AI's Capability in Assisting Scientific Research in Physics, Astrophysics, and Cosmology II: Project Planning and Proposal Evaluation

Jia Liu[1,2],* Veena Krishnaraj[3], Kateryna Vovk[1,2], Kosuke Aizawa[4,1,2], Adrian E. Bayer[5,3], Linda Blot[1,2], Jessica Cowell[6,1,2], Suyog Garg[7], Jonathan Grée[8], Anamaria Hell[1,2], Ben Horowitz[1,2], Masaya Ichikawa[9], Kanyuni Iemoto[10], Keigo Kondo[11,7], Zacharie Lorsin[8], Kevin McCarthy[1,2], Jamie Robinson[3], Miguel Ruiz-Granda[12,13], Leander Thiele[1,2], Ievgen Vovk[14], and Mingshen Zhou[15,16]

(Author affiliations are listed at the end of the paper.)

We investigate how well large language models (LLMs) can assist scientific project planning and proposal evaluation. One-page project plans were independently generated for eight expert-conceived research projects in physics, astrophysics, and cosmology by human researchers and three contemporary LLMs (ChatGPT, Claude, and DeepSeek; mid-2025 models, used with their default tool access). The resulting 32 proposals were blindly evaluated by four human reviewers and two newer frontier LLMs (Claude Opus 4.8 and ChatGPT Pro 5.5) using a four-aspect evaluation rubric. Reviewers were also asked to identify whether each proposal was written by a human or an AI. Human reviewers rated human- and AI-written proposals similarly overall, whereas both AI reviewers scored AI-written proposals about one point higher (on a five-point scale) than human-written proposals. Human reviewers correctly identified human- and AI-written proposals 72% and 79% of the time, respectively, while both AI reviewers correctly classified all 32 proposals (100%). These results suggest that current LLMs can produce project plans comparable to human-written ones in the eyes of human reviewers, but that AI reviewers show a systematic preference for AI-generated proposals. Our results suggest caution when deploying LLMs widely in proposal preparation and evaluation.

I. INTRODUCTION

Artificial intelligence (AI), and in particular large language models (LLMs), is increasingly being used at every stage of the scientific process, from literature review and ideation to coding, analysis, and writing. Yet most assessments of AI capability rely on benchmarks whose answers are already known, such as standardized exams or curated questions, which measure how well a model recovers existing knowledge rather than how well it operates at the research frontier, where neither humans nor machines yet know the answer. In physics, astrophysics, and cosmology, deep learning has been widely applied to accelerate computation and achieve more accurate parameter inferences (e.g., [1, 2]). More recently, the rise of LLMs has extended AI from numerical tasks to language- or reasoning-related ones [3–19]. Notably, [10] found that LLMs can generate research ideas that expert reviewers rate as more novel than human-generated ones; [12] found that LLMs can provide feedback on papers that researchers often find useful; [8] reported low inter-reviewer agreement together with quality-rating inflation by AI (also see [9, 11]). What remains comparatively unexplored is a controlled, blinded comparison of human and AI performance on an open-ended research-planning task, evaluated by both human and AI reviewers, with attention to evaluator bias. In a companion paper, we examine AI's capability in literature review.

This work is motivated by three questions. First, how capable is AI at frontier-science project planning, in a regime where there is no known answer? Second, and more exploratorily, can humans and AI identify whether a piece of scientific text was written by a human or by an AI, and what reasoning underlies those judgments? Third, how do humans and AI evaluate research plans, and do their evaluations carry systematic biases? The last question speaks directly to the future of funding agencies and to how researchers should prepare proposals in an era when both authorship and review may be AI-assisted.

To address these questions, we conducted a controlled study on eight expert-conceived research projects spanning physics, astrophysics, and cosmology. For each project, a human researcher and three AI assistants independently wrote a one-page proposal from the same title, background, and goal, which were provided verbatim as the common starting point. All 32 proposals were then assessed blindly by human reviewers and LLMs. Three findings stand out: (1) humans identified authorship correctly $\sim$ 70% of the time, whereas the most capable AI reviewers did so essentially perfectly; (2) human reviewers judged AI-written proposals to be no worse than expert-written ones; yet (3) the AI reviewers systematically scored AI-written proposals above human-written ones, a pro-AI bias absent from the human panel. These results carry concrete implications for AI-assisted proposal writing and for the design of review processes as AI is increasingly used in both proposal writing and review.

The remainder of this paper is organized as follows. Section II describes the methodology: the eight research projects, the four groups and their roles, the proposal-writing protocol, and the blind-evaluation procedure and rubric. Section III presents the results on origin identifi-

* jia.liu@ipmu.jp

cation and on proposal quality, for both human and AI reviewers. Section IV summarizes our findings and discusses their implications for AI-assisted research planning and proposal review.

II. METHODS

This work addresses the project-planning stage of the scientific workflow: drafting a research proposal that specifies the scientific goal, the methodology, and the resources and timeline required to carry a project through to completion. To this end, we assembled a controlled corpus of proposals for eight expert-conceived research projects—each prepared independently by a human expert and by three AI assistants, under an identical template, prompt, and rubric—and collected blind quality ratings and human-versus-AI origin judgments from both human and AI reviewers.

The study is organized around four groups—human project planners (the experts), AI prompters, human reviewers, and AI reviewers—whose roles are detailed below. Several design choices were made to isolate proposal content from superficial cues: (1) every proposal followed an identical one-page template and section structure (Sec. II C); (2) human-written proposals were passed through a light grammar-and-spelling pass to remove obvious human mistakes (Sec. II C); and (3) all proposals were anonymized and uniformly formatted before review (Sec. II D). The proposals were deliberately kept much shorter than a typical grant proposal (roughly one page) primarily to keep the workload manageable for the experts who designed the projects and for the reviewers who scored every submission. In total, the eight projects yielded $8 \times 4 = 32$ proposals (8 human, 24 AI), each scored by four human reviewers and two AI reviewers.

A. The eight research projects

The eight projects span theoretical, computational, observational, and instrumentation work across cosmology, astrophysics, and physics. Each was conceived and outlined by a domain expert without any AI assistance; the expert provided a project title, a background paragraph, and a goal paragraph (Table I summarizes the set). These three elements were the common starting point given to both the human planner and the AI prompter for each project. The full titles, backgrounds, and goals for all eight projects are given in Appendix A and match those used in the companion Paper I; Table I serves here as a quick reference.

B. Participant roles

The study involved four groups.

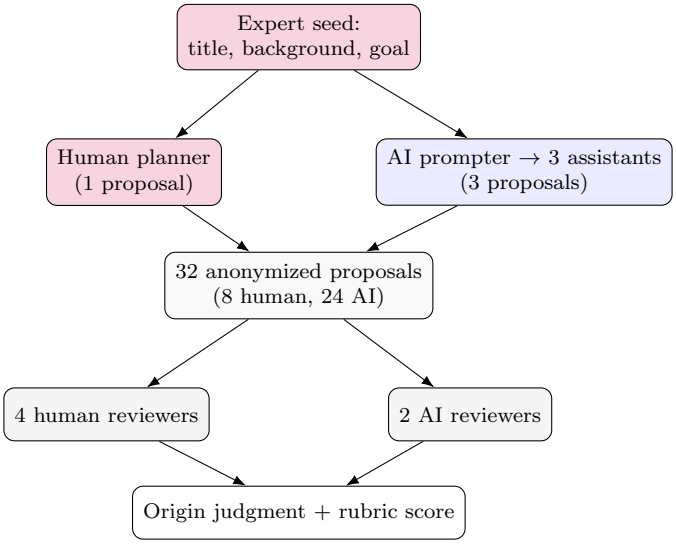

FIG. 1. Overview of the study design. Each expert-provided seed (title, background, goal) is turned into one human-written and three AI-written one-page proposals; the 32 anonymized proposals are then assessed for authorship and quality by four human and two AI reviewers.

Human project planners (experts). For each of the eight projects, one human expert (graduate student or postdoc who works on that topic as part of their routine research) wrote the proposal without any AI assistance. The expert-provided title, background, and goal were the starting point for every proposal, both human and AI. To keep the human and AI proposals format-consistent, the experts followed a shared proposal template and instructions (Sec. II C).

AI prompters. For each project, one AI prompter generated the AI proposals. Prompters were undergraduate or graduate students working outside the specific field of the project but within the broader area of physics, astrophysics, and cosmology. Each prompter took the expert-provided title, background, and goal and prompted three AI assistants with a fixed prompt template (Sec. II C) to produce three AI proposals per project. Each proposal came from a single pass of the fixed prompt template, and the output was taken as-is, without iterative refinement.

Human reviewers. Four of this paper's authors served as blind reviewers. They are faculty members or senior postdocs who, while not necessarily experts on the specific topic of any given proposal, have sufficient knowledge to judge the projects, and as such mirror the composition of a typical grant review panel. They received the full set of anonymized proposals and, for each one, first judged whether it had been written by a human or an AI and then scored it against the predefined rubric (Sec. II D).

AI reviewers. Two AI assistants were given the identical anonymized proposals and rubric and performed the same task as the human reviewers: an origin judgment

ID	Project title	Domain
AGN	MaNGA: AGN Duty Cycle	AGN / galaxy evolution
LBG	The galaxy–dark matter halo connection of Lyman-break galaxies	Large-scale structure
IA	Intrinsic alignments in varying environments	Weak lensing / intrinsic alignments
AR	Prediction of debris emergence on laser-ablated sub-wavelength shapes for millimeter-wave anti-reflection	Instrumentation / laser fabrication
RG	Radio Galaxies with HalfDome	CMB foregrounds / radio galaxies
GW	Studying the environment of gravitational-wave black hole binaries with weak lensing maps	Gravitational waves / multi-messenger
PTA	Forecasting the pulsar timing array sensitivity needed to measure deviations from general relativity	Gravitational waves / tests of GR
SU2	Massive Yang–Mills theory (SU(2))	Theoretical cosmology / inflation

TABLE I. The eight expert-conceived research projects used in this study. Titles are as written by the human experts. Each project was independently turned into one human-written and three AI-written one-page proposals.

(human vs. AI) followed by rubric scores (Sec. II D). We used Claude Opus 4.8 (Anthropic) and ChatGPT Pro 5.5 (OpenAI), two of the most capable models commonly available to general researchers in mid-2026.

C. Proposal preparation

Every proposal, human or AI, followed the same one-page template with four headed sections: (1) Project Title, copied from the expert's title; (2) Background, a one-sentence summary; (3) Goal, a one-sentence summary (the expert's full background and goal were provided verbatim as input and here condensed to one sentence each, whereas the title was reproduced verbatim); and (4) Methodology, broken into no more than five major steps or phases (e.g., data preparation, modeling, analysis), each with an approximate completion time, in under 300 words. Proposals were to use a formal scientific tone appropriate for an academic review panel. This template is intentionally far shorter than a typical research or grant proposal, keeping the writing and review task tractable while still exercising the core planning skills of framing, methodology, and timeline. Whether the same patterns hold for full-length proposals is left to future work. The full instructions and prompt template are in Appendix B.

The human proposals were written by the experts. Starting from the previously prepared background and goal paragraphs, each expert wrote a one-page plan following the template. To remove superficial stylistic and grammatical cues that might reveal human authorship, each human-written plan was then passed through ChatGPT with the instruction to correct typos and grammatical mistakes while making minimal changes to the content. This normalization preserved the scientific content while making the surface style more uniform across human and AI submissions.

The AI proposals were produced by the AI prompters. For each project, the prompter generated three proposals using three different assistants, ChatGPT, Claude, and DeepSeek, each prompted with the same template (Appendix B) and with the expert's title, background, and goal. Because frontier models were updated frequently over the study period, and because benchmarking specific model versions is not the aim of this work, we did not require a fixed model version; prompters used whichever current version of each assistant was available to them, generally through the free or entry-level paid tiers. The AI proposals were generated between June and September 2025. During this window the versions in typical use were GPT-4o and the o-series reasoning models for ChatGPT (with GPT-5 becoming available following its August 2025 release), Claude Sonnet 4 / Opus 4 for Claude, and DeepSeek-V3 / R1 for DeepSeek. This yielded $8 \times 3 = 24$ AI proposals, which together with the 8 human proposals make up the corpus of 32.

D. Proposal evaluation

All 32 proposals were anonymized and standardized to a uniform visual format so that neither origin nor identity could be inferred from formatting. Each proposal was then independently assessed by four human reviewers and, separately, by two AI reviewers. For every proposal, each reviewer performed two tasks: (1) a binary judgment of whether the proposal was written by a human or by an AI; and (2) a quality rating on a predefined rubric.

The rubric scores each proposal from 1 (poor) to 5 (excellent) along four aspects: clarity and structure of the research plan; appropriateness of methods to the scientific goal; resource and tool planning; and feasibility, timeline, and risk awareness. The full scoring rubric, identical to the one shown to both the human planners and the reviewers, is given in Table II. The same rubric

Aspect	5 (Excellent)	4 (Good)	3 (Adequate)	2 (Weak)	1 (Poor)
Clarity and Structure of Research Plan	Clear, logical sequence of steps; structured into coherent phases.	Mostly clear structure; small gaps in logic or detail.	Some structure present, but steps are vague or loosely connected.	Poorly structured or fragmented plan; lacks internal coherence.	No discernible structure; chaotic or missing entirely.
Appropriateness of Methods to Scientific Goal	Methods are well-justified, state-of-the-art, and fit the scientific question.	Methods are suitable, though justification may be thin or options unexplored.	Methods are acceptable but basic or overly generic.	Methods are mismatched or poorly motivated.	Methods are incorrect, unworkable, or irrelevant.
Resource and Tool Planning	Specifies required datasets, software, AI tools, and compute needs with precision.	Most tools and resources identified; some details vague.	Only basic resources mentioned; limited specificity.	Important resources missing or misunderstood.	No mention of data/tools; unrealistic resource assumptions.
Feasibility, Timeline, and Risk Awareness	Timeline is realistic, well-paced; risks and contingencies are clearly addressed.	Feasible plan with some timing uncertainty; minor risk handling.	Rough timeline present but lacks detail or risk mitigation.	Unclear or unrealistic timeline; no mention of risks.	Infeasible timeline or no plan at all.

TABLE II. Rubric used for the blind evaluation of the research proposals. Each proposal was scored from 1 to 5 on each of the four aspects by every reviewer (human and AI).

and origin-judgment task were given to the AI reviewers, allowing a direct comparison between human and AI evaluation.

For the AI evaluation we used two assistants, Claude Opus 4.8 (Anthropic) and ChatGPT Pro 5.5 (OpenAI), each asked to score all 32 proposals on the rubric and to guess whether each proposal was human- or AI-written, using the same instructions provided to the human reviewers. Both AI reviewers were run at their highest available reasoning setting. The agreement between human and AI reviewers, both on origin classification and on rubric scores, is presented in Sec. III.

III. RESULTS

We report two complementary analyses of the 32 proposals. First, we ask how reliably evaluators can identify whether a proposal was written by a human or an AI, and what cues drive those judgments (Sec. III A). Second, we compare the quality scores assigned to human- and AI-written proposals by both human and AI reviewers (Sec. III B). Throughout, "human reviewers" refers to the four expert reviewers, and the two AI reviewers are Claude Opus 4.8 and ChatGPT Pro 5.5.

A. Who wrote the proposal? Human or AI?

By default, each AI reviewer was prompted, in a single pass, to score all 32 proposals on the rubric and to judge their origin. To guard against ordering or priming (earlier items biasing later judgments) effects, we also ran a two-step variant in which the proposals were scored first and their origin judged only afterwards, and we compared randomized versus sorted orderings of the proposal links; neither change significantly altered the ratings or the origin classifications (the two-step, rate-first run is labelled "Codex 5.5 Pro" in Table III).

Evaluator	AI-written	Human-written
Human reviewers	79%	72%
Claude (Opus 4.8)	100%	100%
ChatGPT (5.5 Pro)	100%	100%
Codex (5.5 Pro, rate-first)	100%	100%
Claude (Sonnet 4.6)	67%	100%

TABLE III. Percentage of proposals whose author was correctly identified, for AI-written versus human-written proposals, by each evaluator.

Table III reports the percentage of correct origin judgments for AI-written and human-written proposals, by evaluator. Four human reviewers identified AI-written proposals correctly 79% of the time on average and human-written proposals 72% of the time, well above chance but far from perfect, and slightly more likely to mistake a human proposal for an AI one than the reverse. Accuracy varied widely across the individual reviewers, however, from 59% to 88%, so the human panel is far from a single clean classifier. The variation across the three AI authors was modest: when written by ChatGPT, Claude, and DeepSeek, the AI-written proposals were correctly flagged as AI by the human reviewers 84%, 72%, and 81% of the time, respectively, with no single model consistently easier or harder to detect. The most capable AI reviewers, by contrast, classified every pro-

posal correctly: Claude Opus 4.8, ChatGPT Pro 5.5, and the rate-first Codex 5.5 Pro all reached 100% accuracy on all 32 proposals, for both human- and AI-written cases. We note that the proposals were presented together in a single pass, but no evaluator—human or AI—was told the ratio of human- to AI-written proposals; the perfect AI accuracy therefore does not stem from knowing the class balance. With limited proposals, however, this perfect score should not be read as evidence of reliable authorship detection in general. The near-perfect performance also appears to be a property of the frontier models only: a smaller model, Claude Sonnet 4.6, still identified every human-written proposal correctly but flagged AI-written proposals only 67% of the time, below the human reviewers.

The free-text justifications reveal what drove these judgments. The human reviewers most often flagged a proposal as AI-written when it looked "too clean," had a "very tight" or template-like five-step structure, contained no citations, proposed unrealistic or round-number timelines, listed irrelevant or nonsensical methods and tools, or leaned on machine-learning buzzwords (e.g., "bootstrap resampling," or a tacked-on large-language-model step). Conversely, they read a proposal as human when it carried concrete author–year citations or self-citations, used idiosyncratic jargon ("donuts," "blobs"), adopted a focused and feasible scope, used the first person ("we"), or contained insider domain markers (e.g., a *LiteBIRD* or HSC reference suggesting an author from a specific institute). These cues were applied inconsistently across the four reviewers: for the same proposal, one reviewer often read a feature as a human signature while another read it as a hint for AI, and several judgments were openly heuristic: one reviewer guessed "AI" simply because they "already selected two humans."

The two AI reviewers claimed to base their judgments on a strikingly similar set of features, but applied them consistently and without error. They marked a proposal as AI-written when it had a uniform five-phase template, polished and impersonal prose with repetitive parallel sentence structure, generic survey (large observational program) choices named without specificity, exhaustive tool lists, round month-based timelines, a boilerplate risk-mitigation step, and generic deliverables such as a "public white paper" or "community code release." They marked a proposal as human when it cited specific papers by author and year, gave concrete parameter ranges or observational priors (e.g., a surface density of 0.2 LBGs arcmin^{-2}, or specific laser pulse-energy ranges), used idiosyncratic jargon, or showed cautious, conditional scientific workflow (validating on simulations before real data; beginning from a Helmholtz decomposition). Notably, Claude Opus 4.8 and ChatGPT Pro 5.5 agreed with each other to a remarkable degree, frequently citing the same feature of the same proposal, and their reasoning overlapped with the human justifications, but unlike the humans, they never misclassified.

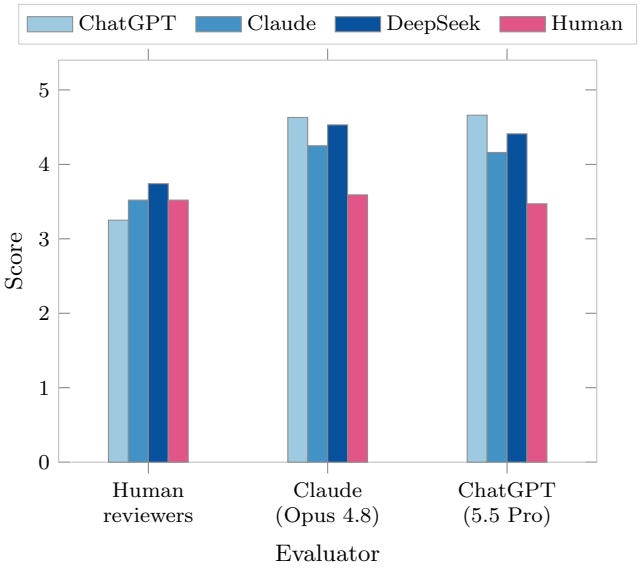

FIG. 2. Mean total proposal score (averaged over the four rubric aspects), grouped by evaluator (human reviewers, Claude Opus 4.8, ChatGPT Pro 5.5). Within each group, the bars show the four proposal authors, with the three AI authors in shades of blue and the human author in pink. Human reviewers score all authors similarly (~ 3.5), whereas both AI reviewers score the AI-written proposals roughly one point above the human-written ones. Means and standard deviations are listed in Table IV.

B. Proposal ratings

Figure 2 shows the mean total score (averaged over the four rubric aspects), grouped by evaluator, with the four proposal authors shown side by side within each group. In this and all subsequent plots, the reported standard deviations (Tables IV and V) are taken across the eight proposals in each evaluator–author category (one per project).

The human reviewers rate all authors into a narrow range (3.25–3.74 out of 5): DeepSeek scored highest (3.74 ± 0.52), ChatGPT lowest (3.25 ± 0.47), and, critically, the human-written proposals (3.52 ± 0.59) in the middle, essentially the same as the AI-written average (3.51). By the human reviewers' own scoring, AI-written proposals were therefore no worse, and in DeepSeek's case marginally better, than the expert-written ones. We note, however, that the four human reviewers often disagreed with one another, so the human reviewers' mean is a relatively noisy measure.

The AI reviewers, in contrast, rated the AI-written proposals far higher: Claude Opus 4.8 gave the AI authors an average of 4.47 versus 3.59 for the human authors, and ChatGPT Pro 5.5 gave 4.41 versus 3.47—roughly one full point in favor of the AI-written proposals. The same gap appears in two further tests not shown in Fig. 2: the rate-first Codex 5.5 Pro (AI-written 4.51 vs. human-written 3.50) and Claude Sonnet 4.6 (4.24

Author	Human rev.	Claude 4.8	ChatGPT 5.5
ChatGPT	3.25 ± 0.47	4.63 ± 0.52	4.66 ± 0.39
Claude	3.52 ± 0.43	4.25 ± 0.54	4.16 ± 0.42
DeepSeek	3.74 ± 0.52	4.53 ± 0.25	4.41 ± 0.39
Human	3.52 ± 0.59	3.59 ± 0.61	3.47 ± 0.50

TABLE IV. Mean total proposal score ± standard deviation across the eight projects, for each proposal author (rows) and evaluator (columns); the data plotted in Fig. 2.

vs. 3.72), confirming that neither scoring the proposals before judging origin nor using a smaller model removes the effect. The human-written proposals received nearly the same score from every evaluator (3.47–3.72), so the disagreement is driven almost entirely by how the AI-written proposals are scored: humans judged them around 3.5, while the AI reviewers judged them around 4.2–4.5. The **pro-AI bias** appears across all four AI models and is absent from human evaluators.

The per-aspect breakdown in Fig. 3 shows that these patterns are not uniform across rubric dimensions. The pro-AI bias of the AI reviewers is most pronounced for Clarity and Structure, where they rate the AI authors near the ceiling ($\sim$ 4.9–5.0) while holding the human author at ~ 3.75, and for Resource and Tool Planning, where the human author is scored lowest of all four authors by both AI reviewers (3.13–3.25). It is weakest for Appropriateness of Methods to Scientific Goal: here the human reviewers actually place the human author at or near the top (3.72), and the AI-reviewer gap, while still present, is the smallest of the four aspects. Feasibility, Timeline, and Risk Awareness is consistently the weakest dimension for the AI-written proposals under every evaluator (e.g., 3.24 from human reviewers, versus 3.46–3.77 on the other three aspects), echoing the qualitative observation that AI plans tend toward unrealistic or round-number timelines and over-ambitious scope; it is also the aspect on which the human and AI proposals are most comparable.

We find no evidence of a same-vendor preference among the AI reviewers: ChatGPT-written proposals were scored highest by both the Claude and the ChatGPT evaluators, so the pro-AI bias is a preference for AI-style writing in general rather than for an evaluator's own model family.

C. Project-level variation

The near-parity between human- and AI-written proposals in the human panel is an aggregate that hides substantial project-to-project variation (Fig. 4). The human reviewers rated the human-written proposal higher in five of the eight projects and the AI-written proposals higher in the remaining three. The size and sign of this gap are strongly anti-correlated with the quality of the human proposal itself (Pearson $r = -0.95$): where the expert produced a detailed, domain-specific plan (e.g., RG, SU2, IA), reviewers ranked it above the AI versions, whereas the AI advantage was concentrated in projects with an unusually thin human plan—most starkly GW, whose human proposal scored only 2.4/5. The two AI reviewers exhibit the same slope but shifted upward, preferring the AI-written proposals in all eight projects; the two panels differ mainly by a near-constant vertical offset—the roughly one-point pro-AI bias identified above, made visual. The human–AI quality comparison and the AI-reviewer bias are therefore driven by the same handful of projects rather than being uniform across the sample, and the comparison is sensitive to the level of individual human effort on each proposal.

IV. CONCLUSION

We studied human and AI performance on scientific project planning, an open-ended research-planning task with no ground-truth answer, across eight expert-conceived projects in physics, astrophysics, and cosmology. For each project, we collected one human-written and three AI-written one-page proposals and had all 32 anonymized proposals blindly evaluated for both authorship and quality by four human reviewers and two advanced AI models. We summarize three major findings:

1. On capability, at the level of a short, structured research proposal, current LLMs produce research plans comparable to those of human researchers, as rated by expert human reviewers (Figures 2, 3). This demonstrates AI's potential to assist with project planning in scientific workflows.

2. Humans and AI are not equally good at telling if a proposal is written by human or AI (Table III). Human reviewers were well above chance ($\gtrsim$70%) but relied on noisy and sometimes conflicting cues. The two most capable AI reviewers applied essentially the same cues as human reviewers but did so consistently and correctly, classifying all 32 proposals without error (100%).

3. Pro-AI bias was observed in all AI evaluations, where all four AI reviewers scored AI-written proposals roughly a point higher than human-written ones (Figures 2). Such bias is absent in human evaluation. This pro-AI bias is a concrete risk for any review system that incorporates LLMs: an AI reviewer may favor AI-written proposals, potentially disadvantaging human applicants and creating a feedback loop that rewards a recognizable "AI style" over scientific substance. The finding is consistent with previous works by [4, 8, 9, 11]. [1]

[1] Major funders have already moved to mitigate such biases: NIH

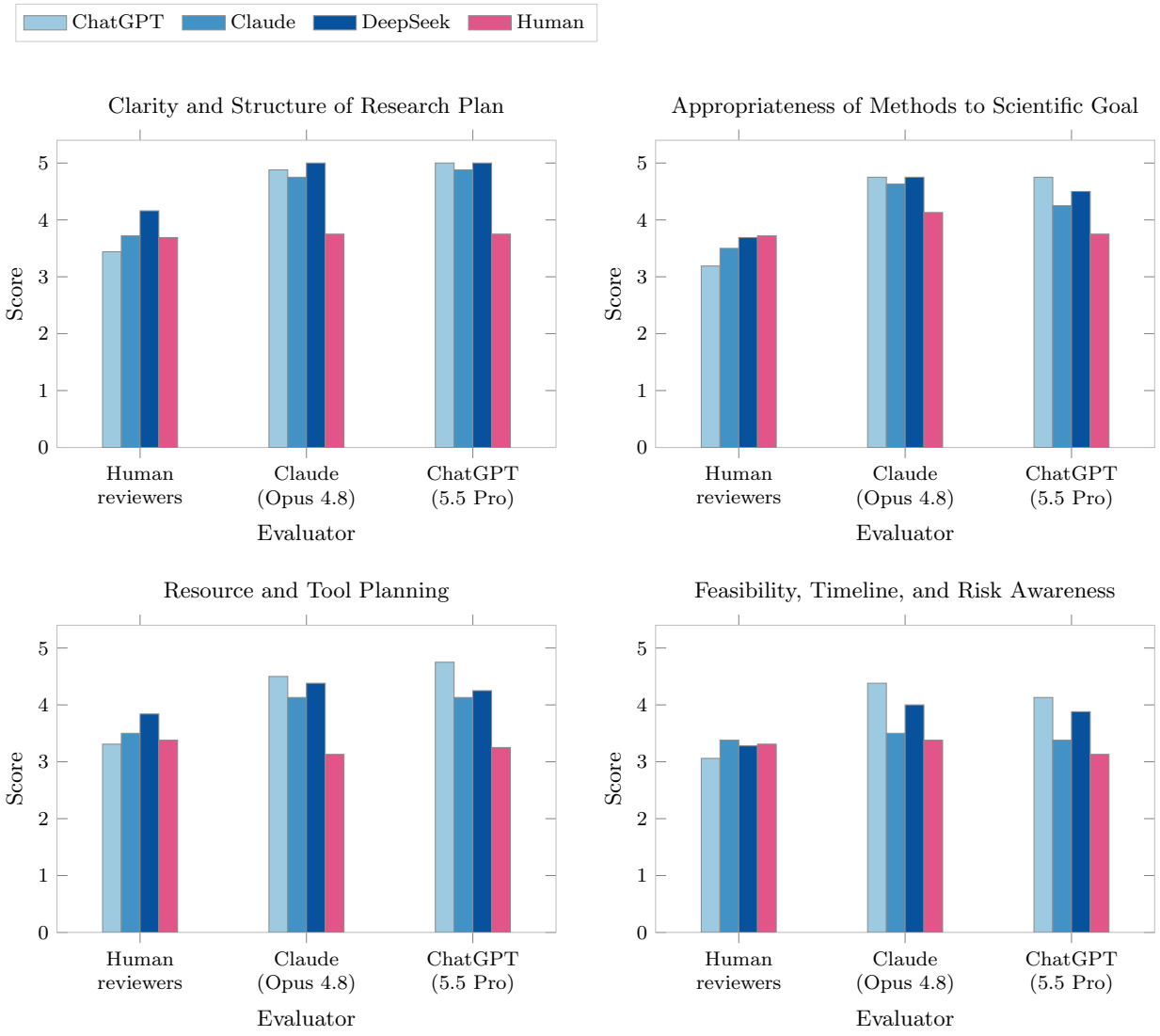

FIG. 3. Mean proposal score broken down by the four rubric aspects, grouped by evaluator (human reviewers, Claude Opus 4.8, ChatGPT Pro 5.5). The three AI authors are shown in shades of blue and the human author in pink. Means and standard deviations are listed in Table V.

4. These aggregate patterns hide strong project-to-project variation (Sec. III C, Fig. 4): the human panel preferred the human-written proposal in five of the eight projects, and the AI–human score gap is tightly anti-correlated with the quality of the human plan ($r = -0.95$). Both the human–AI parity and the pro-AI bias are thus concentrated in a few projects rather than uniform, and the comparison is sensitive to the level of individual human effort on each proposal.

Several caveats remain in our study. First, our sample is limited to eight projects. This reflects the substantial human effort and coordination required to advance many projects on a comparable timeline, rather than a lack of interest. Even so, the set deliberately spans a diverse range of topics and of methodologies—observational, theoretical, instrumental, and numerical. Second, the AI assistants used to generate the proposals changed rapidly over the roughly one-year span of the project and were already outdated by the time of submission. Our results should therefore not be read as a benchmark of the most capable current models, but as a record of a specific point in time at which LLMs became broadly competitive with human researchers on tasks of this kind. This trend is worth documenting precisely because assembling this degree of human coordination is slow and difficult. Third, we used only a few of the most popular models (Chat-

prohibits the use of generative-AI tools to analyze or formulate peer-review critiques of grant applications [20], NSF bars reviewers from uploading proposal content to non-approved AI tools [21], and the ERC requires non-delegation of evaluative judgment and strict confidentiality [22]; our findings supply an empirical rationale for such caution

Aspect	Author	Human reviewers	Claude (Opus 4.8)	ChatGPT (5.5 Pro)
Clarity and Structure	ChatGPT	3.44 ± 0.66	4.88 ± 0.35	5.00 ± 0.00
	Claude	3.72 ± 0.41	4.75 ± 0.46	4.88 ± 0.35
	DeepSeek	4.16 ± 0.40	5.00 ± 0.00	5.00 ± 0.00
	Human	3.69 ± 0.56	3.75 ± 0.46	3.75 ± 0.46
Appropriateness of Methods	ChatGPT	3.19 ± 0.46	4.75 ± 0.46	4.75 ± 0.46
	Claude	3.50 ± 0.64	4.63 ± 0.52	4.25 ± 0.46
	DeepSeek	3.69 ± 0.78	4.75 ± 0.46	4.50 ± 0.76
	Human	3.72 ± 0.49	4.13 ± 0.83	3.75 ± 0.71
Resource and Tool Planning	ChatGPT	3.31 ± 0.48	4.50 ± 0.53	4.75 ± 0.46
	Claude	3.50 ± 0.30	4.13 ± 0.64	4.13 ± 0.35
	DeepSeek	3.84 ± 0.44	4.38 ± 0.52	4.25 ± 0.46
	Human	3.38 ± 0.63	3.13 ± 0.64	3.25 ± 0.46
Feasibility, Timeline, Risk	ChatGPT	3.06 ± 0.29	4.38 ± 0.74	4.13 ± 0.64
	Claude	3.38 ± 0.35	3.50 ± 0.53	3.38 ± 0.52
	DeepSeek	3.28 ± 0.47	4.00 ± 0.00	3.88 ± 0.35
	Human	3.31 ± 0.68	3.38 ± 0.52	3.13 ± 0.35

TABLE V. Mean proposal score ± standard deviation (across the eight projects) for each rubric aspect, proposal author, and evaluator; the data plotted in Fig. 3.

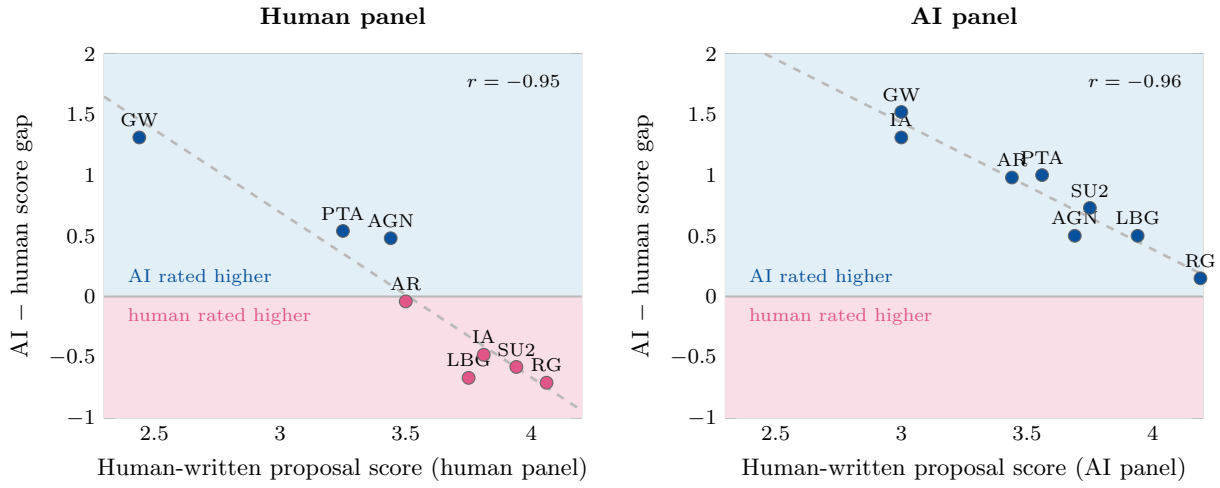

FIG. 4. Project-level breakdown of the AI−human score gap, for the human panel (left) and the AI reviewers (right). Each point is one of the eight projects; the horizontal axis is the score that panel gave the human-written proposal, and the vertical axis is the mean AI-written score minus the human-written score. Both panels show the same strong anti-correlation with the quality of the human proposal ($r \approx -0.95$): the AI advantage is largest where the human plan was weak (e.g., GW) and smallest for the best human proposals (RG). The panels differ mainly by a vertical offset—the human panel rates the human proposal higher in five of eight projects, whereas the AI reviewers rate the AI proposals higher in all eight—so the pro-AI bias appears as an upward shift of the same trend.

GPT, Claude, and DeepSeek) and did not attempt an exhaustive or version-matched model comparison, which was not the aim of this work. Finally, we evaluated only the methodology and planning of each proposal, not the novelty of the underlying scientific idea: in our design, the ideas were all human-conceived and held fixed across the human and AI proposals. Novelty is itself a central criterion in real review, and recent work finds that LLMs can generate research ideas that expert reviewers rate as more novel than those of human researchers, albeit with somewhat weaker feasibility [10]; our findings are complementary to previous studies.

Future work should broaden the disciplinary and demographic scope, test full-length proposals and multi-round review, incorporate an explicit assessment of idea novelty alongside planning quality, probe whether the pro-AI bias persists when evaluators are instructed to ignore style, and extend the analysis to the other stages of the research workflow.

ACKNOWLEDGMENTS

We thank Jingjing Shi, Qiuyue Liang, and Kenta Hotokezaka for help shaping some of the scientific projects. JL acknowledges support from the Kavli Foundation and Google. KV acknowledges support from the Kavli Foundation. AH was supported in part by JSPS KAKENHI Grant No. JP26K17133, the CD3 Google Seed grant, and by the World Premier International Research Center Initiative (WPI), MEXT, Japan. KK acknowledges support from the Forefront Physics and Mathematics Program to Drive Transformation (FoPM), a World-leading Innovative Graduate Study (WINGS) Program, the University of Tokyo. JG and ZL acknowledge support from the International Laboratory for Astrophysics, Neutrino and Cosmology Experiments (ILANCE). AEB is supported by the Simons Foundation. MRG acknowledges financial support from the Formación del Profesorado Universitario program of the Spanish Ministerio de Ciencia, Innovación y Universidades; from the CMB-Inflate project funded by the European Union's Horizon 2020 Research and Innovation Staff Exchange under the Marie Skłodowska-Curie grant agreement No. 101007633; and from MICIU/AEI/10.13039/501100011033 under projects PID2022-139223OB-C21 and PID2022-140670NA-I00 (also funded by FEDER, UE). LT is supported by JSPS under KAKENHI 24K22878 and 26K17136 and by the Royal Society under ICA\R2\252140.

AI usage: the AI-written proposals (Sec. II C) were generated around mid-2025 by the AI prompters using three assistants: ChatGPT (OpenAI; GPT-4o and the o-series reasoning models, with GPT-5 following its August 2025 release), Claude (Anthropic; Claude Sonnet 4 and Opus 4), and DeepSeek (DeepSeek-V3 and R1), generally accessed through free or entry-level paid tiers. The AI evaluations (Sec. II D) were performed with Claude Opus 4.8 (Anthropic) and ChatGPT Pro 5.5 (OpenAI); the additional two-step (rate-first) and cross-model tests used Codex 5.5 Pro (OpenAI) and Claude Sonnet 4.6 (Anthropic). Beyond their role as study instruments, LLMs were also used to assist with language editing, figure preparation, and formatting of this manuscript; all scientific content, analyses, and conclusions are the authors' own.

AUTHOR AFFILIATIONS

[1]Center for Data-Driven Discovery, Kavli IPMU (WPI), UTIAS, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan
[2]Kavli IPMU (WPI), UTIAS, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8583, Japan
[3]Department of Astrophysical Sciences, Princeton University, Peyton Hall, Princeton, NJ 08544, USA
[4]Department of Physics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[5]Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA
[6]Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, United Kingdom
[7]Research Center for the Early Universe, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan
[8]École polytechnique, Institut polytechnique de Paris, Palaiseau, France
[9]Advanced Energy, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan
[10]Department of Astronomy, University of Texas at Austin, Austin, Texas, USA
[11]Department of Physics, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan
[12]Instituto de Física de Cantabria (IFCA, CSIC–UC), Avenida los Castros s/n, 39005 Santander, Spain
[13]Departamento de Física Moderna, Universidad de Cantabria, Avenida los Castros s/n, E-39005 Santander, Spain
[14]Institute for Cosmic Ray Research, The University of Tokyo, 5-1-5 Kashiwa-no-Ha, Kashiwa, Chiba 277-8582, Japan
[15]Department of Astronomy, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China
[16]School of Astronomy and Space Sciences, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China

[1] Y. D. Hezaveh, L. Perreault Levasseur, and P. J. Marshall, Nature **548**, 555 (2017).

[2] F. Villaescusa-Navarro, D. Anglés-Alcázar, S. Genel, D. N. Spergel, R. S. Somerville, R. Dave, A. Pillepich, L. Hernquist, D. Nelson, P. Torrey, *et al.*, The Astrophysical Journal **915**, 71 (2021).

[3] T. D. Nguyen, Y.-S. Ting, I. Ciucă, C. O'Neill, Z.-C. Sun, M. Jabłońska, S. Kruk, E. Perkowski, J. Miller, J. Li, *et al.*, in *Proceedings of the Second Workshop on Information Extraction from Scientific Publications (WIESP), IJCNLP-AACL 2023* (2023) pp. 49–55, arXiv:2309.06126.

[4] H. Zhou, H. Huang, Y. Long, B. Xu, C. Zhu, H. Cao, M. Yang, and T. Zhao, in *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (CCL)* (2024) pp. 1310–1319, arXiv:2409.16788.

[5] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, PMLR, Vol. 202 (2023) pp. 24950–24962, arXiv:2301.11305.

[6] C. A. Gao, F. M. Howard, N. S. Markov, E. C. Dyer, S. Ramesh, Y. Luo, and A. T. Pearson, npj Digital Medicine **6**, 75 (2023).

[7] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha, arXiv e-prints (2024), arXiv:2408.06292.

[8] A. Shcherbiak, H. Habibnia, R. Böhm, and S. Fiedler, Judgment and Decision Making **19**, e21 (2024).

[9] A. Panickssery, S. R. Bowman, and S. Feng, in *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 37 (2024) arXiv:2404.13076.

[10] C. Si, D. Yang, and T. Hashimoto, in *International Conference on Learning Representations (ICLR)* (2025)

arXiv:2409.04109.
[11] K. Wataoka, T. Takahashi, and R. Ri, arXiv e-prints (2024), arXiv:2410.21819. Presented at the NeurIPS 2024 Safe Generative AI Workshop.
[12] W. Liang, Y. Zhang, H. Cao, B. Wang, D. Y. Ding, X. Yang, K. Vodrahalli, S. He, D. S. Smith, Y. Yin, D. A. McFarland, and J. Zou, NEJM AI **1**, 10.1056/AIoa2400196 (2024), arXiv:2310.01783.
[13] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, arXiv e-prints (2023), arXiv:2303.11156.
[14] J. Ren and W. Wang, arXiv e-prints (2025), arXiv:2509.09709.
[15] V. Sikimić, Synthese **206**, 282 (2025).
[16] U. Sandström and M. Thelwall, arXiv e-prints (2026), arXiv:2603.14565.
[17] W. Thorne, J. James, and Y. Wang, arXiv e-prints (2026), arXiv:2603.08281.
[18] F. Villaescusa-Navarro, B. Bolliet, P. Villanueva-Domingo, A. E. Bayer, A. Acquah, C. Amancharla, A. Barzilay-Siegal, P. Bermejo, C. Bilodeau, P. C. Ramírez, M. Cranmer, U. L. França, C. Hahn, Y.-F. Jiang, R. Jimenez, J.-Y. Lee, A. Lerario, O. Mamun, T. Meier, A. A. Ojha, P. Protopapas, S. Roy, D. N. Spergel, P. Tarancón-Álvarez, U. Tiwari, M. Viel, D. Wadekar, C. Wang, B. Y. Wang, L. Xu, Y. Yovel, S. Yue, W.-H. Zhou, Q. Zhu, J. Zou, and Íñigo Zubeldia, The denario project: Deep knowledge ai agents for scientific discovery (2025), arXiv:2510.26887 [cs.AI].
[19] A. Hell and L. Thiele, (2026), arXiv:2605.08212 [cs.LG].
[20] National Institutes of Health, The use of generative artificial intelligence technologies is prohibited for the NIH peer review process, NIH Guide Notice NOT-OD-23-149 (2023), `https://grants.nih.gov/grants/guide/notice-files/NOT-OD-23-149.html`.
[21] National Science Foundation, Notice to the research community: Use of generative artificial intelligence technology in the NSF merit review process (2023), `https://www.nsf.gov/news/notice-to-the-research-community-on-ai`.
[22] European Research Council, The use of AI in grant proposal evaluation (2026), `https://erc.europa.eu/system/files/2026-03/Use-AI-grant-proposal-evaluation.pdf`. See also `https://erc.europa.eu/news-events/news/erc-clarifies-limits-ai-use-grant-evaluation`.

Appendix A: Background and goals of the eight research projects

The following project titles, backgrounds, and goals were written by the human experts without any AI assistance, and were the common input provided to the human planners and the AI prompters for each project (Sec. II A).

1. AGN — MaNGA: AGN Duty Cycle

Background: The time a galaxy spends in the AGN phase, from both general arguments and ensemble studies such as quasar clustering and black-hole mass-function studies and He II proximity-zone analysis, is suggested to last $\sim 10^6$–10^9 yr. Ionization studies of AGN host galaxies and their surroundings indicate that active nuclei can switch on/off on a $\sim$100 kyr timescale—multiple times during their lifetime—also finding support in numerical simulations, indicating that the AGN luminosity may fluctuate by orders of magnitude over its lifetime. For a galaxy size of $R \sim 10$ kpc it takes $t = R/c \sim 30$ kyr for a change in AGN UV luminosity to propagate over the entire galactic disk. We propose to perform a systematic search for fading or, more generally, variable AGN in the most recent MaNGA survey DR17.
Goal: To assess the total fraction of fading/brightening AGN among the sample, indicative of the duty-cycle fraction these objects span in the transitional phase, directly comparable to the predictions of cosmological simulations. We further aim to characterize the AGN UV luminosity evolution on kyr timescales and compare it, where possible, to that derived from the resolved longitudinal brightness profiles of X-ray jets, potentially yielding a first-of-its-kind link between AGN accretion rate and high-energy particle acceleration in AGN jets.

2. LBG — The galaxy–dark matter halo connection of Lyman-break galaxies

Background: A Lyman-break galaxy (LBG) is a galaxy whose broadband photometry shows a "drop-out" in the bluest bands as features in its spectrum move from blue to red through the filter set due to cosmic expansion; the reduction in flux blueward of the Lyman-α and Lyman-limit frequencies is caused by scattering and absorption of UV photons by intervening neutral hydrogen. This feature is a robust way to identify galaxies at $z > 2$ in photometric surveys, which can then be followed up spectroscopically. Current and future surveys, such as the ʻŌnohiʻula Prime Focus Spectrograph Galaxy Evolution Survey (PFS:GE), will measure redshifts for thousands of LBGs at $2.0 < z < 4.0$ to study their properties and how they trace large-scale structure (LSS).
Goal: In preparation for the measurement and modeling of the LBG galaxy–galaxy two-point correlation function (2PCF), we will conduct a theoretical investigation into the potential progenitors of the PFS:GE LBG target sample using the Uchuu–UniverseMachine mock galaxy catalog, to better understand how they trace the LSS via galaxy bias and the halo occupation distribution (HOD) as a function of redshift. Of particular interest is any sign of galaxy assembly bias or other selection-function nuances that could lead to incorrect inferences of the galaxy bias (b_g) and growth rate of LSS ($f\sigma_8$).

3. IA — Intrinsic alignments in varying environments

Background: Weak-lensing surveys are one of the most powerful probes in cosmology; however, the intrinsic alignment (IA) of galaxies contaminates the signal. Many studies have therefore investigated the characteristics of IA in order to eliminate it from the data. So far, researchers believe IA is related to galaxy color and luminosity; in addition, we suspect it is related to the large-scale environment, such as the matter distribution.
Goal: We will evaluate the additional dependence of IA on the large-scale matter density field and estimate the impact on cosmological analysis when such a dependence is neglected.

4. AR — Prediction of debris emergence on laser-ablated sub-wavelength shapes

Background: We have been developing methods to fabricate sub-wavelength structures (SWS) for anti-reflective coating in the millimeter-wave region on hard materials such as ceramics, using ultra-short-pulse laser ablation, which is crucial for machining materials with relatively wide energy band gaps. For efficient fabrication of SWS with a higher ablation rate it is necessary to inject lasers with as high a power as possible; however, at the same time we observe redeposition of ablated debris, particularly at higher energies. Since the physics behind the emergence of debris is mostly unknown, their shape is uncontrollable, and the quantitative conditions under which they appear are ambiguous.
Goal: To predict the shape expected to be fabricated for a given set of laser-scanning parameters (pulse frequency, pulse energy, spacing of scanning lines, etc.), we first accumulate data on fabricated shapes across different parameter sets and make plots of the results including shape information such as depth and the existence of debris. Our goal is to extrapolate to arbitrary parameters using Gaussian-process regression (and, where useful, large language models), comparing predicted data points against real laser-machining data.

5. RG — Radio Galaxies with HalfDome

Background: At low CMB frequencies (around 100 GHz), high-energy radio galaxies act as bright point-source contaminants to CMB maps. The locations of these galaxies are likely correlated with features in the underlying large-scale structure as well as with galaxy properties (e.g., the CIB, radio continuum, X-ray).
Goal: Add realistic radio luminosities to galaxies in N-body simulations that are relevant for CMB foreground-contamination modeling and correlated with the mass of the underlying haloes, incorporating as much accurate physics as possible (such as the synchrotron-spectrum behavior of different radio-galaxy types). The HalfDome simulations aim to create realistic correlated foregrounds across different wavelengths on N-body simulations.

6. GW — Environment of gravitational-wave black hole binaries with weak-lensing maps

Background: The first detection of a gravitational wave (GW) by LIGO opened a new era of multi-messenger astronomy, and around 300 GW events from binary black hole (BBH) mergers have now been observed. However, the origins of these binary black holes and the environments they reside in remain unknown. Combining gravitational waves with cosmological data will provide rich information on the origin and evolution of BBHs.
Goal: This project aims to unveil the environment of BBHs. We combine publicly available gravitational-wave localization data with cosmological data, such as weak-lensing maps and galaxy-overdensity maps, to investigate the correlation between the BBH distribution and the matter distribution in the universe. The results will be used to constrain the astrophysical origins of BBHs.

7. PTA — Forecasting pulsar timing array sensitivity to deviations from general relativity

Background: The Pulsar Timing Array (PTA) is a measurement method relying on the observation of pulsars, fast-rotating neutron stars with well-known timing models. By measuring slight perturbations in the times of arrival (ToAs) of each pulse, computing residuals, and plotting the pulsar-pair correlations as a function of angular separation, one obtains the overlap reduction function (ORF), known in GR as the Hellings–Downs curve. This method allowed, in 2023, evidence for a stochastic gravitational-wave background to be reported by several collaborations (EPTA, NANOGrav, CPTA, ...). In modified gravity, additional polarization modes and a modified dispersion relation lead to a modified ORF.
Goal: In our work we neglect the scalar and vector polarization modes and focus on a modified dispersion relation (which modifies the phase and group velocities). Approximating the gravitational-wave background as plane waves interfering with each other, the relevant quantity to study is the phase velocity, which we do not constrain a priori (parameter space $[0, +\infty)$). Our goal is to forecast the measurement precision required to distinguish, at $1, 2, 3$ to 5σ confidence, a 20%, 10% to 1% deviation from GR (in terms of phase-velocity deviation), and in how many years that precision will be achievable.

8. SU2 — Massive Yang–Mills theory

Background: When exploring mechanisms that drive inflation, non-Abelian gauge fields—such as SU(2) Yang–Mills fields—have been proposed as alternatives to scalar fields. For instance, introducing a Chern–Simons coupling between a pseudo-scalar field and a non-Abelian gauge field can lead to slow-roll inflation, as in Chromo-Natural Inflation. Moreover, higher-order gauge-field corrections, such as $(F\tilde{F})^2$ terms, can emerge in effective field theories and play a significant role in early-universe dynamics, potentially enabling inflationary scenarios.
Goal: The goal of this project is to investigate mechanisms by which vector fields can drive cosmic inflation. We focus on Proca and Yang–Mills SU(2) theories and study how to modify them so that their background solution leads to accelerated expansion on cosmological scales, with particular emphasis on the effect of adding a mass term in both theories compared to the massless case.

Appendix B: Standardized prompts and instructions

Both human planners and AI prompters were asked to produce a one-page proposal, in a formal scientific tone appropriate for an academic review panel, organized under four headings:

1. Project Title, copied from the title provided;
2. Background, a one-sentence summary;
3. Goal, a one-sentence summary; and
4. Methodology, broken into no more than five major steps or phases (e.g., data preparation, modeling, analysis), each with an approximate completion time, in under 300 words.

All parties were told that the proposal would be evaluated on four criteria: clarity and structure of the research plan; appropriateness of methods to the scientific goal; resource and tool planning; and feasibility, timeline, and risk awareness (Table II). The AI prompters prompted each model with the following fixed template:

> You are an expert physics researcher writing a 1-page proposal for a new research project. The audience is an academic review panel.
>
> The project title: [provided by human]
> Background: [provided by human]
> Goal: [provided by human]
>
> Write the proposal using clear, concise academic language and organize it under the following headings and instructions: 1. Project Title: copy from the title provided; 2. Background: one sentence summary; 3. Goal: one sentence summary; 4. Methodology: Break down the work into major steps or phases (e.g., data preparation, modeling, analysis). Include theoretical, computational, or experimental techniques if relevant. Include the approximate time for each step. No more than 5 steps and keep it under 300 words.
>
> Use formal scientific tone appropriate for a grant or academic setting. Your proposal will be evaluated based on the following criteria: 1. Clarity and structure of research plan, 2. Appropriateness of methods to scientific goal, 3. Resource and tool planning, 4. Feasibility, timeline, and risk awareness.

To remove superficial stylistic tells, each human-written proposal was passed through ChatGPT with the following instruction: "Correct typo and grammar mistakes, with minimal change to the content: [text]."

Appendix C: Sample proposals for the SU(2) project

To illustrate the material that reviewers saw, we reproduce the four proposals for the SU(2) project (one human, three AI), all generated from the identical title, background, and goal in Appendix A. They are lightly reformatted for typesetting but otherwise unedited.

Human

Title: Massive Yang–Mills theory (SU2).

Background: In the context of early-universe cosmology, several models have proposed that vector fields—either through direct coupling to the inflaton scalar field or via higher-order terms such as F^4—can act as viable sources of inflation.

Goal: This internship tackles vector-field theories as candidates driving cosmic inflation, starting with Proca theory and extending it to non-Abelian massive SU(2) Yang–Mills theory, with a particular focus on their coupling to gravity and the effect of the addition of the mass.

Methodology:

1. *Theoretical foundations & Helmholtz decomposition (2 weeks):* Review Proca theory in curved spacetime; use the Helmholtz decomposition to separate the vector field into longitudinal and transverse modes; derive the equations of motion and identify constraint equations; compare with the massless (Maxwell) case to isolate physical degrees of freedom.

2. *Energy–momentum tensor & inflation criteria in an FRW background (2–3 weeks):* Apply the theory to a homogeneous, isotropic FRW background; derive the energy–momentum tensor from the Proca Lagrangian and compute the energy density and pressure; analyze whether the configuration can drive inflation (i.e., satisfies $\rho + 3p < 0$).

3. *Extension to massive SU(2) Yang–Mills theory (2–3 weeks):* Generalize to the non-Abelian SU(2) case with mass; derive the equations of motion and compute energy density and pressure in FRW spacetime; compare with Proca theory and evaluate under what conditions the SU(2) configuration leads to inflationary dynamics.

4. *Computational implementation & numerical solutions (3–4 weeks):* Implement the background field equations numerically in Python or Mathematica; study the time evolution of the scale factor, energy density, and field components for various initial conditions and coupling parameters; identify regions in parameter space favorable for inflation.

5. *Perturbation analysis & stability (3–4 weeks):* Linearize the field equations around the inflationary background and classify perturbations (scalar, vector, tensor); study their evolution to assess the stability of the background solution; analyze whether instabilities or anisotropies arise in either the Proca or SU(2) case.

ChatGPT

Title: Massive Yang–Mills theory (SU2).

Background: Non-Abelian SU(2) gauge fields augmented by Chern–Simons and higher-order $F\tilde{F}$ operators offer a viable alternative to scalar inflation models, but the cosmological role of an explicit mass term remains unexplored.

Goal: Quantitatively determine how adding a mass term to Proca and SU(2) Yang–Mills fields alters background dynamics and perturbative stability, identifying parameter regimes that support sustained, isotropic inflation.

Methodology:

1. *Model formulation (4 months):* Derive covariant Lagrangians for massive Proca and SU(2) sectors including F^2, $F\tilde{F}$, and $(F\tilde{F})^2$ corrections; impose cosmological ansätze; obtain Euler–Lagrange equations.

2. *Background dynamics & stability (4 months):* Solve background equations numerically across mass and coupling grids; compute slow-roll parameters, anisotropy measures, and identify fixed points; perform linear stability (tensor, vector, scalar) via the Mukhanov–Sasaki formalism, using Python/Julia ODE solvers and Mathematica for eigen-analysis.

3. *Perturbation spectra & observables (4 months):* Evolve perturbations through horizon crossing; derive power spectra, tensor chirality, and non-Gaussianity; compare with CMB–LiteBIRD sensitivities and PTA gravitational-wave forecasts.

4. *Parameter estimation & feasibility checks (3 months):* Constrain viable regions using Planck+BICEP/Keck data; propagate instrumental uncertainties to inflationary parameters; assess robustness against radiative corrections and strong-coupling bounds.

5. *Risk mitigation & dissemination (5 months):* If the mass term triggers gradient instabilities, pivot to Stückelberg-like UV completions; cross-check with lattice simulations; publish results in JCAP; present at COSMO-26.

Claude

Title: Massive Yang–Mills theory (SU2).

Background: This project investigates non-Abelian gauge fields as inflationary drivers, building on established frameworks like Chromo-Natural Inflation where SU(2) Yang–Mills fields coupled to pseudo-scalar fields can generate slow-roll inflation through Chern–Simons interactions and higher-order gauge corrections.

Goal: The primary objective is to systematically analyze how Proca and Yang–Mills SU(2) vector-field theories can be modified to produce cosmological inflation, with particular emphasis on comparing the dynamical effects of mass terms versus massless configurations in driving accelerated expansion.

Methodology:

1. *Theoretical framework development (3 months):* Derive modified Einstein field equations for SU(2) Proca and Yang–Mills theories in cosmological backgrounds; establish the action formalism incorporating mass terms and analyze the stress–energy tensor components; identify key parameters controlling the transition from massless to massive regimes.

2. *Background solution analysis (4 months):* Solve the coupled Einstein–Yang–Mills/Proca equations numerically (Mathematica, Python); construct a phase-space analysis to identify inflationary attractors and determine slow-roll conditions; map the parameter space where accelerated expansion is viable.

3. *Perturbation theory and stability (3 months):* Perform linear perturbation analysis around inflationary backgrounds to assess stability and compute scalar/tensor power spectra; implement a gauge-invariant formalism for vector perturbations; verify that ghost instabilities are absent in the massive case.

4. *Observational constraints (2 months):* Compare theoretical predictions with Planck CMB data, focusing on spectral indices, tensor-to-scalar ratio, and non-Gaussianity signatures; establish observationally viable parameter ranges and distinguish between Proca and Yang–Mills scenarios.

5. *Comparative analysis and documentation (3 months):* Synthesize results comparing massive versus massless theories, identify novel inflationary mechanisms, and prepare manuscripts for publication; develop numerical codes for community use and present findings at international conferences.

DeepSeek

Title: Massive Yang–Mills theory (SU2).

Background: Non-Abelian gauge fields, notably SU(2) Yang–Mills fields coupled via mechanisms like Chern–Simons terms or through higher-order corrections such as $(F\tilde{F})^2$, present viable alternatives to scalar fields for driving cosmic inflation by enabling sustained slow-roll dynamics in the early universe.

Goal: This project aims to systematically investigate how massive vector fields, specifically within modified Proca and SU(2) Yang–Mills theories, can generate viable cosmic inflation on cosmological scales, comparing the effects of explicit mass terms against their massless counterparts on the background accelerated expansion.

Methodology:

1. *Theoretical framework & model specification (2 months):* Formally define the action for the massive Proca and SU(2) Yang–Mills theories, including the explicit mass term and necessary potential or coupling structures to ensure consistent isotropic background configurations; derive the background equations of motion in a flat FRW spacetime.

2. *Background dynamics computation (3 months):* Implement numerical codes (Python/Mathematica, using ODE solvers like `scipy.integrate.solve_ivp`) to solve the coupled Friedmann and gauge-field background-evolution equations; scan parameter spaces (mass scale, coupling constants, initial field values) to identify regions yielding sufficient e-folds of quasi-de Sitter expansion; quantify inflationary observables for both massive and reference massless cases.

3. *Stability analysis (3 months):* Perform a comprehensive linear stability analysis of the identified inflationary solutions against isotropic perturbations, deriving and numerically solving the perturbation equations to assess the presence of ghosts, Laplacian, or gradient instabilities.

4. *Perturbation spectrum & observables (3 months):* Calculate the primordial scalar and tensor perturbation spectra; compute key observables (scalar spectral index n_s, tensor-to-scalar ratio r, non-Gaussianity $f_{\rm NL}$); compare predictions between massive Proca, massive SU(2), and their massless counterparts.

5. *Synthesis, comparison & publication (1 month):* Integrate results, evaluating the viability and distinguishing features of massive vector inflation; assess the specific role of the mass term in enabling stable inflation and shaping observational signatures; prepare findings for publication.