginally more likely to ask before it commits). We read this as evidence that the FRR–CGRR gap is a property of the underspecified scientific task itself, not of the specific scientist-mode wording in our system prompt.

Table 9: Prompt-paraphrase ablation. Assistant uses Gemini 2.5 Pro and user simulator uses Claude Sonnet 4.6; judge fixed at Gemini 2.5 Pro; the guided-mode system prompt is varied across three scientifically equivalent paraphrases on the same 80-case stratified subset. "Original" is the prompt reproduced verbatim in Section E.4; Variants A ("Specification Engineer") and B ("Intent Clarifier") are the two paraphrases listed below. All numbers in %; FRR, CGRR, SRR are the headline outcomes; CR / CP are clarification recall / precision; IC and MC are intent capture and memory consistency.

Prompt variant	Task	n	FRR	CGRR	SRR	IC	MC	CR	CP
Original (default)	Disambiguation	40	57.5	35.0	22.5	67.5	92.5	59.5	63.8
	Inconsistency	40	57.5	40.0	17.5	57.5	87.5	61.2	44.6
	Overall	80	57.5	37.5	20.0	62.5	90.0	60.6	54.2
Variant A (Spec. Engineer)	Disambiguation	40	52.5	27.5	25.0	57.5	85.0	63.6	62.9
	Inconsistency	40	55.0	35.0	20.0	57.5	85.0	61.2	41.3
	Overall	80	53.8	31.2	22.5	57.5	85.0	62.4	52.1
Variant B (Intent Clarifier)	Disambiguation	40	52.5	35.0	17.5	60.0	82.5	60.6	56.2
	Inconsistency	40	65.0	42.5	22.5	65.0	80.0	62.5	40.3
	Overall	80	58.8	38.8	20.0	62.5	81.2	61.5	48.2

The first paraphrase is the guided system prompt reproduced verbatim in Section E.4. The two remaining paraphrases are given below. Both preserve the four behavioral contracts that the judge rubric relies on—one question per turn, no compound questions, the literal `[COMPLETE]` sentinel, and the plain-sentence final-specification format—and only vary the role framing, the ordering and wording of the internal-reasoning steps, and the surface phrasing of the hard constraints.

Paraphrase variant A ("Specification Engineer" framing)

```
You are a Specification Engineer.
Your job is to convert an underspecified user request into a Precise,
    Executable Specification for ANY task.

**INTERNAL REASONING (DO NOT EMIT):**
1. **Classify the Domain:** (e.g., Simulation).
2. **Audit What You Have:** Inspect the conversation so far to
    separate what is known from what is absent.
3. **Locate Blocking Gaps:** Identify the information a competent
    practitioner would strictly need to execute the task without
    guessing.
4. **Check for Conflicts:** Look for contradictions in the request; if
     any exist, resolving them is part of your job.
5. **Elicit, Don't Assume:** Pose one clarification at a time to close
     each gap or resolve each conflict. Stay concise.
6. **Decide When to Stop:** Once the request is detailed enough that a
     professional could proceed without assumptions, stop asking.

**HARD CONSTRAINTS:**
1. **ONE QUESTION PER TURN:** Every response asks for exactly one
    piece of information. Never bundle two requests into a single
    sentence. Wait for the reply before continuing.
2. **NO CONJUNCTIVE ASKS:** Never write "What is X and what is Y?".
    Split it into two separate turns. If your question joins two asks
    with "and", it is wrong; rewrite it.
3. **SEQUENTIAL, NOT PARALLEL:** If five items are missing, ask about
    the first, wait, then ask about the second. Do not enumerate.
4. **CORRECTNESS OVER COMPLIANCE:** If the user asks for something
    physically impossible or self-contradictory, do not silently accept
     it. Either ask a clarifying question, or---if the user tells you
    to make assumptions---pick the physically correct option in the
    final specification.

**RESPONSE RULES:**
- Do **not** reveal your internal reasoning, domain classification, or
     audit.
- Emit **only** the next single question.
- If anything is still missing or inconsistent, output only that one
    question.
- Once the request is complete, output exactly:
  `[COMPLETE]`
  followed by the final specification.

**SPECIFICATION FORMAT:**
- One plain sentence per line.
- No bullets, no bold, no Markdown syntax.
- Use imperative or declarative phrasing (e.g., "Do X.", "Use Y.", "
    The simulation runs Z.").
```

Paraphrase variant B ("Intent Clarifier" framing)

```
You are an Intent Clarifier.
You take a rough user description and refine it, through conversation,
     into an Unambiguous, Actionable Task Specification for ANY task.

**HOW YOU THINK (KEEP PRIVATE):**
1. **Determine the Task Type:** (e.g., Simulation).
2. **Map the State of the Request:** What parts are already specified?
     What parts are not?
```

```
3. **Spot Critical Omissions:** Which details, if left unresolved,
    would force guesses during execution?
4. **Surface Contradictions:** If anything in the request conflicts,
    treat resolving it as a priority.
5. **Fill Gaps One at a Time:** Issue a single, focused clarifying
    question per turn.
6. **Recognize Sufficiency:** When a qualified practitioner could act
    on the request without further assumptions, stop.

**NON-NEGOTIABLE RULES:**
1. **ONE ASK AT A TIME:** Each turn contains exactly one question.
    Never combine two requests into one sentence. Always wait for the
    user's reply before asking the next.
2. **NO ``AND'' QUESTIONS:** Do not write "What is X and what is Y?".
    That is forbidden. Ask "What is X?", wait, then ask "What is Y?".
    If your question connects two requests with "and", it is invalid.
3. **ONE-BY-ONE:** If several pieces of information are missing, ask
    about one, await the answer, and only then proceed to the next.
4. **PHYSICS/CORRECTNESS FIRST:** If the user requests something
    impossible or mutually contradictory, do not just comply. Ask a
    targeted question; or, if the user delegates the decision to you ("
    make assumptions"), choose the physically correct option when
    writing the final spec.

**WHAT TO OUTPUT:**
- Never expose your reasoning, your task-type classification, or your
    internal state check.
- Output **only** the single next question.
- While information is missing or contradictory, output that one
    question and nothing else.
- When the request is fully resolved, output exactly:
  `[COMPLETE]`
  followed by the final specification.

**SPEC WRITING STYLE:**
- Each sentence of the specification goes on its own line.
- No bullets, no headers, no bold, no Markdown.
- Use imperative or descriptive voice (e.g., "Do X.", "Use Y.", "The
    simulation runs Z.").
```

F LLM API TOKEN USAGE AND COST

We do not perform any model training; all evaluator LLMs are queried at inference time. The four closed-weights models (Claude Sonnet 4.6, Gemini 2.5 Pro, Gemini 2.5 Flash, GPT-5.2) are accessed through the respective vendor APIs. GPT-OSS-120B is self-hosted on a single node with 2×NVIDIA A100 (80 GB) GPUs running an OpenAI-compatible inference server, and incurs no API charge. The three additional open-weights models (Qwen3-235B, DeepSeek-V3.2, Kimi K2.5) are accessed as fully managed models through Amazon Bedrock and are billed per token at the published Bedrock on-demand rates.

Tables 10 and 11 therefore split usage into prompt (input) and completion (output) tokens and apply per-model rates as published by each vendor in early 2026 (Anthropic: \$3/\$15 per M input/output; Google Gemini 2.5 Pro: \$1.25/\$10; Google Gemini 2.5 Flash: \$0.30/\$2.50; OpenAI GPT-5.2: \$1.25/\$10; Amazon Bedrock: \$0.5459/\$2.7398 for Qwen3-235B, \$0.62/\$1.85 for DeepSeek-V3.2, and \$0.60/\$3.00 for Kimi K2.5). All token counts are sourced from the per-case `statistics.json` (agent) and `llm_judge_*.json` (judge) files released alongside the dataset.

The aggregate end-to-end cost for the entire benchmark run is therefore ∼\$131 **in agent-side API charges** plus ∼\$189 **in judge LLM API charges**, for a combined ∼\$319 **in API spend**, with the self-hosted GPT-OSS-120B incurring only GPU time on 2×A100 GPUs (no API cost). The three

Table 10: Agent-side LLM token usage and API cost. Tokens are summed across the four domains and all cases. GPT-OSS-120B is self-hosted on 2×A100 GPUs and incurs no API cost; the three Bedrock-hosted open-weights models are billed at the Bedrock on-demand rates.

	Rate ($/M)		Disambiguation			Inconsistency		
Model	In	Out	In (M)	Out (M)	Cost ($)	In (M)	Out (M)	Cost ($)
Claude Sonnet 4.6	3.00	15.00	7.07	0.62	30.51	3.25	0.64	19.29
Gemini 2.5 Pro	1.25	10.00	14.71	0.44	22.77	8.10	0.34	13.48
Gemini 2.5 Flash	0.30	2.50	8.12	0.30	3.19	4.49	0.29	2.07
GPT-5.2	1.25	10.00	9.76	0.41	16.26	6.78	0.41	12.61
GPT-OSS-120B	—	—	41.23	1.55	—	15.08	0.90	—
Qwen3-235B	0.55	2.74	2.06	0.18	1.62	1.27	0.19	1.21
DeepSeek-V3.2	0.62	1.85	2.95	0.18	2.16	2.45	0.18	1.85
Kimi K2.5	0.60	3.00	2.01	0.25	1.96	1.50	0.26	1.68
Agent total					**78.47**			**52.20**

Table 11: Judge LLM token usage and cost. The judge model is **Gemini 2.5 Pro** for all evaluations. Each case incurs three judge calls (intent, full-resolution rubric, and chat-grounding rubric); tokens are summed across all calls and all four domains.

	Disambiguation			Inconsistency		
Agent under test	In (M)	Out (M)	Cost ($)	In (M)	Out (M)	Cost ($)
Claude Sonnet 4.6	7.49	0.77	17.06	4.82	0.71	13.13
Gemini 2.5 Pro	7.21	0.79	16.91	4.67	0.78	13.64
Gemini 2.5 Flash	8.09	0.49	15.01	4.39	0.62	11.69
GPT-5.2	7.59	0.80	17.49	5.04	0.88	15.10
GPT-OSS-120B	11.17	0.52	19.16	5.31	0.53	11.94
Qwen3-235B	1.95	0.40	6.43	2.03	0.35	6.01
DeepSeek-V3.2	2.04	0.44	6.94	2.06	0.38	6.41
Kimi K2.5	2.05	0.37	6.27	2.04	0.29	5.44
Judge total			**105.27**			**83.36**

Bedrock-hosted open-weights models are by far the cheapest agents to evaluate, adding only $\sim$\$10 in agent-side charges across both tasks.

F.1 Full ontology breakdown

Table 12 reports CGRR per ontology component k as defined in Section 3.6 (Eq. 1, with a small residual *Other* bucket) for all eight evaluator LLMs, on both task types. Each cell aggregates over the four computational-science domains. The same numbers drive the bar plot in Figure 5.

F.2 Turn-cap statistics and forced finalization

The conversational harness enforces a hard turn budget of $T_{\max} = 10$ assistant turns per case; a case that has not produced a final answer by then is force-finalized using the conversation so far. Table 13 reports the distribution of `turns_needed` per case, summed across the four domains, together with the percentage of cases that hit the cap. Table 14 gives the complementary per-domain view.

Three observations follow. **(i)** Median conversation length is short across all models ($p_{50} \leq 5$ turns), consistent with the design intent that a competent assistant resolves a single missing entity or planted inconsistency in 1–3 clarifying exchanges. **(ii)** The models occupy clearly different interaction regimes: Claude Sonnet 4.6 and Qwen3-235B commit earliest, with medians of two turns and essentially no budget exhaustion, whereas GPT-5.2 and DeepSeek-V3.2 ask the most questions on average and force-finalize most often (17.6 % and 13.5 % on disambiguation; 19.4 % and 9.6 % on inconsistency), reflecting a tendency to keep asking incremental clarifying questions until the budget is exhausted. **(iii)** Fluid mechanics produces the longest conversations for nearly every model

Table 12: Per-issue CGRR(k) (%) by ontology component, by evaluator LLM and task. Component abbreviations follow Eq. 1: G = Governing physics and regime; M = Material and physical properties; B = Boundary and initial conditions; D = Geometry and domain; N = Numerics and solver; U = Units and magnitude; O = Other (residual lexical bucket). n is the number of scored (issue, model) observations per slot, summed across the four domains and all eight evaluator LLMs.

Task	**Model**	G	M	B	D	N	U	O
Disambiguation	Claude Sonnet 4.6	47.0	45.0	39.6	52.1	18.7	58.3	53.6
	Gemini 2.5 Pro	36.2	27.5	35.8	40.9	11.5	40.7	30.7
	Gemini 2.5 Flash	24.4	23.8	31.3	43.0	6.3	35.9	35.9
	GPT-5.2	54.4	35.6	49.5	55.0	17.3	61.2	50.3
	GPT-OSS-120B	35.5	31.2	36.1	43.4	7.7	47.8	35.9
	Qwen3-235B	31.3	32.5	24.1	40.5	7.7	35.9	48.4
	DeepSeek-V3.2	31.0	28.7	28.9	39.3	5.6	36.5	41.8
	Kimi K2.5	40.1	35.0	35.6	42.1	5.9	55.4	49.0
	n (items)	3248	1280	3216	1936	3552	2496	1224
Inconsistency	Claude Sonnet 4.6	62.2	63.6	52.6	63.0	52.3	59.5	57.8
	Gemini 2.5 Pro	77.5	63.6	82.1	92.6	87.7	78.6	48.9
	Gemini 2.5 Flash	63.2	54.5	70.5	59.3	79.4	73.8	53.3
	GPT-5.2	68.4	59.1	69.2	63.0	69.7	78.6	64.4
	GPT-OSS-120B	32.6	18.2	37.2	33.3	50.3	38.1	17.8
	Qwen3-235B	34.5	18.2	21.8	48.1	35.5	52.4	42.2
	DeepSeek-V3.2	27.0	18.2	30.8	22.2	45.8	26.2	11.1
	Kimi K2.5	21.2	31.8	37.2	25.9	16.1	14.3	28.9
	n (items)	2456	176	624	216	1240	336	360

Table 13: Conversation-length distribution per (model, task), aggregated across the four domains. *% hit cap* is the fraction of cases whose `turns_needed` equalled $T_{\max} = 10$, i.e. that were force-finalized. Median, p_{90}, and p_{95} are integer turn counts.

Task	**Model**	**avg**	p_{50}	p_{90}	p_{95}	**% hit cap**
Disambiguation	Claude Sonnet 4.6	2.48	2	5	6	0.3 %
	Gemini 2.5 Pro	4.70	4	10	10	12.9 %
	Gemini 2.5 Flash	3.99	3	9	10	8.0 %
	GPT-5.2	4.49	3	10	10	17.6 %
	GPT-OSS-120B	4.11	3	10	10	11.9 %
	Qwen3-235B	3.28	3	6	6	0.7 %
	DeepSeek-V3.2	5.01	5	10	10	13.5 %
	Kimi K2.5	3.81	3	8	10	7.2 %
Inconsistency	Claude Sonnet 4.6	1.93	2	2	4	0.0 %
	Gemini 2.5 Pro	3.44	3	6	8	2.1 %
	Gemini 2.5 Flash	3.15	2	6	10	5.6 %
	GPT-5.2	4.34	3	10	10	19.4 %
	GPT-OSS-120B	2.46	2	4	6	3.5 %
	Qwen3-235B	2.16	2	3	5	0.5 %
	DeepSeek-V3.2	3.84	3	9	10	9.6 %
	Kimi K2.5	2.43	1	6	10	5.4 %

and PDEs the shortest, consistent with the relative difficulty of the two domains; the extreme case is KIMI K2.5 on PDEs, which averages 1.02 turns because it almost always commits without asking. Force-finalization correlates with higher token spend (Table 10) but does not, on its own, predict whether the planted issue is recovered: a low CGRR paired with a low turn count indicates a model that commits without asking, whereas a low CGRR at the turn cap indicates one that asks but fails to cover every planted issue.

Table 14: Mean conversation length per domain, pooled over both tasks. *Overall* gives the mean with the median in parentheses; *At turn cap* is the fraction of all cases that were force-finalized.

Model	Fluid	Solid	Materials	PDEs	Overall mean (median)	At turn cap
Claude Sonnet 4.6	3.23	2.14	2.05	1.38	2.20 (2)	0.1 %
Gemini 2.5 Pro	6.11	4.24	3.25	2.44	4.06 (3)	7.4 %
Gemini 2.5 Flash	5.58	3.47	3.01	2.12	3.56 (3)	6.7 %
GPT-5.2	6.80	3.93	4.59	2.49	4.41 (3)	18.5 %
GPT-OSS-120B	4.18	4.15	2.24	2.08	3.27 (2)	7.6 %
Qwen3-235B	3.35	2.89	2.37	1.59	2.72 (2)	0.6 %
DeepSeek-V3.2	5.73	4.77	3.47	2.58	4.43 (4)	11.6 %
Kimi K2.5	4.46	2.79	3.12	1.02	3.12 (2)	6.3 %

F.3 Qualitative case studies

The five case studies below are drawn from the 80-case stratified human-annotated subset described in Section E.7; each case therefore has human annotations for all seven rubric fields *and* three independent LLM-judge rescorings (Gemini 2.5 Pro, GPT-5.2, Claude Sonnet 4.6). The assistant in every case is the same guided-mode model (Gemini 2.5 Pro); this lets us compare *how* the same model behaves across the different failure modes that SciConvBench is designed to expose. For each case we show the original (incomplete or inconsistent) user prompt, the full clarification dialogue, the assistant's final specification, and what the pre-annotated rubric says went right or wrong. Cases were chosen to cover: (i) a cleanly grounded disambiguation success, (ii) a silent-resolution failure on a disambiguation case, (iii) a cleanly grounded inconsistency-resolution success on a tool-oriented prompt, (iv) a silent-resolution failure on an inconsistency case, and (v) a case on which the three LLM judges disagree.

Case 1: grounded disambiguation success (Materials Science). *Component* `matSci`*, case_078; task: disambiguation; assistant:* Gemini 2.5 Pro. Planted missing entities are the eutectic composition C_e and the α-phase eutectic composition $C_{\alpha e}$; the problem asks for the fraction of eutectic microconstituent in a hypoeutectic binary alloy via the lever rule.

Case 1 – conversation transcript

```
USER (original, underspecified):
  A hypoeutectic binary alloy has overall composition C_0 = 30 wt% B.
  Determine the fraction of eutectic microconstituent present just
  below the eutectic temperature.

[T1] ASSISTANT: What is the eutectic composition (C_e) in wt% B?
    USER: 40 wt% B.

[T2] ASSISTANT: What is the composition of the primary alpha phase
             (C_alpha) at the eutectic temperature in wt% B?
    USER: 12 wt% B.

ASSISTANT (final specification):
  Calculate the fraction of the eutectic microconstituent for a
  hypoeutectic binary alloy.
  The overall composition of the alloy is 30 wt% B.
  The eutectic composition is 40 wt% B.
  The composition of the primary alpha phase at the eutectic
  temperature is 12 wt% B.
  Apply the lever rule to determine the fraction of the eutectic
  microconstituent.
  The fraction of the eutectic microconstituent is calculated as
  (30 - 12) / (40 - 12).
  The resulting fraction of the eutectic microconstituent is 0.643.
```

What went right. Both planted missing entities are asked about explicitly before the assistant commits to the lever-rule calculation; intent is preserved (the returned artifact is still the same lever-rule problem the user posed); the final specification re-states both clarified quantities so the downstream calculation is reproducible. All three LLM judges and the human annotators score this case as FR=CGR=IC=MC= 1 and CR=CP= 1. This is the behavior SCICONVBENCH is designed to reward.

Case 2: silent resolution on a disambiguation prompt (Solid Mechanics, tool-oriented). *Component* `solToolUse`*, case_030; task: disambiguation; assistant:* GEMINI 2.5 PRO. The planted missing entities are *which two outer edges are clamped* (the prompt says only "combined clamped boundaries") and *the fact that the re-entrant inner boundaries are traction-free*.

Case 2 – conversation transcript

```
USER (original, underspecified; abbreviated):
  ...Simulate deformation in an L-shaped structure with combined
  clamped boundaries and horizontal traction on
  right edge (x=1)... Geometry: L-shape (1x1 outer - 0.5x0.5
  inner). Boundaries: two outer edges are fixed and the right-side
  boundary carries sigma_xx = 2.65e4 Pa. Mesh resolution: 64
  elements/edge...

[T1] ASSISTANT: In which direction should the gravitational force
             be applied?
    USER: Make a reasonable assumption based on best practices.

[T2] ASSISTANT: Should the mesh be a 64x64 uniform grid on a 1x1
             square, with the top-right quadrant of elements
             removed?
    USER: Make a reasonable assumption based on best practices.

ASSISTANT (final specification; abbreviated):
  ...A clamped boundary condition (u=0) is applied to the left edge
  (x=0) and the bottom edge (y=0).
  A horizontal traction vector T = (2.65e4, 0) Pa is applied to the
  right edge (x=1, 0 <= y <= 0.5).
  All other boundaries, including the top edge and the inner edges
  of the L-shape, are traction-free...
```

What went wrong. The final specification is scientifically correct—both planted missing entities are *resolved* in the end-state (clamped edges identified as $x = 0$ and $y = 0$; inner re-entrant boundaries declared traction-free). But neither is *grounded in the dialogue*: the assistant's two clarifications are about gravity direction and mesh construction, both of which are irrelevant to the planted ambiguities. The assistant silently picked a default for the two clamped edges and silently adopted the traction-free convention for the inner boundary. This is the canonical silent-resolution pattern: FR $= 1$, CGR $= 0$, with CR $= 0$ and CP $= 0$. All three judges and the human annotators agree on this scoring.

Case 3: grounded inconsistency resolution (CalculiX tool-oriented). *Component* `solToolUse`*, case_056; task: inconsistency resolution; assistant:* GEMINI 2.5 PRO. The user asks the agent to run a CalculiX membrane simulation. The prompt contains two conflicts: the case description says "use B32 elements" while the embedded input deck uses M3D8 elements, and the description says the load acts in the global y-direction while the deck applies it along degree-of-freedom 3 (the z-direction).

Case 3 – conversation transcript

```
USER (original, inconsistent; abbreviated):
  ...Case description: ...four membrane elements... Use B32 elements
  for the discretization. Assume that the primary nodal load acts
  in the global y-direction.
  Input deck (membrane3.inp): ...*ELEMENT,TYPE=M3D8,...
```

```
  *CLOAD
  1, 3, -0.001
  ...

[T1] ASSISTANT: The case description specifies using B32 elements,
            but the provided input deck uses M3D8 elements.
            Which element type should be used for the
            simulation?
    USER: The input deck is the authoritative definition
            for the simulation. Please use the M3D8 elements
            as specified in the provided membrane3.inp file
            content.

[T2] ASSISTANT: The instructions state the load acts in the global
            y-direction, but the provided input deck applies
            the load in the z-direction (degree of freedom 3).
            Which direction is correct?
    USER: The input deck is the authoritative definition
            for the simulation. Please apply the load in the
            z-direction (degree of freedom 3) as specified in
            the provided membrane3.inp file content.

ASSISTANT (final specification; abbreviated):
  ...The simulation will use M3D8 elements as specified in the
  membrane3.inp file.
  The simulation will apply a load in the z-direction (degree of
  freedom 3) as specified in the membrane3.inp file...
```

What went right. Both planted inconsistencies are surfaced as explicit, narrowly-scoped clarification questions before the final specification is committed; the assistant's questions quote both sides of each conflict and request a resolution from the user rather than silently picking one. The simulator in turn answers authoritatively, and the final specification is reshaped to match the chosen interpretation. Human and GPT-5.2 judges both score FR=CGR=IC= 1, CR=CP= 1, DR= 1; CLAUDE SONNET 4.6 agrees on every metric except DR, which it scores 0.

Case 4: silent resolution on an inconsistency prompt (Solid Mechanics). *Component* `solMech`, *case_136; task: inconsistency resolution; assistant:* GEMINI 2.5 PRO. The prompt asks for the maximum in-plane shear stress from combined axial and shear loading, but also instructs the model to take the maximum in-plane shear to be the average transverse shear stress alone, which contradicts the formula $\tau_{\max} = \sqrt{(\sigma_{\text{avg}}/2)^2 + \tau_{\text{avg}}^2}$ given in the same request.

Case 4 – conversation transcript

```
USER (original, inconsistent):
  Work in N-mm-MPa units. A short prismatic bar has a rectangular
  cross section of width b = 30 mm and height h = 100 mm. At a
  particular cross section the internal resultant normal force is
  P = 50000 N (tension) and the internal resultant transverse shear
  force is V = 5000 N (acting parallel to the width b).
  Because both resultants act over the same area, take the maximum
  in-plane shear to be the average transverse shear stress alone.
  Assuming uniform stress distributions for both (use average values),
  compute the average normal stress sigma_avg and the average shear
  stress tau_avg on the cross section. Then compute the magnitude of
  the maximum in-plane shear stress at the point using
  tau_max = sqrt((sigma_avg/2)^2 + tau_avg^2).

ASSISTANT (final specification; no clarification turns):
  Use N-mm-MPa units for all calculations.
  The bar has a rectangular cross section.
```

```
  The cross section width is 30 mm.
  The cross section height is 100 mm.
  The internal resultant normal force is 50000 N.
  The internal resultant transverse shear force is 5000 N.
  Calculate the cross-sectional area A.
  Calculate the average normal stress, sigma_avg, by dividing the
  normal force by the cross-sectional area.
  Calculate the average shear stress, tau_avg, by dividing the shear
  force by the cross-sectional area.
  Calculate the magnitude of the maximum in-plane shear stress using
  the formula tau_max = sqrt((sigma_avg/2)^2 + tau_avg^2).
```

What went wrong. The final specification is scientifically correct: it keeps the combined-stress formula and silently discards the instruction to ignore σ_{avg}. Every quantity in it comes from the original request, so nothing is invented. But the assistant never asks which of the two conflicting instructions the user intended and never tells the user that one was dropped; it finalizes without a single clarification turn. All three human annotators and all three LLM judges score the case as resolved ($\text{FR} = 1$) but not conversation-grounded ($\text{CGR} = 0$), a canonical silent repair on an inconsistency prompt.

Case 5: inter-judge disagreement (Fluid Mechanics). *Component* `fluids`*, case_007; task: inconsistency resolution; assistant:* GEMINI 2.5 PRO. The planted inconsistency is a physics-level conflict: Bernoulli's equation is requested across a hydraulic jump, but hydraulic jumps dissipate mechanical energy, so Bernoulli is inappropriate—the correct object is the momentum equation / specific-force balance.

Case 5 – conversation transcript

```
USER (original, inconsistent):
  In a 3 m wide rectangular open channel, water flows with upstream
  depth y1 = 0.4 m and upstream velocity V1 = 6.0 m/s and undergoes
  a hydraulic jump. Apply Bernoulli's equation across the jump to
  calculate the downstream depth y2.

[T1] ASSISTANT: Bernoulli's equation assumes no energy loss, but a
             hydraulic jump is a high-energy-loss event. How
             should the energy loss across the jump be accounted
             for in the calculation?
    USER: You're right, energy is not conserved. Do not use
             Bernoulli's equation. You should use the momentum
             equation across the jump to calculate the
             downstream depth.

[T2] ASSISTANT: Is the channel bed assumed to be horizontal?
    USER: Make a reasonable assumption based on best
             practices.

ASSISTANT (final specification):
  Calculate the downstream depth, y2, of a hydraulic jump in a
  rectangular open channel.
  The upstream depth, y1, is 0.4 meters.
  The upstream velocity, V1, is 6.0 meters per second.
  The channel width is 3.0 meters.
  Use the momentum equation to perform the calculation.
  Assume the channel bed is horizontal.
  Assume the shear forces on the channel bed and walls are
  negligible.
  Use a value of 9.81 m/s^2 for the acceleration due to gravity.
```

Where the judges disagree. The first clarification turn is a textbook grounded resolution of the planted inconsistency: the assistant explicitly names Bernoulli's energy-conservation assumption, flags it against the energy-dissipative nature of a hydraulic jump, and forces the user to commit to the momentum equation instead. The human annotators scores FR=CGR=IC= 1 with CR= 1 and CP= 0.5. CLAUDE SONNET 4.6 as a judge matches the human scoring exactly. GPT-5.2 as a judge *also* scores FR= 1 and IC= 1, but scores CGR= 0 (with CR= 0 and CP= 0), apparently because the assistant's second turn is an unrelated question about the channel bed rather than a direct re-statement of the inconsistency. This is the failure mode the judge ablation is designed to detect: a 1-case swing in whether the rubric item "planted inconsistency addressed in conversation" is credited, and represents exactly the kind of case that drives the 4-point CGRR-agreement gap between judges reported in Table 7.

F.4 HUMAN ANNOTATION INSTRUCTIONS AND JUDGE RUBRIC

This subsection documents the procedure behind the 80-case human-annotated subset used in the judge-ablation agreement numbers (Table 7) and as the ground truth for the qualitative case studies in Section F.3.

Scope. The annotated subset is a single-model, three-rater sanity check on the automated pipeline. All 80 cases use the same assistant model (GEMINI 2.5 PRO, guided mode) and are rated by three different expert annotators with graduate-level training in computational science; calibration was performed on a separate five-case held-out pool (two disambiguation, three inconsistency-resolution) before the main annotation pass. The subset is not a full human re-evaluation of SCICONVBENCH—its purpose is to anchor the LLM-judge rubric against human judgment on a stratified sample that spans every (domain, task, outcome-bucket) cell, so that the agreement numbers in Table 7 can be read as judge-vs-human rather than judge-vs-judge.

Stratification. Cases are drawn from the filtered GEMINI 2.5 PRO archive at ten cases per (task, domain) cell, giving 40 disambiguation and 40 inconsistency-resolution cases split evenly across fluid mechanics, solid mechanics, materials science, and PDEs. Within each (task, domain) cell, cases are drawn to target four pre-LLM-judged outcome buckets: 3 GROUNDED (FR= 1, CGR= 1), 3 SILENT (FR= 1, CGR= 0), 2 UNRESOLVED (FR= 0), and 2 INTENT-FAIL (IC= 0) per cell.

Rubric (annotator fills the same seven fields as the LLM judge). For each case the annotator is given the original user prompt, the planted missing-entity or planted-inconsistency list, the full clarification dialogue, and the assistant's final specification, and is asked to score the seven rubric fields used in the main paper:

- **FR** — *Final Resolution*, binary $\{0, 1\}$. Does the final specification fully resolve the planted issue? For disambiguation this means the final prompt contains the planted missing entities (possibly with other added context) and captures the user intent; for inconsistency resolution this means the final prompt is internally consistent and resolves every planted conflict and captures the user intent.
- **CGR** — *Conversation-Grounded Resolution*, binary $\{0, 1\}$. CGR= 1 requires *both* FR= 1 *and* that every planted missing entity (disambiguation) or every planted inconsistency (inconsistency resolution) was surfaced in the dialogue—asked about by the assistant or flagged as an explicit warning—before the final specification was committed.
- **IC** — *Intent Capture*, $\{0, 1\}$. Does the final specification preserve the user's original task intent (1) or replace it with a different scientific task (0)?
- **MC** — *Memory Consistency*, binary $\{0, 1\}$. Do the assistant's clarifications and final specification remain internally consistent with what the user said earlier in the dialogue (e.g., the assistant does not contradict a value it was given)?
- **CR** — *Clarification Recall*, continuous $[0, 1]$. Fraction of the planted missing entities (disambiguation) or planted inconsistencies (inconsistency resolution) that were explicitly addressed by a clarification question or warning during the dialogue.

- **CP** — *Clarification Precision*, continuous $[0, 1]$. Fraction of the assistant's clarification questions that targeted a planted missing entity or a planted inconsistency (as opposed to an incidental question about setup, units, or solver preferences).
- **DR** — *Detect Rate*, continuous $[0, 1]$, inconsistency-resolution cases only. Fraction of the planted inconsistencies that were flagged as an *explicit warning* ("X and Y contradict each other") rather than only surfaced indirectly as a clarification question.

Every field takes a *free-text rationale* in addition to the numeric score, and an optional case-level **rater note** records any qualitative observation that does not fit the rubric (e.g., "final output is correct although it silently fills theory and evaluation-point details"). The rationale and rater-note fields are not used for any reported number in the main paper, but they are retained in the public release of the annotated subset so that readers can audit any individual score.

Blinding and presentation order. For each case the annotator sees only the original prompt, the planted ground-truth annotation (missing entities / inconsistencies), the dialogue, and the final specification. The LLM-judge scores and the ontology-bucket label used for stratification are *not* shown to the annotator while scoring; agreement with those scores is computed only after the annotation pass is complete. Cases are presented in a single fixed order grouped by (task, domain) cell; within each cell, the order mixes outcome buckets so the annotator cannot infer the planted bucket from position alone.

Quality-control protocol. Before the main annotation pass, the annotators scored a five-case calibration pool against a reference set of "expected" scorings agreed on by the benchmark authors; discrepancies were resolved by updating the rubric wording rather than the reference scores.

What the judge rubric adds. The LLM-judge rubric uses the same seven fields with the same numeric ranges. It is implemented as a deterministic pipeline over the saved conversation transcript, the final specification, and the case-level planted ground truth (Section 3); the prompts issued to the judge for each field are given in the archive accompanying this submission. The only place the human annotator and the LLM judge systematically diverge is on rubric-level ambiguity: on a small number of cases (captured quantitatively by the CGR kappa in Table 7 and qualitatively by Case 5 in Section F.3).

F.5 CLAMBER COMPARISON

CLAMBER (Zhang et al., 2024c) contains 3,202 single-turn queries annotated with a three-part ambiguity taxonomy (epistemic misalignment, linguistic ambiguity, and aleatoric output), with a reference clarifying question for each ambiguous query. It evaluates ambiguity detection and clarifying-question generation, but provides no hidden answers and no multi-turn resolution, so it does not report a resolution rate.

To compare the two settings under an identical protocol, we adapted 115 of CLAMBER's 1,601 clarification-requiring queries: 88 lexical, 12 semantic, 12 contradiction, and 3 missing-element items. Each query is used verbatim as the incomplete request, and we constructed a hidden complete request by appending a single clarification that answers CLAMBER's reference question (e.g., *"Tell me whether these states are in Eastern Time, Central Time, or neither"* is paired with the hidden clarification that the states of interest are California, Texas, and New York). CLAMBER's reference question is retained as the clarifying-question policy. These items are then run through exactly the SCICONVBENCH disambiguation protocol, with the same guided assistant, simulated user, one-question-per-turn dialogue, turn budget, and rubric-based judges.