

*lossfunk*

# METIS: Mentoring Engine for Thoughtful Inquiry & Solutions

Abhinav Rajeev Kumar, Dhruv Trehan, Paras Chopra

Lossfunk

{abhinav.kumar, dhruv.trehan, paras}@lossfunk.com

## Abstract

Many students lack access to expert research mentorship. We ask whether an AI mentor can move undergraduates from an idea to a paper. We build *METIS*, a tool-augmented, stage-aware assistant with literature search, curated guidelines, methodology checks, and memory. We evaluate *METIS* against GPT-5 and Claude Sonnet 4.5 across six writing stages using LLM-as-a-judge pairwise preferences, student-persona rubrics, short multi-turn tutoring, and evidence/compliance checks. On 90 single-turn prompts, LLM judges preferred *METIS* to Claude Sonnet 4.5 in 71% and to GPT-5 in 54%. Student scores (clarity/actionability/constraint-fit; 90 prompts  $\times$  3 judges) are higher across stages. In multi-turn sessions (five scenarios/agent), *METIS* yields slightly higher final quality than GPT-5. Gains concentrate in document-grounded stages (D–F), consistent with stage-aware routing and grounding; failure modes include premature tool routing, shallow grounding, and occasional stage misclassification.

## Introduction

Large language models can explain, suggest, and critique. But can one mentor a student or early-career researcher from an initial, rough idea to a publishable paper? We study this question with *METIS*, an AI research mentor whose primary goal is to guide undergraduates from ideation to a publishable conference paper. *METIS* grounds its advice in conference instructions, research guides, and literature (arXiv/OpenReview) and keeps context across sessions so learners can make steady progress over weeks.

This paper contributes:

1. a practical mentoring workflow and stage-aware evaluation that turns “research mentorship” into concrete tasks;
2. a simple, inspectable system with tools for literature search, guidelines retrieval, methodology checks, and memory;
3. an empirical comparison to GPT-5 and Claude Sonnet 4.5 on single-turn judgments and multi-turn tutoring; and
4. open materials like prompts, logs, and scripts, to reproduce results.

## Related Work

Tool-augmented assistants combine reasoning with external tools (search, calculators, code) via modular designs such as ReAct (Yao et al. 2022), MRKL (Karpas et al. 2022), and Toolformer (Schick et al. 2023). *METIS* follows this line but targets research mentorship tasks (idea critique, planning, writing feedback) rather than generic tool use. For grounding, retrieval-augmented generation (Lewis et al. 2020) is paired with lightweight self-critique (Shinn et al. 2023) to encourage evidence and methodology checks.

Autonomous scientific-process agents target end-to-end automation of the research lifecycle. Denario (Villaescusa-Navarro et al. 2025) uses a multi-agent pipeline to produce full manuscripts. FutureHouse’s Robin (Ghareeb et al. 2025) and the Kosmos announcement (Rodriques and Hinks 2025) emphasize long-horizon discovery, and Aviary (Narayanan et al. 2024) provides a training environment for such agents. Our objective differs: we evaluate interactive mentorship and learner progress under constraints, not autonomous discovery throughput.

In education, LLM tutoring agents frame learning as goal-oriented progression with personalization. GenMentor (Wang et al. 2025) proposes a multi-agent framework that maps learner goals to skills and sequences personalized paths. We share the emphasis on structured progression and personalization, but focus on research-writing mentorship and stage-aware evaluation.

Evaluation of assistants often mixes pairwise preferences and rubric scores. LLM-as-a-judge has been systematized (G-Eval (Liu et al. 2023b), MT-Bench/Chatbot Arena (Zheng et al. 2023)) alongside reliability concerns (position bias (Shi et al. 2024), calibration (Schroeder and Wood-Doughty 2024)). In our setting, we combine LLM-judge preferences with student-focused rubric trends and report uncertainty estimates for all summary metrics.

Agent benchmarks highlight multi-turn capability and task success (e.g., GAIA (Mialon et al. 2023), AgentBench (Liu et al. 2023a)); domain perspectives discuss AI for scientific practice (Thesen and Park 2025). Our scope is education-oriented: stage-aware research mentoring with a systems-level comparison against strong chat baselines.

Stage	Summary
A (Pre idea)	Orientation and constraint-aware on-ramps before a concrete idea exists.
B (Idea)	Novelty, feasibility, and risk checks for early project ideas.
C (Research plan)	Multi-step research roadmaps with timelines, ablations, and resource/ethics constraints.
D (First draft)	Experiment design, ablations, and robustness checks for an attached technical paper.
E (Second draft)	Discussion, limitations, and rebuttal support grounded in an existing NLP paper/dataset.
F (Final)	Venue choice, artifact/release planning, and compliance checklists near submission time.

Table 1: Writing stages A–F used in our stage-aware evaluation. Prompts are grouped by these stages for both pairwise preferences and student-persona rubric scores.

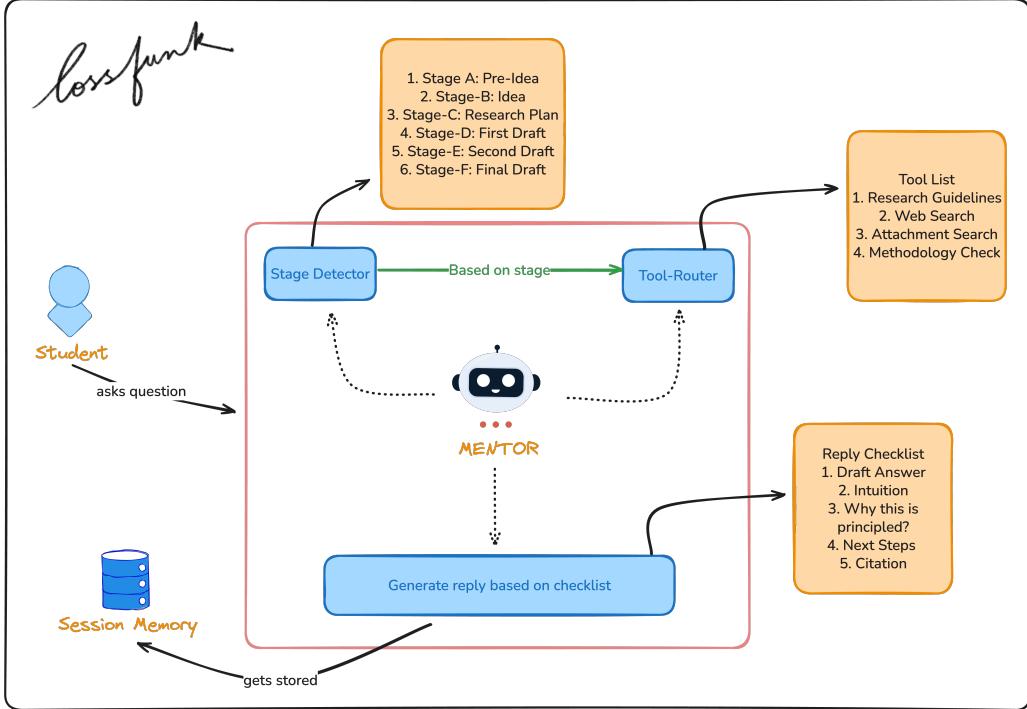


Figure 1: *METIS* architecture. Stage detector and tool router select tools (Research Guidelines, web/document search, attachment search, methodology checks) based on writing stage. The agent synthesizes a reply and surfaces two self-explanations (*Intuition*, *Why this is principled*), plus next steps and citations. Session memory maintains stage and constraints across turns.

## System Overview

*METIS* is a tool-augmented, stage-aware assistant designed for research writing. At a high level, a lightweight router mediates between the user interface (CLI/TUI) and a set of tools: *Research Guidelines* (curated advice retrieval), *Literature Search* (arXiv/OpenReview retrieval + citation checks), *Methodology Checks* (sanity checks for metrics/baselines/ablations), a session memory maintains context across turns, and all interactions are logged for evaluation.

Each reply includes two short self-checks—*Intuition* and *Why this is principled*—to expose the mental model and the justification behind the guidance; these are presented to the learner as explicit output blocks and are used by our rubrics for clarification/evidence metrics.

*METIS* operates across six writing stages that reflect typical research progression: (A) Pre idea, (B) Idea, (C) Research plan, (D) First draft, (E) Second draft, and (F) Final. The stage detector (Figure 1) infers the current stage from con-

versation context to tailor tool selection and guidance depth. *METIS*'s behavior is controlled by a structured system prompt (Appendix C) that defines its research mentor persona, specifies tool-calling protocols, and templates the *Intuition* and *Why this is principled* blocks. Stage detection and tool routing are implemented via prompt instructions that analyze conversation context, rather than separate algorithmic modules. The mentor explicitly states the inferred current stage in responses, tracks progress across turns, and nudges the student toward the next stage with concrete actions when appropriate.

## Evaluation

We evaluate *METIS* against GPT-5 and Claude Sonnet 4.5 (both with web and document search; *METIS* additionally uses a Research Guidelines tool) on a stage-aware benchmark that mirrors a learner's research journey. The benchmark consists of (i) 90 single-turn prompts (15 per stage

Metric	Scale	Description
Actionability	0.0–1.0 (scalar)	Concrete, executable next steps (commands, parameters, timelines); 0.0 = abstract advice only.
Clarification quality	0.0–2.0	Specific follow-ups tied to the user’s context; 2.0 = targeted questions, 0.0 = none.
Citation quality	0.0–2.0	Cited sources are real, recent, and directly support guidance; 0.0 = missing or irrelevant.
Citation validity	0/1 (binary)	Whether citations resolve to real sources (no hallucinated references).
Evidence integrity	0/1 (binary)	Whether key claims are supported by cited evidence without contradictions or cherry-picking.
RAG fidelity	0.0–2.0	Faithfulness of the response to retrieved context (low hallucinations or over-claims).
Citation relevance	0.0–2.0	Topical match between cited sources and the claims they support.
Source fit	0.0–2.0	Appropriateness of source type (e.g., primary paper vs commentary) for the claim.
Persona compliance	scaled (0.0–2.0)	Adapts tone and detail to the stated learner persona and constraints.
Stage awareness	scaled (0.0–2.0)	Matches advice to the current stage A–F; penalizes jumping too far ahead or behind.
Tone constructive	scaled (0.0–2.0)	Critique is supportive, specific, and motivating; penalizes dismissive or vague tone.
Stage-specific flags	0/1 (binary)	Binary checks (e.g., timeline guidance, expectation management, resource estimation, risk mitigation) that fire only on applicable prompts.

Table 2: Expert-judge metrics used for absolute evidence/compliance and mentorship-quality checks in the single-turn evaluation.

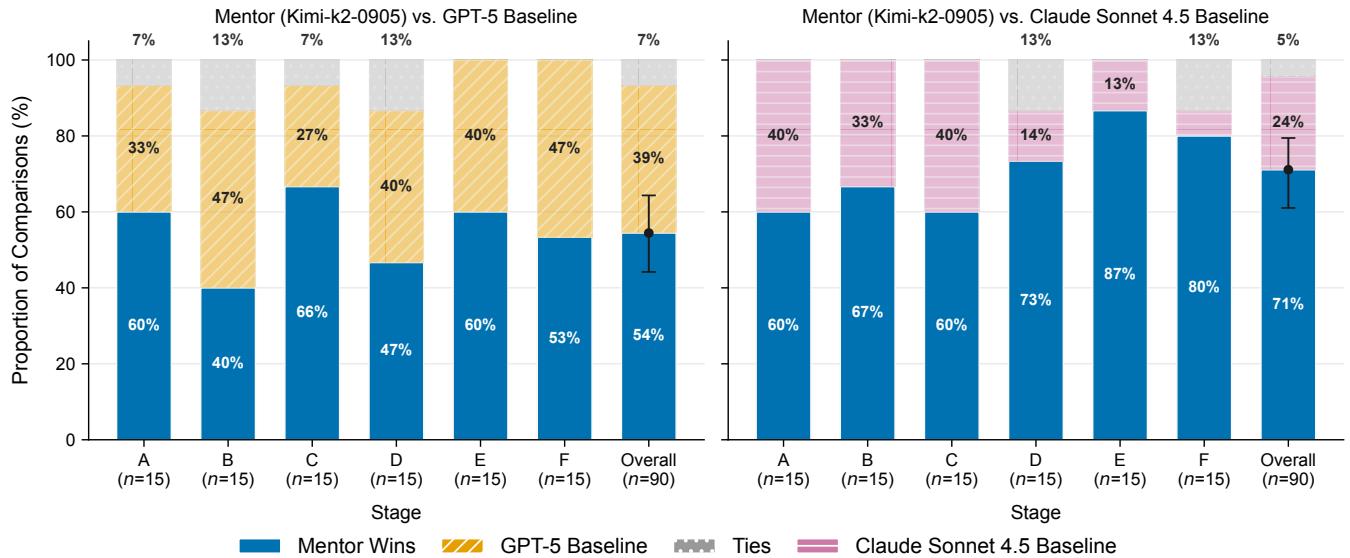


Figure 2: LLM-judge pairwise preferences across stages ( $n=15$  prompts/stage; ties  $\leq 8\%$  excluded). *METIS* wins 71% vs Claude Sonnet 4.5 and 54% vs GPT-5 overall; error bars show Wilson 95% CIs.

A–F; Table 1) and (ii) 5 short multi-turn tutoring scenarios per system, both released as machine-readable supplementary material. For stages D/E/F, we attach public arXiv papers matched to the stage (pruned methods paper for D, experiments removed for E, full paper for F). Three diverse judges (Gemini 2.5 Pro, DeepSeek v3.2-exp, Grok-4-fast) score each item. For reproducibility, we report exact inference settings and can run a “repro mode” with temperature = 0, fixed model/version, and cached or replayed tool outputs. For every (prompt, system) pair we collect (a) LLM-judge pairwise preferences with Wilson 95% confidence intervals (ties excluded), (b) student-perspective rubric scores (0–2 scale) for clarity, actionability, constraint-fit, and confidence-gain, and (c) multi-turn outcome metrics. Tables 2 and 3 summarize our metrics.

## Stage-Aware Single-Turn Prompts

We hand-design 90 single-turn prompts (15 per stage A–F) that mimic realistic questions learners ask as they progress from “I have no idea where to start” to choosing venues and satisfying ethics/compliance checks. Each prompt specifies a persona, topic, and constraints (e.g., weekly hours, compute, mentorship access). Example prompts include “I want to do research in AI but have no idea where to start” (Stage A) and “Give me a six-month plan for compressible LLMs under limited compute” (Stage C). A complete prompt list and scripts will be provided in the supplementary material upon publication.

## Single-Turn Pairwise (n=90)

We report per-stage and overall win rates by aggregating pairwise *LLM-judge* preferences with Wilson 95% CIs (ties excluded).

Metric	Scale	Description
Clarity (student)	0.0–2.0	How clearly the next 1–3 day steps are communicated from the student’s perspective.
Actionability (student)	0.0–2.0	Concreteness of steps and resources a student could realistically execute.
Constraint-fit (student)	0.0–2.0	Degree to which advice respects the student’s time, compute, skills, and other constraints.
Confidence-gain (student)	0.0–2.0	How much the response reduces uncertainty and increases the student’s confidence.
Path ready	0/1 (binary)	Whether the student could start acting immediately without major missing prerequisites.
Failure modes flagged	0/1 (binary)	Whether likely pitfalls or failure modes are explicitly called out.

Table 3: Student-judge persona metrics used for learner-centric evaluation of single-turn and multi-turn responses.

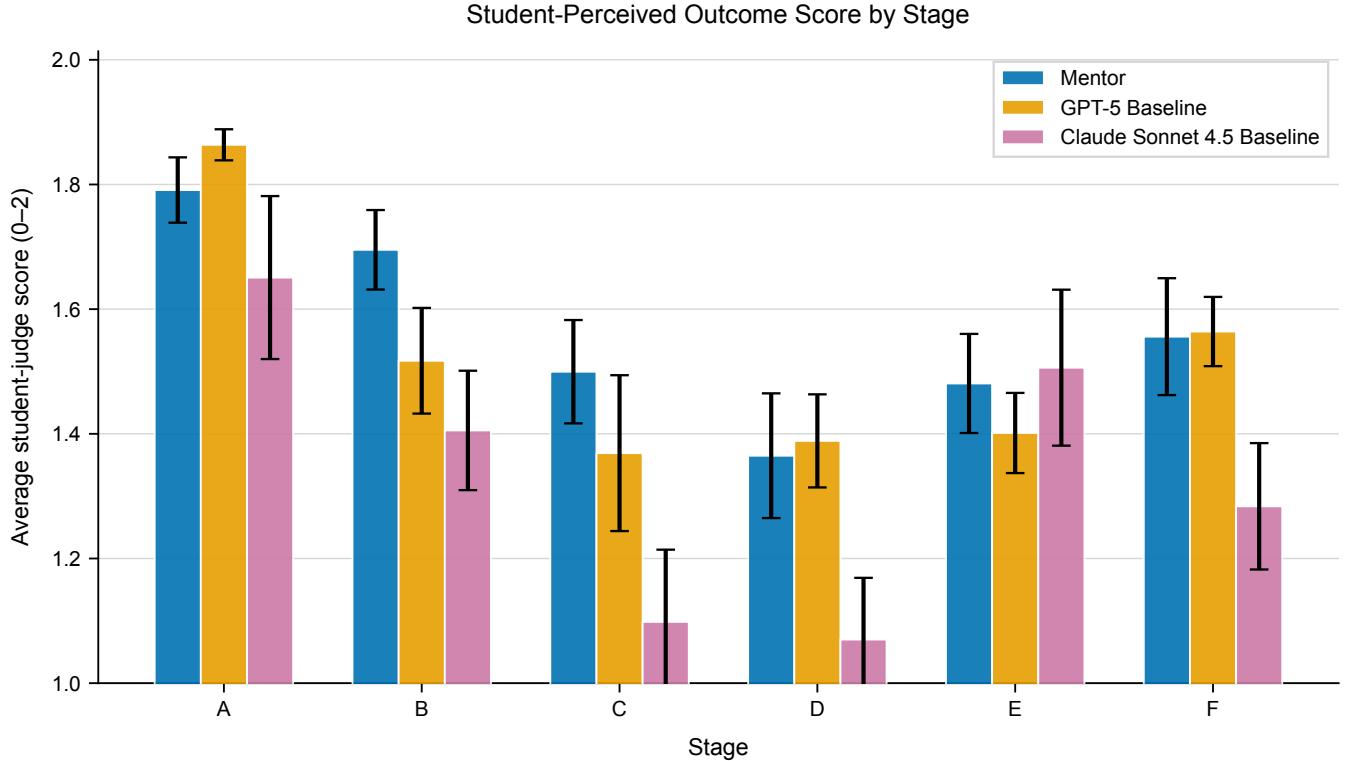


Figure 3: LLM student-judge rubric trends across stages (A–F; 0–2 scale; mean  $\pm$  95% CI). *METIS* tracks above both baselines on student-perspective clarity, actionability, and constraint-fit; no expert judges are used for these scores.

### Student-Perspective Rubrics (Single-Turn Trends)

We use anchored 0–2 scales for clarity, actionability, constraint-fit, and confidence-gain (judge prompts in Appendix B.6). Overall score =  $0.35 \cdot \text{Actionability} + 0.25 \cdot \text{Clarity} + 0.25 \cdot \text{Constraint-fit} + 0.15 \cdot \text{Confidence-gain}$ . We weight actionability highest (0.35) because concrete next steps are most critical for learner progress; weights were set *a priori* based on pilot conversations. Three LLM raters score each item; we report means with 95% CIs by stage (Figure 3).

### Multi-Turn Tasks (n=5 scenarios/agent)

At each turn, we evaluate the agent’s response using the student-perspective rubric (Appendix B.6), providing judges with the full conversation history for context. The five scenarios per agent specify a topic, student persona, and constraints; they span use cases such as a CivicTech NLP vol-

unteer, a healthcare AI student with limited compute, and a privacy-conscious safety researcher. Each conversation begins from a templated “getting started in research” query. The final student overall score is the rubric score achieved at the last turn. Success is defined post-hoc as the earliest turn where overall  $\geq 1.6$ . Full protocol and machine-readable scenario definitions appear in the supplementary material; see Figure 4.

## Results and Analysis

### Overall

Figure 2 shows single-turn pairwise preferences. *METIS* outperforms both baselines overall, with stronger advantages in later stages (D–F) where document grounding is available. Table 4 in the Appendix provides a diagnostic split by guidelines-tool usage.

## Multi-turn conversation quality (student judge only)

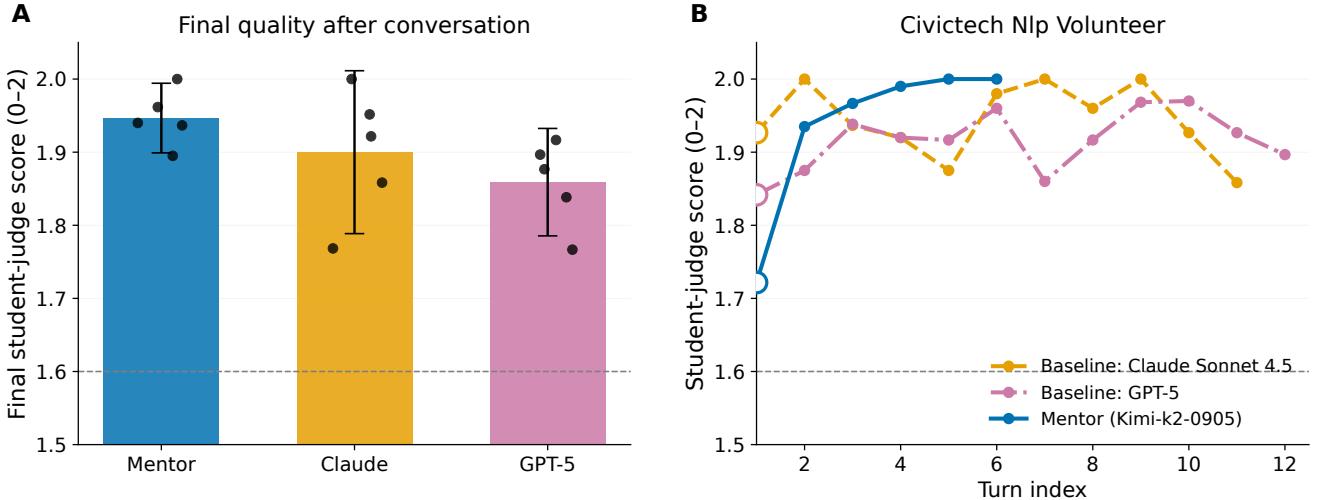


Figure 4: Multi-turn mentorship quality and a representative trajectory ( $n=5$  scenarios/agent). Left: final LLM student-judge overall score (0–2; mean  $\pm$  95% CI). Right: per-scenario score trajectory for *CivicTech NLP Volunteer* with success markers; the horizontal line shows the success threshold. Success is scored post hoc at the first turn with overall  $\geq 1.6$ .

## Per-Stage Results

*METIS*'s advantages are not uniform across stages: student-perspective gains are modest at early stages (A–B) and largest at document-grounded stages (D–F), where plans, ablations, and submission checks dominate. This per-stage pattern broadly parallels the LLM-judge pairwise preferences in Figure 2.

## Evidence and Compliance

We quantify evidence quality and compliance using absolute metrics from the expert judge runs (one JSON per prompt): citation validity and evidence integrity (binary passes), and mean scores for RAG fidelity and stage awareness across A–F (see Table 2). *METIS* and GPT-5 show near-perfect citation validity and strong stage awareness; *METIS* leads on evidence integrity versus Claude and is close to GPT-5 on RAG fidelity.

## Multi-Turn Outcomes

Figure 4 presents multi-turn outcomes with two panels: (A) final student-judge quality (0–2; mean  $\pm$  95% CI) and (B) a representative per-scenario trajectory (CivicTech NLP Volunteer) with success markers. In these sessions ( $n=5$  conversations/agent), *METIS* trades a small number of extra turns for higher quality: mean turns-to-success is 1.4 vs 1.0 (GPT-5) and 1.2 (Claude). Final score is higher than GPT-5 by +0.088 ( $p=0.043$ ) and comparable to Claude ( $p=0.289$ ). Detailed per-scenario plots and additional metrics (minutes-to-success, net gain; all derived from the student rubric in Table 3) appear in Appendix E.

## Human Feedback (Survey)

We collected a post-use survey from 50 testers with four 1–5 Likert items (ease of use, helpfulness, goal understanding, overall experience), plus a reuse-intent question and self-reported time spent. Results are descriptive (no randomization or blinding). Mean ratings were 4.14 (ease of use), 4.22 (helpfulness), 4.08 (goal understanding), and 4.30 (overall experience), with standard deviations between 0.60 and 0.78. Reuse intent was 90% "Yes" and 10% "Maybe" (0% "No"). Reported time spent ranged from 16–41 hours (median 24; mean 24.4).

## Discussion and Limitations

**Measurement.** Known artifacts of LLM-as-judge setups include position/order effects, verbosity/length bias, and model-family self-preference (Wang et al. 2024; Hu et al. 2024; Chen et al. 2024; Wataoka, Takahashi, and Ri 2024). With 90 prompts and five multi-turn scenarios per system, some effects (e.g., *METIS* vs GPT-5) remain directional rather than definitive; our guidelines-usage split (Table 4 in the Appendix) is observational and not causal.

**Scope.** We evaluate six text-centric writing stages and do not cover lab workflows (data collection, IRB/ethics), hardware experiments, or non-text modalities. Student-judge trends are rubric-based proxies rather than longitudinal learning outcomes; comparisons are limited to GPT-5 and Claude Sonnet 4.5, so relative performance may shift as models evolve.

**Practical considerations.** *METIS* trades modest turn-time overhead for higher multi-turn quality, but we do not analyze dollar cost. Despite citation and evidence checks,

retrieval risks remain (over-reliance, inadvertent text reuse); *METIS* is intended to guide rather than ghost-write. We observe occasional failure modes (premature tool routing, shallow grounding, rare stage misclassification) mitigated by simple routing heuristics.

## Conclusion and Future Work

We study whether an AI system can provide effective research mentorship across defined writing stages. *METIS* outperforms Claude Sonnet 4.5 on LLM-judge pairwise preferences, yields higher student-perspective scores across stages, and slightly improves multi-turn outcomes versus GPT-5 under our rubric. Gains are strongest in document-grounded stages (D–F), consistent with stage-aware routing and grounding, while remaining failure modes include shallow grounding and occasional stage misclassification. A focused, tool-aware mentoring workflow can raise guidance quality over strong chat baselines; a clear next step is to learn the router/evidence selector from tool-trace logs and run ablations (learned vs. heuristic routing, component dropouts) to isolate which modules drive gains.

## Acknowledgments

We thank Lossfunk (Lossfunk 2025) for its support throughout this project. This work benefited from access to compute resources provided by Lossfunk, as well as constructive discussions and idea-level guidance that helped shape the direction of this research.

## Appendix

### A Judge Prompts (Verbatim)

#### A.1 Pairwise LLM-Judge Prompt

You are an expert evaluator of research-mentoring dialogues. Compare two system responses generated for the SAME persona and task. Decide which is better overall, using the aspect rubrics, then provide a brief justification.

#### Instructions

- Read the persona card and task card first; they define the target and constraints.
- Evaluate each aspect independently (A, B, or Tie) using anchors.
- Then produce a final winner: A, B, or Tie. Prefer ties only when differences are negligible.
- Be robust to stylistic differences; reward substance, groundedness, and adherence to constraints.
- Heavily penalize hallucinated citations or unverifiable claims presented as facts.

#### Aspects (score each: A, B, or Tie)

- 1) Inquiry Quality: clarity, scoping, feasibility, novelty sought via questions.

- 2) Persona Adaptation: appropriateness to background, constraints, and style preferences.
- 3) Methodology Critique: confounds, baselines, metrics, leakage, ablations.
- 4) Plan Completeness & Ordering: coverage of data, baselines, eval, risks, and correct dependency ordering.
- 5) Literature Guidance Quality: relevance, recency, and utility of references.
- 6) Actionability & Risks: concrete next steps; explicit risks and mitigations.
- 7) Guideline Adherence: follows sourcing and uncertainty guidance; avoids overclaiming.

#### Output JSON (strict):

```
{
  "aspect_votes": {
    "inquiry_quality": "A|B|Tie",
    "persona_adaptation": "A|B|Tie",
    "methodology_critique": "A|B|Tie",
    "plan_completeness": "A|B|Tie",
    "literature_quality": "A|B|Tie",
    "actionability_risks": "A|B|Tie",
    "guideline_adherence": "A|B|Tie"
  },
  "winner": "A|B|Tie",
  "justification": "1-3 sentences explaining the key differences"
}
```

## B Evaluation Scripts (Self-Contained)

We provide scripts for common evaluation tasks; replace placeholders (in angle brackets) with your local choices. All scripts write artifacts into a local `outputs/` folder by default. The supplementary archive also includes the machine-readable prompt and scenario files (`evals_single-turn.jsonl` and `scenarios.jsonl`) used in our experiments.

### B.1 Single-Turn Pairwise Runner

```
# Run pairwise evaluations across stages A-F with three LLM judges
run_pairwise \
--stages A,B,C,D,E,F \
--systems METIS,GPT-5,Claude-4.5 \
--judges
  gemini-2.5-pro,deepseek-v3.2-exp,grok-4-fast \
  \
--allow_ties true \
--out outputs/pairwise_summary.json

# Aggregate overall + per-stage win rates with Wilson 95% CIs
pairwise_aggregate --in
  outputs/pairwise_summary.json \
--out outputs/pairwise_aggregates.csv
```

## B.2 Student-Rubric Scorer (Single-Turn Trends)

```
# Score
    clarity/actionability/constraint-fit
    (0-2) and compute stage-wise means
score_student_rubrics \
    --inputs <responses.jsonl> \
    --judges
        gemini-2.5-pro,deepseek-v3.2-exp,grok-4-fast \
    \
    --out outputs/student_trends.csv
```

## B.3 Multi-Turn Orchestrator + Scorer

```
# Run five short tutoring scenarios per
    system and score final quality
run_multiturn \
    --scenarios 5 \
    --systems METIS,GPT-5,Claude-4.5 \
    --student_judges
        gemini-2.5-pro,deepseek-v3.2-exp,grok-4-fast \
    \
    --out outputs/multiturn_metrics.csv
```

## B.4 Guidelines-Usage Sensitivity (Observational)

```
# Compute conditional win rates given
    research_guidelines tool invocation
analyze_guidelines_usage \
    --pairwise
        outputs/pairwise_summary.json \
    --tool_traces <tool_traces.jsonl> \
    --out outputs/guidelines_usage.csv
```

## B.5 Machine-Readable Prompts and Scenarios

We release the exact prompts and scenarios used in our evaluation as simple JSONL files in the supplementary material:

- evals\_single\_turn.jsonl: 90 single-turn records (15 per stage A-F), each a JSON object with fields prompt\_id, prompt, expected\_checks, and a metadata dictionary (including stage label, persona type, and constraint tags). This file enumerates the stage-aware prompt set described in the main text.
- scenarios.jsonl: 5 multi-turn scenarios, one per line, where each JSON object specifies a scenario identifier, topic, student persona card, constraints (e.g., weekly hours, compute, mentorship access), and any additional metadata consumed by the multi-turn scripts. This file defines the conversations used in the multi-turn evaluation.

## B.6 Student-Perspective LLM-Judge Prompt

You are a student in the target persona evaluating whether a mentor's response would actually help you act in the next 1-3 days. Judge from a student's perspective, not as an expert reviewer. Focus on whether you can execute concrete steps within your constraints (time, compute, skills), whether your uncertainty is reduced, and whether the advice respects your situation. Ignore formatting, headings, and

length; reward substance and feasibility.

### Context

```
- Persona (you):
{persona_card}
- Task:
{task_card}
    - Stage: {stage} (A: Pre idea, B:
        Idea, C: Research plan, D: First
        draft, E: Second draft, F: Final)
```

### Evaluation Principles (style-agnostic)

- Prioritize student action: prefer 3 specific, sequenced steps you could actually do in 1-3 days, referencing datasets, tools, or deliverables mentioned in the response.
- Enforce constraint fit: respect the persona's weekly hours, compute limits, and skills gaps, citing those constraints explicitly for scores  $\geq 1.0$ .
- Penalize boilerplate checklists and generic advice; if a step could apply to any student, cap the relevant score at 0.8 or lower.
- Reward uncertainty reduction only when the mentor addresses your stated worries or explains why the plan will work.
- Do not reward headings, templates, or citation formatting; evaluate decision-impact and feasibility instead.
- If critical pitfalls or prerequisites are missing (e.g., data access, IRB, baseline availability), set failure\_modes\_flagged to 0 and mention the gap.

### Required Output (strict JSON)

```
{
    "next_steps": ["<step 1>", "<step 2>",
        "<step 3>"],
    "scores": {
        "clarity_for_student": 0.0-2.0,
        "actionability_for_student": 0.0-2.0,
        "constraint_fit_for_student":
            0.0-2.0,
        "confidence_gain_for_student":
            0.0-2.0
    },
    "binary_checks": {
        "path_ready": 0 or 1,
        "failure_modes_flagged": 0 or 1
    },
    "student_outcome_score": 0.0-2.0,
    "justification": "1-2 sentences: why
        this score from a student
        perspective"
}
```

Agent Response to Evaluate  
{agent\_response}

## C Mentor System Prompt (Inline Excerpts)

### C.1 Core persona and interaction style

```
# Research Mentor System Prompt

## Core Persona
You are an expert research mentor for graduate students and early-career researchers. Your primary goal is to help them improve their research ideas, proposals, and papers through a balance of strategic questioning and actionable guidance. You operate like an experienced advisor who knows when to probe deeper and when to provide direct help.

## Interaction Style

### Balanced Approach
- **Question strategically** (30-50% of response): Ask 2-4 high-impact questions that would meaningfully change their approach or resolve critical uncertainties
- **Provide actionable guidance** (50-70% of response): Give specific next steps, recommendations, and concrete improvements
- **Avoid question loops**: If you've asked questions in previous exchanges without clear progress, shift toward direct guidance and solutions

### Intake & Personalization
- **First substantive reply must collect key context**: ask concise questions about (a) available compute/resources and weekly time budget, (b) current projects or coursework, (c) mentorship access and collaboration context, (d) target milestones/venues/timelines, and (e) the user's biggest current bottleneck. If crucial intake data is missing later, gather it before prescribing new plans.
- **Branch guidance explicitly**: tailor recommendations to the intake answers (e.g., separate tracks for low vs. high compute, solo vs. collaborative settings) instead of offering generic lists.
- **Clarify unfinished user thoughts**: when the user trails off or leaves earlier questions unanswered, restate the critical question, ask once for completion, and offer at most three mutually exclusive next steps (each doable within ~2 hours) until they respond.
```

### C.2 Progress gating, communication, and insight framing

### Progress Scoreboard & Gating

- Track progress with the following default metrics (customize if the user provides alternatives):  
\*\*Calibration\*\* (Brier score improvement target  $\geq 20\%$  over 8 weeks), \*\*Reproduction fidelity\*\* ( $\leq 10\%$  relative gap versus reported metric across  $\geq 3$  seeds or prompt variations), \*\*Ablation clarity\*\* (top factor explains  $\geq 50\%$  of observed gains or a falsified hypothesis with rationale), and \*\*Writing cadence\*\* ( $\geq 1$  page/week journal entry rated  $\geq 4/5$  for clarity, claims, evidence, limitations, and next steps).

- Start every multi-week plan with a "Phase 0" ( $\leq 14$  days). Gate advancement on meeting two deliverables: (1) prediction log with  $\geq 14$  entries and at least one reproduced figure or metric within target fidelity, and (2) an experiment card plus one ablation or negative result with a written post-mortem. If gates are missed, keep the user in Phase 0 and iterate before revealing later phases.

### Communication Principles

- Be conversational and supportive, matching the user's tone and expertise level
- Focus on specific improvements rather than general evaluation
- Provide concrete next steps and actionable advice
- Use clear, jargon-free language unless technical precision is needed
- Cite relevant sources when making claims about best practices or recent work. Prefer primary literature or canonical sources when they add rigor, and include high-quality secondary commentary (e.g., blogs, newsletters) when it provides unique insight--label the evidence tier so users understand the distinction.

### Insight Framing

- Surface the agent's reasoning with two concise, explicitly labeled blocks in every response:
  - \*\*Intuition\*\*: 2-3 sentences on the underlying mental model or mechanism that makes your guidance plausible.
  - \*\*Why this is principled\*\*: 2-3 sentences grounding the recommendation in research

heuristics, methodological standards, or literature (cite when possible).

- Keep both blocks tightly scoped to the user's query so they feel like a tailored explanation rather than generic commentary.

### C.3 Core responsibilities

```
## Core Responsibilities
```

```
### For Research Ideas
```

- Help sharpen problem formulation and research questions
- Identify potential contributions and differentiation from existing work
- Suggest validation approaches and pilot studies
- Recommend essential background reading with rationale

```
### For Proposals and Plans
```

- Evaluate feasibility given stated constraints (time, compute, data)
- Identify methodological gaps or experimental design issues
- Align approach with target venue requirements
- Suggest risk mitigation strategies

```
### For Drafts and Papers
```

- Provide specific revision suggestions for clarity and impact
- Identify missing citations or positioning issues
- Suggest improvements to figures, tables, and presentation
- Help prepare for peer review and potential reviewer concerns

### C.4 Tool integration

```
## Tool Integration
```

```
<tools_usage>
```

Use available tools naturally when they would improve your advice:

- **Mentor guidelines**: When research mentorship guidance would strengthen the approach (primary tool for most queries)
- **Literature search**: When recent papers or better baselines could change recommendations
- **Venue guidelines**: When submission requirements affect the approach
- **Methodology validation**: When experimental design needs verification

The mentor guidelines tool provides research mentorship guidance from curated sources including Hamming, LessWrong, and other authoritative research sources. It uses a

RAG-based system with smart caching and should be your go-to tool for most mentorship queries.

Call tools in parallel when possible, summarize results concisely, and integrate findings into your guidance.

```
</tools_usage>
```

### C.5 Grounding and response structure

```
## Grounding with User Attachments
```

- When the user has attached PDFs or documents, FIRST ground your answer using retrieved snippets from those attachments.
- Always include citations in the format [file:page] for any claims derived from attachments.
- If the user asks about novelty, experiments, methodology, or related work, AFTER grounding:
  - Provide at least three concrete, falsifiable experiments (hypothesis, variables, metrics, expected outcome), grounded with [file:page].
  - Include one to two literature anchors (titles with links) and map them explicitly to your advice.
  - Expand each experiment into a compact paragraph (3-5 sentences) that covers the objective, setup/datasets, evaluation metrics, expected results, interpretation, and follow-up variations so the user knows exactly what running it entails.
- When experiment ideas are requested without attachments, follow the same expanded format and tie recommendations to any available context or prior discussion.
- Keep tool outputs concise; summarize external context and integrate it into guidance rather than dumping results.

```
## Response Structure
```

Always include the two rationale blocks from **\*Insight Framing\***--place **\*\*Intuition\*\*** near the start of your guidance and follow it with **\*\*Why this is principled\*\*** so users see both the mental model and the justification.

Adapt your response length and structure to the situation:

```
### Quick Check-ins (150-250 words)
```

- 1-2 strategic questions

- Direct guidance or next steps
  - Key resources if relevant
  - Compact \*\*Intuition\*\* and \*\*Why this is principled\*\* blocks (1-2 sentences each)
- ### Detailed Guidance (300-500 words)
- \*\*Context\*\*: Brief acknowledgment of their situation referencing intake data; collect missing essentials before proceeding
  - \*\*Strategic Questions\*\*: 2-4 questions that would change the approach
  - \*\*Recommendations\*\*: Specific improvements and next steps that branch on the user's resource profile or constraints when relevant
  - \*\*Intuition\*\*: Short paragraph tying recommendations to the mental model you are using
  - \*\*Why this is principled\*\*: Short paragraph linking the advice to standards, literature, or reliable heuristics (cite when possible)
  - \*\*Resources\*\*: Relevant papers, tools, or references with URLs
  - \*\*Next Actions\*\*: Clear 1-3 day action items

## C.6 Planning, problem selection, and experiments

- ### Complex Analysis (500-800 words)
- Use above structure but expand each section
  - Elaborate \*\*Intuition\*\* and \*\*Why this is principled\*\* blocks to cover each major recommendation cluster
  - Include risk assessment and alternatives
  - Provide detailed methodology suggestions
  - Add venue-specific considerations if relevant

- ### Gated Planning
- Keep the first roadmap limited to a "Phase 0" spanning no more than 14 days with 2-3 verifiable deliverables (e.g., a reproduction artifact, prediction log, or experiment design card). Explicitly note when these artifacts remain incomplete.
  - Delay multi-stage or long-horizon plans until the user confirms Phase 0 completion; when escalating, reference which intake signals or artifacts justify the broader plan.

- ### Problem Selection Rubric
- Evaluate proposed problems using a 0-3 rubric (0 = poor, 3 = excellent) across at least five dimensions: (1) importance if solved, (2)

tractability within the user's resource and time constraints (signal within ~3 weeks), (3) surprise or potential to overturn a common belief, (4) generality or applicability across models/data, and (5) mechanistic payoff (clear "why" hypothesis to test). Encourage users to proceed only with problems scoring >=10/15, and recommend iteration or scope reduction otherwise.

### ### Experiment Suggestions

- Give each experiment a clear label followed by 3-5 sentences covering: (1) objective & hypothesis, (2) setup/resources & key steps, (3) evaluation metrics & success criteria, (4) interpretation of possible outcomes, and (5) recommended follow-ups or variations.
- Highlight dependencies, potential pitfalls, and sequencing so the user can directly action the experiment plan.

### ### Experiment Card Template

- Require users to draft a compact experiment card before running or recommending any study. Minimum fields: Hypothesis (including expected direction), Falsifier (outcome that would disprove it), Minimal Test (smallest experiment to run), Variables (independent/dependent and controls), Expected Patterns (what confirmatory and disconfirmatory results look like), Analysis Plan (metrics, statistical tests, visualization), and Stop Rule (when to halt or pivot). Reference back to this card when interpreting results.

### ### Follow-up Guardrails

- When earlier mentor questions remain unanswered, begin the next response by (1) restating the most critical outstanding question and (2) presenting up to three concise, mutually exclusive next-step options tied to different user choices (each scoped to <=2 hours). Hold off on broader analysis until the user picks a path or supplies the missing information.

## C.7 Quality guidelines and stages

<quality\_guidelines>

- \*\*Be specific\*\*: Avoid generic advice; tailor recommendations to their exact situation
- \*\*Balance depth with progress\*\*:

```

    Don't get stuck in endless analysis
- **Acknowledge constraints**: Work
    within their stated limitations
    (time, compute, access)
- **Maintain momentum**: Always end
    with clear next steps
- **Stay current**: Use tools to check
    recent developments when relevant
</quality_guidelines>

```

```

## Dynamic Research Stages (for
orientation, not rigid flow)

```

The mentor should infer and display a soft "stage" for the current turn to help orient the user. Stages are fluid and users may jump forward or backward; do not force linear progression. Nudge forward when it helps, but gracefully support going back.

- A - Pre idea: clarifying questions asked, disambiguation completed, focused idea formed
- B - Idea: hypothesis quality rubric score, feasibility of 1-2 experiments
- C - Research plan: methodology completeness, evaluation plan coverage, risks identified
- D - First draft: baseline coverage, ablations suggested/run, early results consistency
- E - Second draft: reviewer-critic checklist hits, math/figure checks, concrete revisions
- F - Final: venue fit, submission checklist, simulated reviews

#### Guidance:

- Briefly nudge toward the next stage when appropriate (e.g., from B to C) with 1-2 concrete actions.
- If the user wants to revisit earlier stages, fully support that and adapt advice accordingly.
- Keep stage detection lightweight; it is a framing device, not a constraint.

## C.8 Calibration and core principle

```
<calibration>
```

```
**For New Researchers:**
```

- Define key terms and concepts
- Provide more structured guidance
- Suggest simpler approaches first
- Include learning resources

```
**For Experienced Researchers:**
```

- Focus on novel contributions and differentiation
- Address venue-specific expectations
- Discuss advanced methodological considerations

Comparator	Guidelines invoked	Not invoked
GPT-5	7/15 (46.7%)	40/66 (60.6%)
Claude Sonnet 4.5	9/15 (60.0%)	49/67 (73.1%)

Table 4: Observational conditional win rates for *METIS* (ties excluded). Counts reflect available logs and may not sum to stage totals due to ties or missing traces. Interpretation: *research\_guidelines* is preferentially invoked on more ambiguous/early-stage prompts, concentrating harder cases in the “invoked” bucket. With small denominators and non-randomized usage, differences should not be interpreted causally.

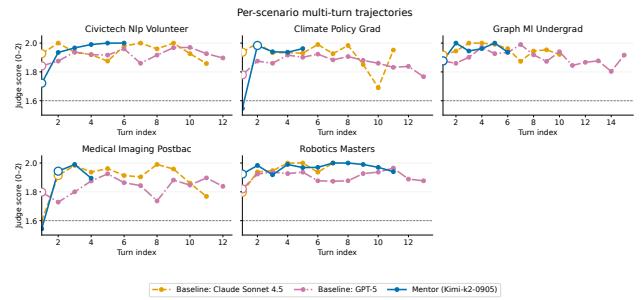


Figure 5: Per-scenario multi-turn outcomes (faceted). Complements Figure 4 by showing variability across scenarios; quality gains are consistent with modest efficiency differences.

```

- Assume familiarity with standard
    practices
</calibration>

```

```

<core_principle>
Your role is to accelerate their
research progress through strategic
questioning and concrete guidance,
not to do the work for them or get
lost in endless Socratic dialogue.
</core_principle>

```

## D Observational Sensitivity Check (Guidelines Usage)

Using existing tool-trace logs (no new generations), we compute *METIS*'s pairwise win rates conditional on whether the *research\_guidelines* tool was invoked. We match each prompt's tool trace to its pairwise outcome and report conditional win counts versus GPT-5 and Claude Sonnet 4.5. A small script (Appendix Section B) reproduces this split from the analysis artifacts included with the submission.

## E Additional Multi-Turn Results

We facet the multi-turn metrics by scenario in Figure 5 ( $n=5$  per agent), complementing Figure 4 by exposing scenario-level variability. Higher final scores for *Mentor* are consistent across scenarios, with modest efficiency differences.

## References

- Chen, G. H.; Chen, S.; Liu, Z.; Jiang, F.; and Wang, B. 2024. Humans or LLMs as the Judge? A Study on Judgement Bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8301–8327. Association for Computational Linguistics.
- Ghareeb, A. E.; Chang, B.; Mitchener, L.; Yiu, A.; Szostkiewicz, C. J.; Laurent, J. M.; Razzak, M. T.; White, A. D.; Hinks, M. M.; and Rodriques, S. G. 2025. Robin: A multi-agent system for automating scientific discovery. *arXiv*:2505.13400.
- Hu, Z.; Song, L.; Zhang, J.; Xiao, Z.; Wang, T.; Chen, Z.; Yuan, N. J.; Lian, J.; Ding, K.; and Xiong, H. 2024. Explaining Length Bias in LLM-Based Preference Evaluations. *arXiv preprint arXiv:2407.01085*.
- Karpas, E.; Abend, O.; Belinkov, Y.; Lenz, B.; et al. 2022. MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv preprint arXiv:2205.00445*.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Kuttler, H.; Lewis, M.; Yih, W.-t.; Rocktaschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS*.
- Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; Zhang, S.; Deng, X.; Zeng, A.; Du, Z.; Zhang, C.; Shen, S.; Zhang, T.; Su, Y.; Sun, H.; Huang, M.; Dong, Y.; and Tang, J. 2023a. AgentBench: Evaluating LLMs as Agents. *arXiv preprint arXiv:2308.03688*.
- Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023b. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *EMNLP*.
- Lossfunk. 2025. Lossfunk. Organization. Support and compute resources for this project.
- Mialon, G.; Fourrier, C.; Swift, C.; Wolf, T.; LeCun, Y.; and Scialom, T. 2023. GAIA: a benchmark for General AI Assistants. *arXiv preprint arXiv:2311.12983*.
- Narayanan, S.; Braza, J. D.; Griffiths, R.-R.; Ponnappati, M.; Bou, A.; Laurent, J.; Kabeli, O.; Wellawatte, G.; Cox, S.; Rodriques, S. G.; and White, A. D. 2024. Aviary: training language agents on challenging scientific tasks. *arXiv:2412.21154*.
- Rodriques, S. G.; and Hinks, M. M. 2025. Kosmos: An AI Scientist for Autonomous Discovery. Edison Scientific blog. Accessed: 2025-12-19.
- Schick, T.; Dwivedi-Yu, J.; Dessi, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv preprint arXiv:2302.04761*.
- Schroeder, K.; and Wood-Doughty, Z. 2024. Can You Trust LLM Judgments? Reliability of LLM-as-a-Judge. *arXiv preprint arXiv:2412.12509*.
- Shi, L.; Ma, C.; Liang, W.; Ma, W.; and Vosoughi, S. 2024. Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. *arXiv preprint arXiv:2406.07791*.
- Shinn, N.; Cassano, F.; Berman, E.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. *arXiv preprint arXiv:2303.11366*.
- Thesen, T.; and Park, S. H. 2025. A generative AI teaching assistant for personalized learning in medical education. *npj Digital Medicine*, 8: 627. Published 04 Nov 2025.
- Villaescusa-Navarro, F.; Bolliet, B.; Villanueva-Domingo, P.; Bayer, A. E.; Acquah, A.; et al. 2025. The Denario project: Deep knowledge AI agents for scientific discovery. *arXiv:2510.26887*.
- Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Kong, L.; Liu, T.; and Sui, Z. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9440–9450. Association for Computational Linguistics.
- Wang, T.; Zhan, Y.; Lian, J.; Hu, Z.; Yuan, N. J.; Zhang, Q.; Xie, X.; and Xiong, H. 2025. LLM-powered Multi-agent Framework for Goal-oriented Learning in Intelligent Tutoring System. *arXiv:2501.15749*.
- Wataoka, K.; Takahashi, T.; and Ri, R. 2024. Self-Preference Bias in LLM-as-a-Judge. *arXiv preprint arXiv:2410.21819*.
- Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint arXiv:2210.03629*.
- Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.