

SCICoQA: Quality Assurance for Scientific Paper–Code Alignment

Tim Baumgärtner and Iryna Gurevych

Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science,
TU Darmstadt and National Research Center for Applied Cybersecurity ATHENE

Code: <https://github.com/ukplab/scicoqa>
Data: <https://hf.co/datasets/ukplab/scicoqa>
Blog: <https://ukplab.github.io/scicoqa>

Abstract

Discrepancies between scientific papers and their code undermine reproducibility, a concern that grows as automated research agents scale scientific output beyond human review capacity. Whether LLMs can reliably detect such discrepancies has not been systematically measured. To this end, we present SCICoQA, a dataset of 635 paper-code discrepancies (92 real, 543 synthetic) for this cross-modal verification task. Across 22 evaluated models, even the best-performing LLMs, Gemini 3.1 Pro and GPT-5 Mini, detect only 46.7% of real-world discrepancies, revealing a critical gap in automated scientific quality assurance. We construct SCICoQA from GitHub issues and reproducibility papers, and propose a synthetic generation pipeline to scale beyond AI to Physics, Quantitative Biology, and other computational sciences. We further introduce a taxonomy of discrepancy types and categories to characterize the occurring mismatches. Our analysis shows that models particularly struggle with omitted paper details, long-context inputs, and papers outside their pre-training corpus.¹

1 Introduction

The “reproducibility crisis” in AI and across science casts doubt on the reliability of research (Baker, 2016; Hutson, 2018). To address this, the computational sciences have long recognized that a paper alone is insufficient and that publishing code, data, and instructions is a prerequisite to ensure experimental findings are reproducible (Buckheit and Donoho, 1995; Peng, 2011; Raff, 2019; Pineau et al., 2021). However, the availability of code guarantees neither reproducibility nor consistency with the scientific text (as showcased in Fig. 1). In practice, implementation details can diverge from their descriptions, introducing performance variations that go unreported (Henderson et al., 2018). These

¹Data also available at: <https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/4994>

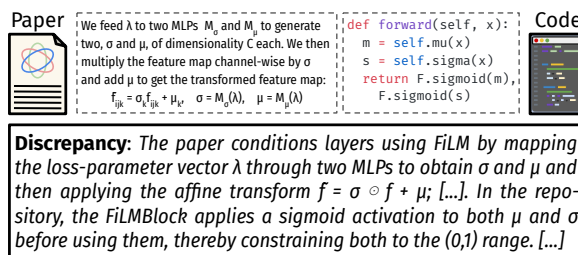


Figure 1: Example from SCICoQA, showing a specific model implementation in the paper and its implementation in the code (simplified for readability). The paper’s description and the code’s implementation mismatch, creating a *paper-code discrepancy*.

discrepancies manifest in troubling ways: from “mathiness,” where equations simulate technical depth while actual gains stem from undocumented tricks (Lipton and Steinhardt, 2019), to evaluation metrics that differ in implementation (Post, 2018; Marie et al., 2021), rendering scientific comparisons invalid.

Reproducibility issues and paper-code inconsistencies are typically only detected during reproduction efforts and post-publication, resulting in a waste of resources and eroding trust in science. While checklists during peer review can encourage authors to provide more reproducible code (Pineau et al., 2021; Dodge et al., 2019), ideally, reviewers check the correctness of the implementation and ensure reproducibility. However, with the rapid growth of submission numbers, reviewers are already under severe time pressure (Rogers and Augenstein, 2020) and conducting intricate code reviews is time-consuming (Collberg and Proebsting, 2016).

Furthermore, the reliance on manual review is becoming increasingly impractical as science begins to scale via agentic systems that develop ideas, generate and execute code, and produce scientific articles autonomously (Boiko et al., 2023; Lu et al., 2024; Tang et al., 2025; Weng et al., 2026; Lu

et al., 2026), a trend recently exemplified by peer-reviewed acceptance to an ICLR workshop (Yamada et al., 2025). While this direction has the potential to accelerate scientific discovery, it makes oversight increasingly challenging as humans cannot verify the rapidly expanding volume of output (Bowman et al., 2022). Moreover, the faithfulness of these systems is not guaranteed as LLMs suffer from limited context (Liu et al., 2024b), experience compounding errors (Dziri et al., 2023; Mirzadeh et al., 2025), and struggle with independent self-correction (Tyen et al., 2024; Wu et al., 2024). Consequently, ensuring the reliability of science at this scale demands automated tools capable of verifying that the code faithfully implements the methods reported in the scientific paper.

To this end, we introduce SCICOQA, the first benchmark for detecting discrepancies between scientific papers and their code as a cross-modal verification task. Benchmarking LLMs on this task allows us to measure their potential for automated quality assurance workflows in science. To construct our dataset, we leverage issues reported in code repositories and papers dedicated to reproducing existing work, such as those from reproducibility challenges and conference tracks. While these sources are realistic and of high quality, they are sparse and limited to CS and AI. Therefore, we scale the data to various computational science domains (e.g., Physics and Quantitative Biology) by generating modifications in codebases, creating discrepancies between the paper and its implementation. Finally, we benchmark open-weight and proprietary LLMs to determine whether current systems can be deployed for detecting paper-code discrepancies. Our evaluation reveals that the best-performing models, while precise in identifying discrepancies, suffer from insufficient recall. The best models in our evaluation detect only 46.7% of real-world cases, highlighting the need to develop more effective models and systems for this critical task.

2 Related Work

Error Detection in Science Recently, LLMs have been employed to detect errors in scientific papers. Early work created short CS papers with logical errors and prompted LLMs to identify the issues (Liu and Shah, 2023). However, their reliance on manually crafted papers limits scalability. Therefore, Dycke and Gurevych (2026) and

Xi et al. (2025) propose dedicated pipelines that generate modifications to papers, invalidating their key claims. They evaluate whether the soundness score of generated reviews decreases or whether LLMs can spot these errors. Both find that LLMs struggle to detect most introduced errors. Besides constructing erroneous papers, recent research focuses on identifying real-world issues. Specifically, Son et al. (2025) and Zhang and Abernethy (2025) use the authors’ retraction notes from WithDrawXiv (Rao et al., 2024) and PubPeer to obtain faulty papers. While realistic, these notes are often un-specific (e.g., “Theorem 2 is incorrect”). Recently, Bianchi et al. (2025) developed a system to detect inconsistencies within the text of AI papers (e.g., incorrect calculations in tables or errors in equations) and then manually verified the detections. They found their system to be relatively precise but to struggle with recall, i.e., detecting all issues in a paper. They also noted an increasing trend of errors in scientific publications over the years, highlighting the importance of quality assurance in science.

While these works share a similar spirit to ours, i.e., identifying errors in science, they are limited by solely analyzing the paper. SCICOQA extends beyond the scientific text and considers the code as well, enabling the detection of cross-modal error types, such as implementations that deviate from the paper description or omissions in the paper or the code that are crucial for understanding and reproducibility. Conceptually, our task can be viewed as a form of cross-modal claim verification (Thorne et al., 2018; Wadden et al., 2020), where the evidence source is a codebase rather than text. Furthermore, SCICOQA utilizes constructed errors to scale the quantity and diversity in the computational science domain, while also providing challenging real-world discrepancies between paper and code.

Code-Comment Inconsistency Our task is also related to Code-Comment Inconsistency (CCI) detection, where natural language comments that no longer match a piece of code need to be identified, e.g., after adding a feature to a function. Early CCI research relied on rule-based systems to identify outdated comments (Ratol and Robillard, 2017). Subsequently, bi-encoders and cross-encoders have been employed (Rabbi and Siddik, 2020; Steiner and Zhang, 2022). Panthaplackel et al. (2021) further extended CCI to dynamic settings, learning

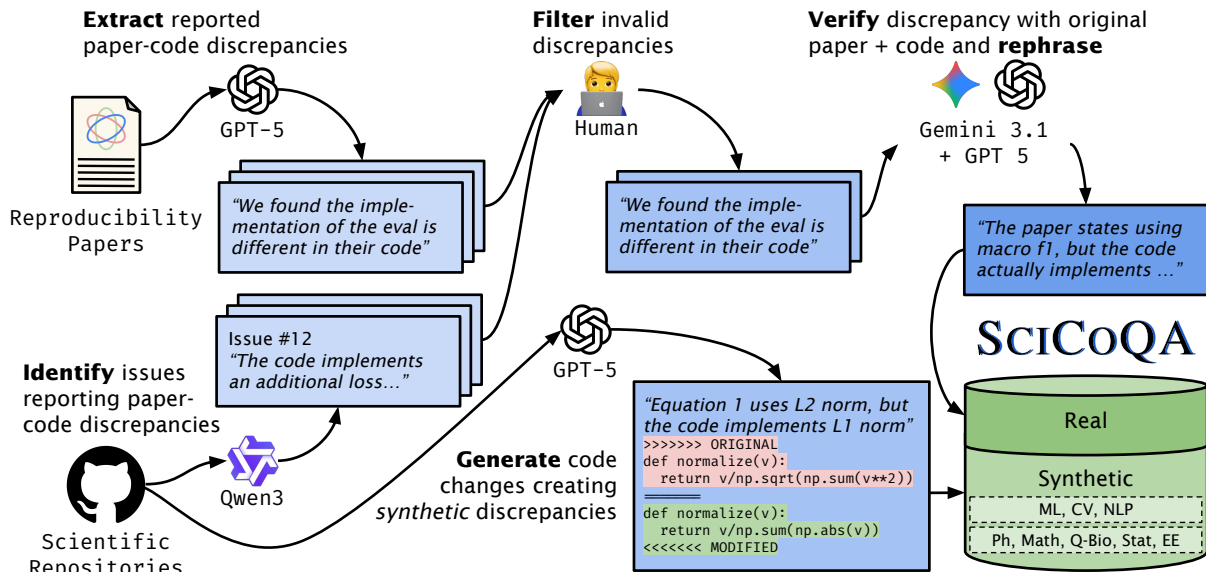


Figure 2: Overview of the data collection process of SCiCOQA. We source real-world data from reproducibility papers and GitHub issues. For the former, paper-code discrepancies are extracted from the paper with GPT-5, for the latter, issues are pre-filtered using Qwen3. Next, all candidates are manually filtered to remove any that do not fit our discrepancy definition. Finally, all paper-code discrepancies are verified with Gemini 3.1 and GPT-5. For synthetic data, we generate discrepancies using GPT-5 for AI and other computational domains.

to detect inconsistency arising from code updates. Recently, LLMs have been deployed for both detection and rectification (Dau et al., 2024; Rong et al., 2025). These methods deal with local and small inputs, typically taking a single function and an inline comment or docstring as input to detect the CCI. SCiCOQA is substantially more ambitious: it requires performing a global alignment by reasoning over a dense scientific paper and a long, multi-file code repository to find inconsistencies.

Science Automation Automation in science is rapidly accelerating (Zheng et al., 2025). Early work targeted specific components of the research cycle, including ideation (Si et al., 2025; Baek et al., 2025), literature review (Wang et al., 2024; Liang et al., 2025), coding (Tian et al., 2024; Gandhi et al., 2025), and writing (Liang et al., 2024; Weng et al., 2025). A recent trend involves generating entire code repositories from a paper (Starace et al., 2025; Seo et al., 2026; Xiang et al., 2025) or automating the scientific method end-to-end, giving rise to “AI Scientists.” However, for both these tasks, validating whether the generated code faithfully reflects the paper is challenging. For example, PaperBench (Starace et al., 2025) relies on high-quality, but expensive, manual rubrics specific to each paper to verify implementation nuances. Others use generic LLM-judges without ground truth validation (Tang et al., 2025), evaluate whether the code

runs (Weng et al., 2026; Xiang et al., 2025), or assign reproducibility scores based on execution outcomes (Siegel et al., 2024; Hu et al., 2025; Ye et al., 2025). Critically, these metrics can be misleading; a generated codebase may execute perfectly and achieve high performance while implementing a method that differs fundamentally from the paper’s scientific description (Beel et al., 2025). SCiCOQA addresses this gap by providing a ground truth dataset of paper-code discrepancies, enabling the development of robust quality assurance models that can verify the faithfulness of (generated) papers to (generated) code. More broadly, while code benchmarks in both software engineering (Jimenez et al., 2024) and science (Tian et al., 2024) predominantly evaluate code generation, SCiCOQA evaluates the complementary task of code comprehension and verification against a scientific specification. Thus, SCiCOQA takes a step toward verification and reproducibility tooling essential for sustaining peer review integrity and science at scale (Wei et al., 2025b; Woodruff et al., 2026; You et al., 2026).

3 SCiCOQA

The objective of the SCiCOQA task is to identify discrepancies between a paper and its code. We first define what constitutes a discrepancy:

Paper-Code Discrepancy We define a paper-code discrepancy as a semantic conflict between the scientific method described in the publication and its actual implementation in the codebase, such that the code does not faithfully reproduce the reported method. This mismatch must be meaningful, implying a fundamental alteration to the scientific logic, experimental protocol, or mathematical formulation described in the text. These discrepancies manifest as three distinct types: *Differences*, where the code implements a logic distinct from the paper’s description (e.g., L1 vs. L2 normalization), *Paper Omissions*, where the code includes critical components missing from the text, or *Code Omissions*, where a step described in the paper is absent from the repository. We distinguish these discrepancies from engineering artifacts: We exclude bugs, as these are independent of the paper’s scientific description. Similarly, mismatches in default hyperparameters are not considered discrepancies if the code supports the paper’s settings via configuration files or CLI arguments. Finally, we exclude trivial implementation details that are standard engineering practices typically omitted from scientific descriptions (e.g., adding noise to a denominator for numerical stability).

The paper-code discrepancies, according to our definition, can arise from distinct authorship (where the paper writer differs from the engineer implementing the code), simplifications made in the text for readability, or code updates or experiments that were not propagated to the manuscript.

3.1 Data Collection

Fig. 2 provides an overview of our data collection. To obtain real-world instances of paper-code discrepancies, we draw from GitHub issues and reproducibility papers. Furthermore, synthetic discrepancies are generated in real scientific codebases.

GitHub Issues We first identify repositories that reference a research paper in their homepage or description field on GitHub, restricted to projects published between 2020 and 2025 and repositories with at least one issue (see §A.1 for details).² For each, we crawl all associated issues, yielding 1,890 repositories with a total of 10,636 issues. To pro-

²The crawled repositories are not always the official implementation by the authors, but sometimes reproductions. They were published, for example, when there is no official implementation or the reimplementation uses a different framework. We still consider these repositories, as researchers might use them to base their experiments on.

cess this large volume efficiently, we automatically classify issues with Qwen3 4B Thinking (Yang et al., 2025), prompting the model to determine whether an issue reports a paper-code discrepancy or not, resulting in 232 candidates (prompt in §F.1). All candidates are subsequently manually filtered to ensure they meet our definition of a paper-code discrepancy, yielding 59 discrepancies.

Reproducibility Papers We collect reproducibility reports from the ML Reproducibility Challenge (Pineau et al., 2019; Sinha et al., 2020, 2022, 2023), which invites participants to reproduce papers from leading ML, CV, and NLP conferences (e.g., NeurIPS, ICML, ICLR, CVPR, ICCV, ACL, EMNLP) and papers from the reproducibility tracks of SIGIR and ECIR from 2020–2025. We retain only reports that reproduce a single, open-access paper with an open-source implementation, resulting in 171 reproducibility papers. From each report, we extract mentions of paper-code discrepancies using GPT-5, resulting in 132 candidates (prompt in §F.3). All extracted discrepancies are manually verified to confirm adherence to our definition, yielding 65 discrepancies.

Validation and Phrasing Finally, all manually filtered discrepancies are validated by Gemini 3.1 Pro and GPT-5, checking the discrepancy’s existence given the original paper, codebase, and the GitHub issue or the description extracted from the reproducibility paper. For the cases where the LLMs disagreed, we manually verified the discrepancy. During this step, the LLMs are further tasked with generating a standardized discrepancy description. We prompt the models to output 3-8 sentences and to state what the paper describes, what the code implements, and where the difference is. We use these generated descriptions as the ground truth to ensure the same format and level of verbosity for paper-code discrepancies (prompt in §F.2 and §F.4). We use two state-of-the-art LLMs to perform the validation and rephrasing to prevent potential bias in the selection and phrasing (see §C.5 for an analysis). The verification step ensures adherence to the discrepancy definition, acting as a high-precision filter. Therefore, we retain only discrepancies that can be verified given the explicit evidence, and prioritize the precision and objectivity of the ground truth. In total, we obtain 92 discrepancies from 72 papers, where 42 discrepancies originate from GitHub issues and 50 from reproducibility papers.

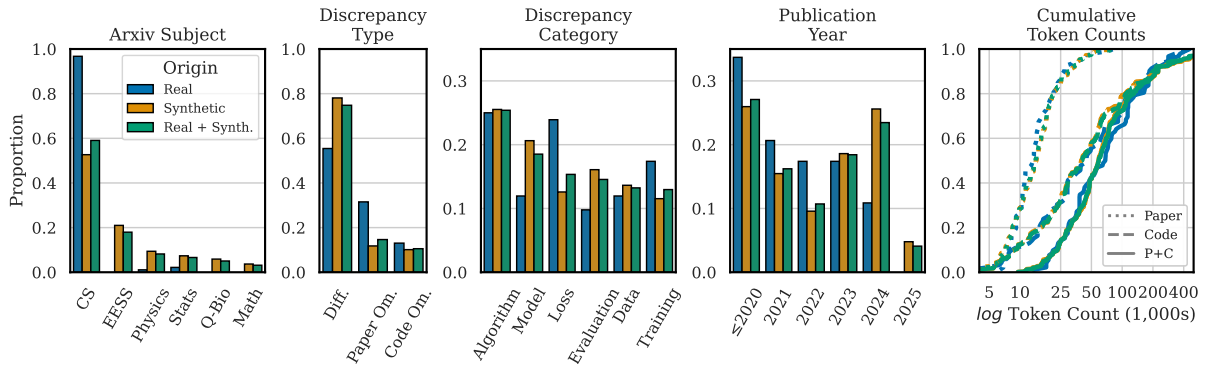


Figure 3: Analysis of the real (blue), synthetic (orange), and combined (green) discrepancy data. The y-axis shows the proportion of the data. The arXiv subjects stand for Computer Science (CS), Electrical Engineering and Systems Science (EESS), Physics, Statistics (Stats), Quantitative Biology (Q-Bio), and Mathematics (Math). The *Discrepancy Category* chart shows only the distribution over computer science papers.

Synthetic Data To scale the data in size and beyond the CS/AI domain, we employ synthetic data generation. Specifically, based on our GitHub crawl, we sample repositories linked to arXiv papers and those with permissive code licenses (i.e., MIT, Apache 2.0, CC-BY, BSD), as we will be redistributing their code partially. We randomly select 102 repositories where the paper’s arXiv subject is CS (balanced over Machine Learning (ML, cs.LG), Computer Vision (CV, cs.CV), and Natural Language Processing (NLP, cs.CL)), and 102 for non-CS papers.³ Next, we prompt GPT-5 with the paper and code to generate five code diffs according to our discrepancy definition (prompt in §F.5). We then sample up to three of these changes, ensuring that they do not manipulate the same file and that the generated diff can be found by exact match in the original code. Finally, we add 13 additional discrepancies that are labeled as Paper Omissions as we later find these to be most critical (see §3.2 and §5). This yields a total of 543 discrepancies across the 204 papers, of which 286 discrepancies are from the CS domain, and 257 are from other computational domains, such as Electrical Engineering and Systems Science (114), Physics (51), or Statistics (40).

Example data, including discrepancy type and category annotations, are shown in Table 12. Examples from the synthetic subset, including the code changes, are shown in Table 13. We release SCICOQA publicly under CC-BY-4.0.

³We manually checked the repositories and removed codebases from our sample that were not suitable, such as repositories for survey papers or with very few files.

3.2 Data Analysis

Domain Fig. 3 analyzes several dimensions of SCICOQA. While our real-world data is mostly from computer science (specifically from AI and its subdomains such as ML, CV, and NLP), the synthetic data contains papers and code from Electrical Engineering and Systems Science, Physics, Statistics, Quantitative Biology, and Mathematics, making the data diverse across computational sciences.

Taxonomy We further analyze the discrepancy type (as defined in §3) by annotating each discrepancy in the real set as *Difference*, *Paper Omission*, or *Code Omission*. For the synthetic data, we provide type definitions in the prompt and generate the label along with the discrepancy. We find that 55% of discrepancies in the real data are Differences, which is also the majority class in the synthetic data, accounting for 78% of the data. Beyond the discrepancy type, we analyze the affected component within the research pipeline. We introduce a taxonomy of six discrepancy categories, namely: *Algorithm*, *Model*, *Loss*, *Evaluation*, *Data*, and *Training*. Table 1 provides an overview of our proposed taxonomy. We apply this taxonomy to real-world data and the CS subset of synthetic data, excluding the other domains where these ML-specific concepts may not always apply cleanly. Similar to the discrepancy type annotation, we manually label the real instances and automatically generate labels for the synthetic ones. In the real data, discrepancies in the Algorithm and Loss categories dominate (25% and 24%), whereas in the synthetic data Algorithm is the largest category (26%) and Model is the second-largest (21%).

	Label	Definition
Type	Difference	The code implements logic distinct from the paper’s description
	Paper Omission	The code includes critical components not described in the paper
	Code Omission	A step described in the paper is absent from the code
Category	Algorithm	Changes to step order, operations, or core logic
	Model	Architectural or weight initialization deviations
	Loss	Alterations to loss definitions or terms
	Evaluation	Modifications to evaluation logic, metrics, or scripts
	Data	Dataset usage, preprocessing, augmentation, or filtering
	Training	Changes to the learning process, schedule, or optimization

Table 1: The SciCoQA discrepancy taxonomy. *Types* describe the structural nature of the mismatch; *Categories* describe which component of the research pipeline is affected.

Input Length Finally, we visualize the token count distributions of the papers and codebases. The medians are 14,200 tokens for papers, 39,272 for codebases, and 56,903 when concatenated. Further, 73 out of 276 papers have more than 100k tokens combined with their codebase, making SciCoQA a challenging task for measuring the models’ long-context abilities. In the appendix, we analyze the dataset by programming languages and the synthetic code diffs (§B.3 and §B.2).

4 Experiments

Given a paper and its code, we prompt the model to generate a list of discrepancies between the two (prompt in §F.6). We then parse the model output into individual discrepancies. The generation prompt contains the same instructions as those used to construct the ground-truth discrepancies. For implementation details, we refer to §A. We further ablate the discrepancy prediction experiment by providing only the code as input. This allows us to quantify the contribution of the paper, distinguishing between discrepancies that require cross-modal reasoning and those that can be inferred from the code (e.g., through comments or readmes) or from the model’s parametric knowledge of the paper.

Evaluation We employ LLM-as-a-Judge (Zheng et al., 2023) to evaluate whether the predicted discrepancy matches the reference. Inspired by Wei et al. (2025a), we use a reasoning model, specif-

ically GPT-OSS 20B (OpenAI, 2025b) (prompt in §F.7), because it is open-weight, enabling reproducibility, and offers a favorable speed-performance trade-off. For the real data where two ground truth descriptions exist (generated by Gemini 3.1 Pro and GPT-5), we consider matches against either as correct. To verify the evaluation setup, we annotate the predictions from 20 discrepancies, where 10 originate from GitHub and 10 from reproducibility papers. To streamline our evaluation, we compute the top three most similar predicted discrepancies per model using EmbeddingGemma (Vera et al., 2025). We then manually assess whether each prediction matches the reference discrepancy. In total, this yields 1,039 annotations, of which 143 predictions match the reference. On this data, the LLM judge achieves an F1 score of $87.5\% \pm 1.1$, showing strong alignment with our annotations.⁴ We further validate the setup by running Qwen3 32B as an alternative judge. We find a Cohen’s Kappa of $\kappa = 0.838$ between them and an F1 score of 91.9%. However, since the F1 score of Qwen3 32B against our human evaluation is only 70.0%, we retain GPT-OSS 20B as a judge. While the evaluation is a binary classification (reference and predicted discrepancies match or do not match), we refer to the performance as *recall*, as the predictions may contain unannotated, but valid, discrepancies, and we only evaluate whether the annotated ones are detected.

Models We evaluate several state-of-the-art model families, including commercial and open-weight models, as well as reasoning, instruction-tuned, and code-specific variants, specifically: GPT-5, Gemini, GPT-OSS, Qwen3, DeepSeek R1, Nemotron, and Mistral (exact model variants in Table 3).

5 Results

Fig. 4 reports the recall of the top-performing models (full results in Table 10). Overall, Gemini 3.1 Pro and GPT-5 Mini perform best, achieving a recall of 46.7%, leaving considerable room for improvement. On the synthetic data, GPT-5 detects 70.0% of discrepancies. Crucially, we observe a strong correlation ($r = 0.94$) between recall on real and synthetic data, validating the synthetic subset as a reliable proxy for ranking models (see §C.4).

⁴We observe a slight variance in predictions despite a fixed seed in vLLM; we therefore report the mean and standard deviation over five runs.

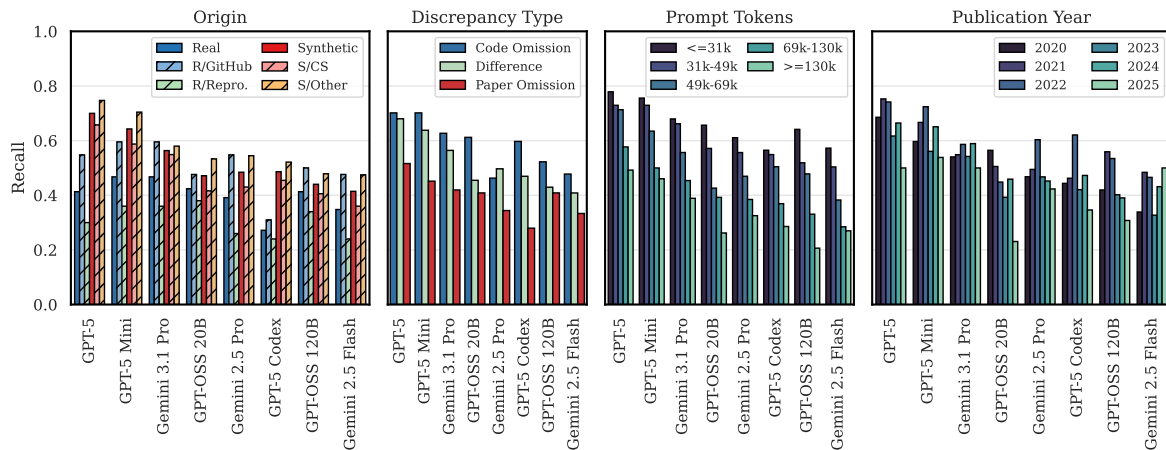


Figure 4: Results of the top 8 best-performing models (sorted by average recall on the real and synthetic data) on the discrepancy dataset by different analyses. From left to right: we analyze the origin of the discrepancy, the type of discrepancy, the number of tokens in the prompt, and the publication year of the paper.

While general scaling laws can be observed (i.e., models trained with more compute and data perform better), GPT-5 Mini and GPT-OSS 20B are exceptions in our task, slightly outperforming their larger variants. We attribute this partly to the self-preference bias of the judge for GPT-OSS, rating its own outputs better (Panickssery et al., 2024), and partly to the model’s verbosity. GPT-5 Mini and GPT-OSS 20B on average make 5.3 and 5.8 predictions per paper, while GPT-5 and GPT-OSS 120B make only 4.6 and 5.0 predictions, giving the smaller variants a higher chance to match the discrepancy at the expense of precision. Additionally, we find GPT-5 Codex to be inferior, despite being a larger model than GPT-5 Mini (Codex is based on GPT-5). While Codex is generally superior in code generation, for SCICOQA, code and natural language understanding are both crucial, and we conjecture that the general instruction-following and reasoning abilities of GPT-5 and GPT-5 Mini are more helpful than specialized coding knowledge.

Origin and Type Analyzing detection rates reveals a clear hierarchy: Code Omissions are the easiest to detect, followed by Differences, while Paper Omissions are the most challenging. This order also explains the performance gap between data sources. Discrepancies from GitHub are easier to detect because they primarily consist of Differences (71.4%) and Code Omissions (19.0%). In contrast, discrepancies from reproducibility papers are more challenging as they are often dominated by Paper Omissions (50%). The challenging nature of Paper Omissions likely stems from the asymmetry between modalities: the paper acts as a specifica-

tion, where all described components must exist in the code. Code Omissions and Differences benefit from this explicit grounding. Conversely, the code contains many implementation details not required in the paper; thus, detecting Paper Omissions requires the harder task of distinguishing deviations from permissible engineering artifacts without a reference in the paper.

Input Length We further analyze the performance by the number of input tokens.⁵ We split the dataset by prompt length into five approximately equally sized bins. We find a consistent pattern where the longer the input, the lower the performance becomes, consistent with findings in prior work (Zhang et al., 2024; Levy et al., 2024). This sensitivity to input length also suggests an explanation for the performance gap between domains in the synthetic data: models struggle more with CS papers, which is likely driven by the larger repository size of CS vs. non-CS repositories (median: 53k vs. 31k tokens). This increased volume creates a larger search space, making the task more challenging. Unlike long-context evaluations that primarily test retrieval of isolated facts from extended contexts (Kamradt, 2023; Hsieh et al., 2024), SCICOQA requires cross-document reasoning over real scientific papers and multi-file codebases to identify semantic mismatches, making it a natural testbed for long-context abilities (similar to the question answering task of Liu et al. (2024a)).⁶

⁵§C.3 conducts a further performance analysis by the position of the relevant code file in the prompt.

⁶The synthetic data generation can further be used to inject discrepancies at controlled positions in the codebase, enabling targeted analysis of context-length effects.

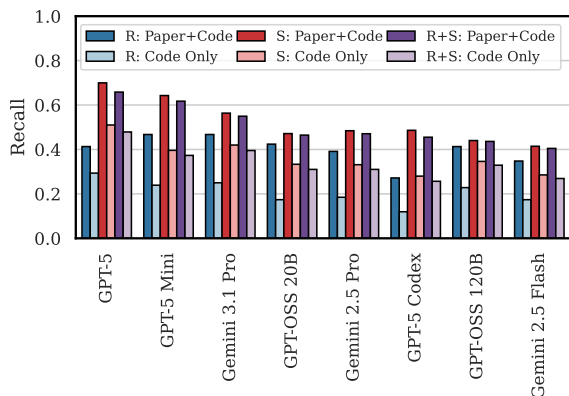


Figure 5: Performance of top 8 models when given paper and code, and only the code, split by data origin: Real (R), Synthetic (S), and combined (R+S).

Publication Year Lastly, we split the data by publication year. Among the top models, the one with the most recent knowledge cutoff is the Gemini family, which was trained using data up to January 2025 (other models have cutoffs in 2024). Therefore, the 2025 split of our data is largely outside the pre-training data, although Gemini models have partial coverage. Most models (except Gemini 2.5 Flash) perform worst on the most recent data. This suggests that models benefit from the inclusion of specific papers and codebases in their pre-training sets, lending support to concerns regarding data contamination discussed in the literature (OpenAI, 2023; Zhou et al., 2023; Sainz et al., 2023). This further emphasizes the importance of our synthetic data generation pipeline, which enables us to continuously update our dataset with uncontaminated data that is not part of the pre-training data of future models.

Open-Weight Models Beyond the top-performing models, we find the other open-weight models (Qwen3, Nemotron, DeepSeek R1, Mistral) severely limited. Nemotron 49B and Qwen3 30B Coder perform best, but only achieve an average recall on the combined data of 23.9% and 23.5%, respectively.

In summary, SCICOQA provides a significant challenge for state-of-the-art LLMs. We find the recall of all models to be limited; the best-performing models can only detect 46.7% of all discrepancies in the real-world data. Consequently, these systems cannot yet be relied upon to ensure the faithfulness of scientific publications to their codebases.

5.1 Code-Only Ablation

To test the multimodality of our data, we remove the paper from the input, leaving only the code. This makes it more difficult to identify discrepancies, but not impossible, since the repositories also frequently contain abstracts, summaries, or details of the paper in their readme files. Furthermore, scientific papers are typically part of the pre-training data; therefore, while the model does not directly receive the paper as input, it may be able to recall it from its pre-training. We show the results of the top models in this experiment in Fig. 5 (full results in Table 11). When removing the paper from the input, all models perform worse, confirming that the paper is necessary to perform the task. We observe an average drop of 19.2 percentage points, a relative drop of 48.3% on the real data. For the synthetic data, the drop is less pronounced; on average, the models’ performance drops by 16.3 percentage points (relative drop of 30.8%).

5.2 Validation of Unlabeled Discrepancies

To better understand the precision of the models, we analyze GPT-5’s, Gemini 2.5 Pro’s, and GPT-OSS 20B’s generations. We select predictions from 20 papers in NLP and CV that were not matched to a ground truth, yielding a total of 224 evaluations. We select these papers to leverage our own domain expertise, as validating discrepancies is challenging and requires a detailed understanding of the paper and its implementation. We report results of this precision analysis in Table 4.

Overall, we find that Gemini 2.5 Pro does not make significant errors, achieving a precision of 94.1%. Its three errors were due to preprocessing (an omitted term in a formula during OCR), an ambiguity within the paper (a nuance even noted in the model’s generation), and the incorrect recall of another paper’s formula. GPT-5 follows (85.7%), primarily failing due to paper ambiguities or misinterpretations, or incorrect assumptions about library functions. Finally, GPT-OSS 20B struggles most (67.0%), frequently misunderstanding the code logic (e.g., failing to align variables named differently in the paper and code or misinterpreting conditional execution paths).

Taking inspiration from information retrieval benchmarks where incomplete relevance judgments are common (Buckley and Voorhees, 2004), we construct a pooled ground truth from the 20 papers. We aggregate the discrepancies originally

Model	GPT-5	Gemini	GPT-OSS
True Positives	66	53	72
False Positives	9	3	31
Precision	88.0	94.6	69.9
Recall	51.2	41.1	55.8
F1	64.7	57.3	62.1

Table 2: Performance of GPT-5, Gemini 2.5 Pro, and GPT-OSS 20B on 129 pooled discrepancies from the annotation of unlabeled predictions and discrepancies from SCICOQA from 20 NLP and CV papers.

annotated in SCICOQA with the verified ones identified by the models (excluding minor ones, which are mostly discrepancies related to mismatching settings that can typically be resolved via configuration or the CLI). This results in a comprehensive set of 129 distinct paper-code discrepancies, and we report the results in Table 2. We find that the number of discrepancies uniquely identified by the models is high (see also Table 4), yet recall remains low (41–56%), confirming our main experiment’s findings that model recall is limited. GPT-5 achieves the best precision-recall tradeoff, obtaining an F1 score of 64.7%. We evaluate all 22 models against this pooled ground truth in §C.6, confirming that recall remains the primary bottleneck. We make the annotations from the precision analysis public, so future work can leverage them as well for detailed analysis.

6 Conclusion

We introduced SCICOQA, a dataset of 635 discrepancies designed to evaluate the alignment between a scientific paper and its code. Our analysis reveals a critical gap: while models like GPT-5 and Gemini demonstrate high precision, they suffer from insufficient recall, detecting only 46.7% of real-world discrepancies. For a verification use case, low recall is a particularly consequential failure mode: missed discrepancies silently provide false assurance, while false positives can be filtered by a human reviewer. We further find that models struggle with Paper Omissions, where code logic is absent from the paper text, and fail to maintain performance on recent, uncontaminated publications. Consequently, while current LLMs show promise as assistants, they cannot yet serve as autonomous verifiers of paper-code faithfulness. As research scales toward “AI Scientists,” SCICOQA provides the essential ground truth to ensure these agents remain faithful to the scientific method. Be-

yond this primary use case, SCICOQA can serve multiple roles: as a natural long-context reasoning benchmark complementing synthetic retrieval probes, as a testbed for cross-modal claim verification between text and code, and as an evaluation suite for code comprehension in scientific domains. We release the dataset, model generations, evaluation pipeline, and synthetic generation code to support these applications. To further advance this field, we encourage the broader adoption of reproducibility tracks to systematically capture these discrepancies and expand future data collection.

7 Limitations

Domains Our real-world data is predominantly skewed towards Computer Science and Artificial Intelligence. Applying our real-world pipeline to non-CS domains, searching GitHub repositories for URLs linking to PubMed, bioRxiv, and medRxiv, yielded very few issues reporting paper-code discrepancies, and we did not identify comparable reproducibility tracks outside of CS with paper-code discrepancy reports that we could use. The infrastructure for surfacing paper-code discrepancies thus appears to be largely absent outside CS/AI. Our synthetic pipeline is a short-term response, which we believe provides more value than excluding these domains entirely. Long-term, we call for broader adoption of reproducibility tracks, journals, and projects, such as ReScience C (Rougier et al., 2017) and the new Replication Research journal (Röseler et al., 2025). Consequently, the distribution of errors outside CS/AI may differ from our synthetic approximations, and performance on non-CS domains should be interpreted with this synthetic nature in mind.

Discrepancy Definition Our discrepancy definition focuses on “meaningful mismatches” that impact reproducibility, explicitly excluding simple bugs, hyperparameter configuration mismatches, or documentation nits. While this focuses the task on scientific validity, it means SCICOQA does not cover the full spectrum of software engineering defects that may exist in research code.

Dataset Size With 635 discrepancies, our dataset is relatively small compared to large-scale pre-training corpora. This size is a deliberate trade-off to ensure high quality: real-world discrepancies are naturally sparsely documented, and we employed a rigorous manual verification process to guarantee

that every entry constitutes a meaningful mismatch rather than a trivial error or noise.

Synthetic GPT-5 Performance Our synthetic discrepancies are generated by GPT-5, which is also evaluated on the benchmark. We find that GPT-5 achieves disproportionately higher recall on synthetic data. This suggests that, on this subset, GPT-5 may benefit from its involvement in the data generation. However, the strong correlation between real and synthetic recall across all 22 models ($r = 0.94$, Fig. 9) confirms that relative model comparisons remain valid. Notably, excluding the GPT-5 family from the correlation analysis increases the correlation to $r = 0.98$, confirming that the overall trend is not driven by GPT-5’s generation advantage; rather, the GPT-5 family sits slightly off the shared trend line that the other models follow closely. To mitigate this limitation, absolute model performance should be assessed on the real subset, while the synthetic subset reliably supports relative model comparison and allows for targeted generation of new, uncontaminated data.

8 Ethical Considerations

Data Release We constructed the synthetic portion of SCICOQA using repositories with permissive licenses (MIT, Apache 2.0, BSD, CC-BY) to ensure respectful redistribution of code. For real-world discrepancies derived from GitHub issues and reproducibility reports, we leverage the publicly available data to create our ground truth data by contextualizing and rephrasing the original issues. None of our data contains any personal information; however, since we work with scientific publications, they are closely associated with the authors of the respective papers. We acknowledge that highlighting discrepancies in a specific author’s work may be perceived negatively. We emphasize that these discrepancies are treated as scientific artifacts for improving community reproducibility standards, rather than criticisms of individual researchers. Discrepancies frequently arise from benign causes, including concurrent code updates, simplifications for readability, or distinct authorship of paper and code, and their presence does not imply negligence or misconduct.

Automation Risks The benchmarked models in this work are intended to assist in quality assurance of scientific papers and their codebases. However, there is a risk of over-reliance; given the low re-

call rates demonstrated in our experiments (46.7% for Gemini 3.1 Pro and GPT-5 Mini), automated tools should not yet be used as the sole arbiter of a paper’s validity. Relying blindly on these systems could lead to a false sense of security regarding the reproducibility and validity of a paper. We emphasize that model outputs should support human review and reproducibility efforts, not serve as automatic reject signals or as evidence of bad faith. Automated detections, particularly given the false positive rates observed in our experiments, require expert verification before any conclusions about research quality are drawn.

Acknowledgments

We thank Jan Buchmann, Bhavyajeet Singh, and Vatsal Venkatkrishna for their helpful comments throughout the paper-writing process, as well as the anonymous reviewers and area chair for their constructive feedback.

This work has been funded by the LOEWE Distinguished Chair “Ubiquitous Knowledge Processing,” LOEWE initiative, Hesse, Germany (Grant Number: LOEWE/4a//519/05/00.002(0002)/81) and by the German Federal Ministry of Research, Technology and Space and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

- Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. [ResearchAgent: Iterative research idea generation over scientific literature with large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6709–6738, Albuquerque, New Mexico. Association for Computational Linguistics.
- Monya Baker. 2016. [1,500 Scientists Lift the Lid on Reproducibility](#). *Nature*, 533(7604):452–454.
- Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduchintala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, Aleksander Ficek, Alex Kondratenko, Alex Shaposhnikov, Alexander Bukharin, Ali Taghibakhshi, Amelia Barton, Ameya Sunil Mahabaleshwarkar, Amy Shen, Andrew Tao, Ann Guan, and 80 others. 2025. [NVIDIA Nemotron Nano 2: An Accurate and Efficient Hybrid Mamba-Transformer Reasoning Model](#). *CoRR*, abs/2508.14444.

- Joeran Beel, Min-Yen Kan, and Moritz Baumgart. 2025. **Evaluating Sakana’s AI Scientist: Bold Claims, Mixed Results, and a Promising Future?** *SIGIR Forum*, 59(1):1–20.
- Federico Bianchi, Yongchan Kwon, Zachary Izzo, Linjun Zhang, and James Zou. 2025. **To Err Is Human: Systematic Quantification of Errors in Published AI Papers via LLM Analysis.** *CoRR*, abs/2512.05925.
- Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. **Autonomous Chemical Research with Large Language Models.** *Nature*, 624(7992):570–578.
- Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosiute, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, and 27 others. 2022. **Measuring Progress on Scalable Oversight for Large Language Models.** *CoRR*, abs/2211.03540.
- Jonathan B. Buckheit and David L. Donoho. 1995. *WaveLab and Reproducible Research*, pages 55–81. Springer New York, New York, NY.
- Chris Buckley and Ellen M. Voorhees. 2004. **Retrieval Evaluation with Incomplete Information.** In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’04, page 25–32, New York, NY, USA. Association for Computing Machinery.
- Christian Collberg and Todd A. Proebsting. 2016. **Repeatability in Computer Systems Research.** *Communications of the ACM*, 59(3):62–69.
- Anh Dau, Jin L.c. Guo, and Nghi Bui. 2024. **Doc-Checker: Bootstrapping code large language model for detecting and resolving code-comment inconsistencies.** In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 187–194, St. Julians, Malta. Association for Computational Linguistics.
- DeepSeek-AI. 2025. **DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.** *CoRR*, abs/2501.12948.
- DeepSeek-AI, Qihao Zhu, Daya Guo, Zhihong Shao, Dejian Yang, Peiyi Wang, Runxin Xu, Y. Wu, Yukun Li, Huazuo Gao, Shirong Ma, Wangding Zeng, Xiao Bi, Zihui Gu, Hanwei Xu, Damai Dai, Kai Dong, Liyue Zhang, Yishi Piao, and 21 others. 2024. **DeepSeek-Coder-V2: Breaking the Barrier of Closed-Source Models in Code Intelligence.** *CoRR*, abs/2406.11931.
- Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. **Show your work: Improved reporting of experimental results.** In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.
- Nils Dycke and Iryna Gurevych. 2026. **Automatic Reviewers Fail to Detect Faulty Reasoning in Research Papers: A New Counterfactual Evaluation Framework.** *Transactions of the Association for Computational Linguistics*. To appear.
- Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. **Faith and Fate: Limits of Transformers on Compositionality.** In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Shubham Gandhi, Dhruv Shah, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. 2025. **Research-CodeAgent: An LLM Multi-agent System for Automated Codification of Research Methodologies.** In *AI for Research and Scalable, Efficient Systems*, pages 3–37, Singapore. Springer Nature Singapore.
- Gemini Team. 2025. **Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.** *CoRR*, abs/2507.06261.
- Google DeepMind. 2025. **Gemini 3.1 Pro Model Card.** Technical report, Google DeepMind.
- Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. **Deep Reinforcement Learning That Matters.** In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3207–3214. AAAI Press.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. **RULER: What’s the Real Context Size of Your Long-Context Language Models?** In *First Conference on Language Modeling*.
- Chuxuan Hu, Liyun Zhang, Yeji Lim, Aum Wadhvani, Austin Peters, and Daniel Kang. 2025. **REPRO-bench: Can agentic AI systems assess the reproducibility of social science research?** In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23616–23626, Vienna, Austria. Association for Computational Linguistics.
- Matthew Hutson. 2018. **Artificial Intelligence Faces Reproducibility Crisis.** *Science*, 359(6377):725–726.

- Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. 2024. [SWE-bench: Can Language Models Resolve Real-world Github Issues?](#) In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Greg Kamradt. 2023. [Needle In A Haystack - Pressure Testing LLMs](#). https://github.com/gkamradt/LLMTest_NeedleInAHaystack. GitHub repository.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient Memory Management for Large Language Model Serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM.
- Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. [Same task, more tokens: the impact of input length on the reasoning performance of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.
- Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, and 1 others. 2024. [Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis](#). *NEJM AI*, 1(8):AIoa2400196.
- Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, Bo Tang, Feiyu Xiong, Keming Mao, and Zhiyu Li. 2025. [SurveyX: Academic Survey Automation via Large Language Models](#). *CoRR*, abs/2502.14776.
- Zachary C. Lipton and Jacob Steinhardt. 2019. [Troubling Trends in Machine Learning Scholarship](#). *ACM Queue*, 17(1):80.
- Jiawei Liu, Jia Le Tian, Vijay Daita, Yuxiang Wei, Yifeng Ding, Yuhan Katherine Wang, Jun Yang, and Lingming Zhang. 2024a. [RepoQA: Evaluating Long Context Code Understanding](#). *CoRR*, abs/2406.06025.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Ryan Liu and Nihar B. Shah. 2023. [ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing](#). *CoRR*, abs/2306.00622.
- Llama-Nemotron Team. 2025. [Llama-Nemotron: Efficient Reasoning Models](#). *CoRR*, abs/2505.00949.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob N. Foerster, Jeff Clune, and David Ha. 2024. [The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery](#). *CoRR*, abs/2408.06292.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. 2026. [Towards end-to-end automation of AI research](#). *Nature*, 651(8107):914–919.
- Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. [Scientific credibility of machine translation research: A meta-evaluation of 769 papers](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.
- Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. [GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Mistral AI & All Hands AI. 2025. [Devstral: Fine-tuning Language Models for Coding Agent Applications](#). *CoRR*, abs/2509.25193.
- Mistral AI Team. 2025. [Mistral OCR](#). <https://mistral.ai/news/mistral-ocr>. Blog.
- OpenAI. 2023. [GPT-4 Technical Report](#). *CoRR*, abs/2303.08774.
- OpenAI. 2025a. [GPT-5 System Card](#). System card, OpenAI.
- OpenAI. 2025b. [GPT-OSS-120B & GPT-OSS-20B Model Card](#). *CoRR*, abs/2508.10925.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. [LLM Evaluators Recognize and Favor Their Own Generations](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Sheena Panthaplackel, Junyi Jessy Li, Milos Gligoric, and Raymond J. Mooney. 2021. [Deep Just-In-Time Inconsistency Detection Between Comments and Source Code](#). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 427–435. AAAI Press.
- Roger D. Peng. 2011. [Reproducible Research in Computational Science](#). *Science*, 334(6060):1226–1227.

- Joelle Pineau, Koustuv Sinha, Genevieve Fried, Rosemary Nan Ke, and Hugo Larochelle. 2019. [ICLR Reproducibility Challenge 2019](#). *ReScience C*, 5(2):5.
- Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d’Alche Buc, Emily Fox, and Hugo Larochelle. 2021. [Improving Reproducibility in Machine Learning Research \(A Report from the NeurIPS 2019 Reproducibility Program\)](#). *Journal of Machine Learning Research*, 22(164):1–20.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Fazle Rabbi and Md. Saeed Siddik. 2020. [Detecting Code Comment Inconsistency using Siamese Recurrent Network](#). In *ICPC ’20: 28th International Conference on Program Comprehension, Seoul, Republic of Korea, July 13-15, 2020*, pages 371–375. ACM.
- Edward Raff. 2019. [A Step Toward Quantifying Independently Reproducible Machine Learning Research](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5486–5496.
- Delip Rao, Jonathan Young, Thomas Dietterich, and Chris Callison-Burch. 2024. [WithdrawXiv: A Large-Scale Dataset for Retraction Study](#). *CoRR*, abs/2412.03775.
- Abhinav Rastogi, Albert Q. Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmantlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, Léonard Blier, Lucile Saulnier, Matthieu Dinot, Maxime Darrin, Neha Gupta, Roman Soletskyi, Sagar Vaze, Teven Le Scao, Yihan Wang, and 80 others. 2025. [Magistral](#). *CoRR*, abs/2506.10910.
- Inderjot Kaur Ratol and Martin P. Robillard. 2017. [Detecting fragile comments](#). In *Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering, ASE 2017, Urbana, IL, USA, October 30 - November 03, 2017*, pages 112–122. IEEE Computer Society.
- Anna Rogers and Isabelle Augenstein. 2020. [What can we do to improve peer review in NLP?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262, Online. Association for Computational Linguistics.
- Guoping Rong, Yongda Yu, Song Liu, Xin Tan, Tianyi Zhang, Haifeng Shen, and Jidong Hu. 2025. [Code Comment Inconsistency Detection and Rectification Using a Large Language Model](#). In *47th IEEE/ACM International Conference on Software Engineering, ICSE 2025, Ottawa, ON, Canada, April 26 - May 6, 2025*, pages 1832–1843. IEEE.
- Nicolas P. Rougier, Konrad Hinsén, Frédéric Alexandre, Thomas Arildsen, Lorena A. Barba, Fabien C. Y. Benureau, C. Titus Brown, Pierre de Buyl, Ozan Caglayan, Andrew P. Davison, Marc-André Delsuc, Georgios Detorakis, Alexandra K. Diem, Damien Drix, Pierre Enel, Benoît Girard, Olivia Guest, Matt G. Hall, Rafael Neto Henriques, and 26 others. 2017. [Sustainable Computational Science: the ReScience Initiative](#). *PeerJ - Computer Science Journal*, 3:e142.
- Lukas Röseler, Lukas Wallrich, Susanne Adler, Prince Oppong Boakye, Thomas Rhys Evans, Janik Goltermann, Tamarinde Haven, Jan Horstmann, Max Korbacher, Miriam Müller, and et al. 2025. [A Community Model for Rigorous and Inclusive Scholarship: Inaugural Editorial of Replication Research \(R2\)](#). *Replication Research*, 1.
- Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. [NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.
- Minju Seo, Jinheon Baek, Seongyun Lee, and Sung Ju Hwang. 2026. [Paper2code: Automating code generation from scientific papers in machine learning](#). In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, April 23-27, 2026*. OpenReview.net.
- Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. [Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Zachary S. Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebel, and Arvind Narayanan. 2024. [CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark](#). *Transactions on Machine Learning Research*, 2024.
- Koustuv Sinha, Maurits Bleeker, Samarth Bhargav, Jessica Zosa Forde, Sharath Chandra Raparthy, Jesse Dodge, Joelle Pineau, and Robert Stojnic. 2023. [ML Reproducibility Challenge 2022](#). *ReScience C*, 9(2).
- Koustuv Sinha, Jesse Dodge, Sasha Luccioni, Jessica Zosa Forde, Sharath Chandra Raparthy, Joelle Pineau, and Robert Stojnic. 2022. [ML Reproducibility Challenge 2021](#). *ReScience C*, 8(2).
- Koustuv Sinha, Joelle Pineau, Jessica Forde, Rosemary Nan Ke, and Hugo Larochelle. 2020. [NeurIPS 2019 Reproducibility Challenge](#). *ReScience C*, 6(2):11.
- Guijin Son, Jiwoo Hong, Honglu Fan, Heejeong Nam, Hyunwoo Ko, Seungwon Lim, Jinyeop Song, Jinha

- Choi, Gonçalo Paulo, Youngjae Yu, and Stella Biderman. 2025. [When AI Co-Scientists Fail: SPOT-A Benchmark for Automated Verification of Scientific Research](#). *CoRR*, abs/2505.11855.
- Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Pawardhan. 2025. [PaperBench: Evaluating AI's Ability to Replicate AI Research](#). In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.
- Theo Steiner and Rui Zhang. 2022. [Code Comment Inconsistency Detection with BERT and Longformer](#). *CoRR*, abs/2207.14444.
- Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. 2025. [AI-Researcher: Autonomous Scientific Innovation](#). *CoRR*, abs/2505.18705.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Shengzhu Yin, and 10 others. 2024. [SciCode: A Research Coding Benchmark Curated by Scientists](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. 2024. [LLMs cannot find reasoning errors, but can correct them given the error location](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13894–13908, Bangkok, Thailand. Association for Computational Linguistics.
- Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panayam, Sara Smoot, Iftexhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 70 others. 2025. [EmbeddingGemma: Powerful and Lightweight Text Representations](#). *CoRR*, abs/2509.20354.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. [Fact or fiction: Verifying scientific claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.
- Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. [AutoSurvey: Large Language Models Can Automatically Write Surveys](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Anjiang Wei, Jiannan Cao, Ran Li, Hongyu Chen, Yuhui Zhang, Ziheng Wang, Yuan Liu, Thiago S. F. X. Teixeira, Diyi Yang, Ke Wang, and Alex Aiken. 2025a. [EquiBench: Benchmarking large language models' reasoning about program semantics via equivalence checking](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33868–33881, Suzhou, China. Association for Computational Linguistics.
- Qiyao Wei, Samuel Holt, Jing Yang, Markus Wulfmeier, and Mihaela van der Schaar. 2025b. [The AI Imperative: Scaling High-Quality Peer Review in Machine Learning](#). *CoRR*, abs/2506.08134.
- Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. [CycleResearcher: Improving Automated Research via Automated Review](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Yixuan Weng, Minjun Zhu, Qiuqie Xie, QiYao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. 2026. [Deepscientist: Advancing frontier-pushing scientific findings progressively](#). In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, April 23-27, 2026*. OpenReview.net.
- David P. Woodruff, Vincent Cohen-Addad, Lalit Jain, Jieming Mao, Song Zuo, MohammadHossein Bateni, Simina Brânzei, Michael P. Brenner, Lin Chen, Ying Feng, Lance Fortnow, Gang Fu, Ziyi Guan, Zahra Hadizadeh, Mohammad T. Hajiaghayi, Mahdi JafariRaviz, Adel Javanmard, Karthik C. S., Ken-ichi Kawarabayashi, and 17 others. 2026. [Accelerating Scientific Research with Gemini: Case Studies and Common Techniques](#). *CoRR*, abs/2602.03837.
- Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, and Meng Jiang. 2024. [Large language models can self-correct with key condition verification](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12846–12867, Miami, Florida, USA. Association for Computational Linguistics.
- Sarina Xi, Vishisht Rao, Justin Payan, and Nihar B. Shah. 2025. [FLAWS: A Benchmark for Error Identification and Localization in Scientific Papers](#). *CoRR*, abs/2511.21843.

Yanzheng Xiang, Hanqi Yan, Shuyin Ouyang, Lin Gui, and Yulan He. 2025. [SciReplicate-Bench: Benchmarking LLMs in Agent-driven Algorithmic Reproduction from Research Papers](#). In *Second Conference on Language Modeling*.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob N. Foerster, Jeff Clune, and David Ha. 2025. [The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search](#). *CoRR*, abs/2504.08066.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 39 others. 2025. [Qwen3 Technical Report](#). *CoRR*, abs/2505.09388.

Christine Ye, Sihan Yuan, Suchetha Cooray, Steven Dillmann, Ian L. V. Roque, Dalya Baron, Philipp Frank, Sergio Martin-Alvarez, Nolan Koblichke, Frank J. Qu, Diyi Yang, Risa Wechsler, and Ioana Ciuca. 2025. [ReplicationBench: Can AI Agents Replicate Astrophysics Research Papers?](#) *CoRR*, abs/2510.24591.

Lei You, Lele Cao, and Iryna Gurevych. 2026. [Preventing the Collapse of Peer Review Requires Verification-First AI](#). *CoRR*, abs/2601.16909.

Tianmai M. Zhang and Neil F. Abernethy. 2025. [Reviewing Scientific Papers for Critical Problems With Reasoning LLMs: Baseline Approaches and Automatic Evaluation](#). In *NeurIPS 2025 AI for Science Workshop*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. [∞Bench: Extending long context evaluation beyond 100K tokens](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277, Bangkok, Thailand. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Tianshi Zheng, Zheyang Deng, Hong Ting Tsang, Weiqi Wang, Jiabin Bai, Zihao Wang, and Yangqiu Song. 2025. [From automation to autonomy: A survey on large language models in scientific discovery](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17733–17750, Suzhou, China. Association for Computational Linguistics.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. [Don't Make Your LLM an Evaluation Benchmark Cheater](#). *CoRR*, abs/2311.01964.

A Implementation Details

A.1 GitHub Crawl Details

We use the GitHub API⁷ to search for code repositories that provide code for a research paper. To obtain only repositories for a single paper, we search the description and homepage fields of the repositories, which are commonly used to indicate which paper the code is intended to implement. Specifically, we search for the following URLs: `arxiv.org`, `openreview.net`, `aclanthology.org`, `doi.org/10.1145`. We further limit our search to only include code repositories that have been published since 2020.

A.2 Issue Processing

For the initial GitHub issue classification with Qwen3 4B Thinking, we used a low temperature of 0.2 to minimize generation variance and promote strict adherence to the classification schema. For the verification of discrepancies from GitHub issues with GPT-5, we checked whether an issue description contained screenshots of the paper or code. If so, we replaced these with their text equivalents.

A.3 Versioning

To ensure that the paper and code still contain the discrepancy, we provide links to versioned publications for arXiv, using the version at the time of the reporting of the discrepancy, i.e., publication of the reproducibility report or creation of the GitHub issue. Similarly, for the codebases, we use the commit history to obtain the version of the code at the time of the reporting of the discrepancy. Our dataset contains the versioned links to both the paper (for arXiv) and the codebase. For the synthetic data, we take a version of the codebase as of 31-10-2025.

A.4 Paper Processing

We provide the paper text in markdown as input. We use Mistral's OCR ([Mistral AI Team, 2025](#)) to convert the PDF to markdown. We exclude figures,

⁷<https://docs.github.com/en/rest?apiVersion=2022-11-28>

Model	Paper	Model Card	Context Window	Knowledge Cutoff	Release Date
DeepSeek Coder 16B v2	DeepSeek-AI (2024)	deepseek-ai/DeepSeek-Coder-V2-Lite-Instruct	131k	Nov 2023	Jun 2024
DeepSeek R1 32B	DeepSeek-AI (2025)	deepseek-ai/DeepSeek-R1-Distill-Qwen-32B	131k	Jul 2024	Jan 2025
DeepSeek R1 8B	DeepSeek-AI (2025)	deepseek-ai/DeepSeek-R1-Distill-Llama-8B	131k	Jul 2024	Jan 2025
Gemini 2.5 Flash	Gemini Team (2025)	gemini-2.5-flash	1,047k	Jan 2025	Aug 2025
Gemini 2.5 Flash Lite	Gemini Team (2025)	gemini-2.5-flash-lite	1,047k	Jan 2025	Aug 2025
Gemini 2.5 Pro	Gemini Team (2025)	gemini-2.5-pro	1,047k	Jan 2025	Aug 2025
Gemini 3.1 Pro	Google DeepMind (2025)	gemini-3.1-pro-preview	1,047k	Jan 2025	Feb 2026
GPT-OSS 120B	OpenAI (2025b)	openai/gpt-oss-120b	131k	Jun 2024	Aug 2025
GPT-OSS 20B	OpenAI (2025b)	openai/gpt-oss-20b	131k	Jun 2024	Aug 2025
GPT-5	OpenAI (2025a)	gpt-5-2025-08-07	272k	Sep 2024	Aug 2025
GPT-5 Codex	OpenAI (2025a)	gpt-5-codex	272k	Sep 2024	Aug 2025
GPT-5 Mini	OpenAI (2025a)	gpt-5-mini-2025-08-07	272k	May 2024	Aug 2025
GPT-5 Nano	OpenAI (2025a)	gpt-5-nano-2025-08-07	272k	May 2024	Aug 2025
Devstral 24B Small	Mistral AI & All Hands AI (2025)	mistralai/Devstral-Small-2507	131k	Mar 2025	Jul 2025
Magistral 24B Small	Rastogi et al. (2025)	mistralai/Magistral-Small-2509	131k	Jun 2025	Sep 2025
Nemotron Nano 9B v2	Basant et al. (2025)	nvidia/NVIDIA-Nemotron-Nano-9B-v2	131k	Apr 2025	Aug 2025
Nemotron Super 49B v1.5	Llama-Nemotron Team (2025)	nvidia/Llama-3_3-Nemotron-Super-49B-v1_5	131k	Dec 2023	Jul 2025
Qwen3 30B Coder	Yang et al. (2025)	Qwen/Qwen3-Coder-30B-A3B-Instruct	262k	Mar 2025	Jul 2025
Qwen3 30B Instruct	Yang et al. (2025)	Qwen/Qwen3-30B-A3B-Instruct-2507	262k	Mar 2025	Jul 2025
Qwen3 30B Thinking	Yang et al. (2025)	Qwen/Qwen3-30B-A3B-Thinking-2507	262k	Mar 2025	Jul 2025
Qwen3 4B Instruct	Yang et al. (2025)	Qwen/Qwen3-4B-Instruct-2507	262k	Mar 2025	Jul 2025
Qwen3 4B Thinking	Yang et al. (2025)	Qwen/Qwen3-4B-Thinking-2507	262k	Mar 2025	Jul 2025

Table 3: Model versions including their context window, knowledge cutoff (date of most recent pre-training data), and release date.

although the captions are included. We further remove the references section from all papers.

A.5 Code Processing

To provide the code in the prompt, we obtain all files from the GitHub repository at a target date using the commit history, which we set to the date when the original GitHub issue was raised or the reproducibility paper was published. For the synthetic data, we set the cutoff to 31-10-2025. Having a versioned and static code repository is crucial to ensure reproducibility of our experiments and to make sure paper-code discrepancies have not been fixed in our dataset.

We retain only files that contain relevant content, mainly excluding dataset files. Specifically, we exclude files that are larger than 1MB and include files with the following extensions: `.c`, `.cc`, `.cpp`, `.cu`, `.h`, `.hpp`, `.java`, `.jl`, `.m`, `.matlab`, `Makefile`, `.md`, `.pl`, `.ps1`, `.py`, `.r`, `.sh`, `config.txt`, `.rs`, `readme.txt`, `requirements_dev.txt`, `requirements-dev.txt`, `requirements.dev.txt`, `requirements.txt`, `.scala`, `.yaml`, `.yml`. We obtain this list by manually inspecting the file extensions in the repositories of our dataset. We further convert all Jupyter notebooks to Python files by removing their outputs. For the prompt, we further include a list of all files in the repository.

A.6 Context Processing

The markdown paper and processed code repository are inserted into the prompt. We fill the prompt up to 90% of the context size to leave space for the reasoning (if supported by the model) and output. If the paper and code would exceed the context window, we truncate entire files from the end of the code prompt until it fits into the context window.

A.7 Model and Decoding Configuration

We deploy the open-weight models with FP16, except for the GPT-OSS models, which were only published in MXFP4. We utilize the Ollama and vLLM libraries (Kwon et al., 2023) on 1-4 A100 80GB GPUs, depending on the model’s VRAM requirements; the commercial models are queried via their respective APIs. We generally set the temperature to 1.0, and for the GPT-5, GPT-OSS, and Gemini models, set a high reasoning effort or budget to maximize performance. For the Gemini models, we set a high reasoning budget of 24k tokens; for the GPT-5 models and GPT-OSS, we set a “high” reasoning budget. Further details about each model can be obtained from the respective model card linked in Table 3.

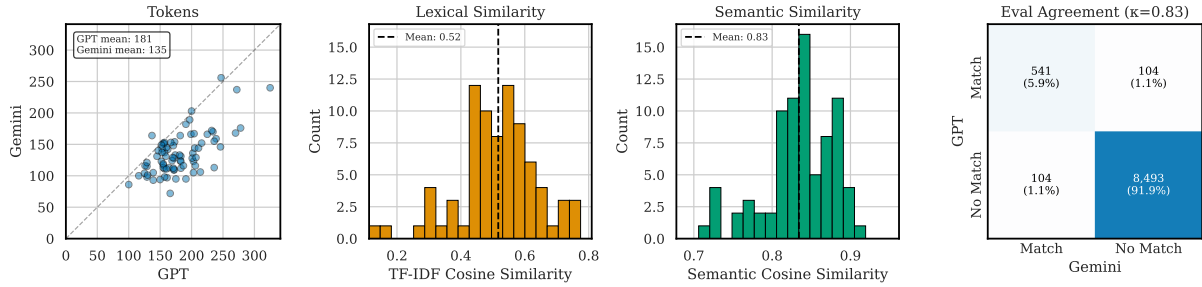


Figure 6: Comparisons between the two ground truth descriptions generated by Gemini 3.1 Pro and GPT-5.

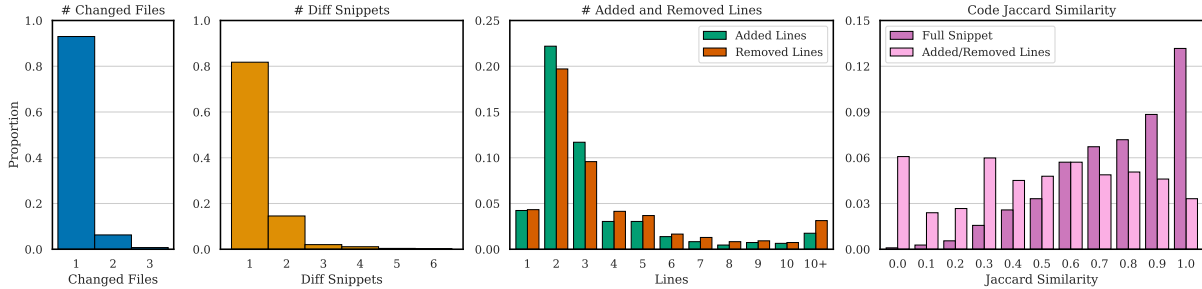


Figure 7: Quantitative analysis of synthetic code modifications. We show the distribution of number of changed files, number of generated diff snippets, added (green) vs. removed (orange) lines in those snippets, and Jaccard similarity between the original and modified code.

B Extended Analysis

B.1 Gemini and GPT Ground Truth Comparison

For the real data, we rephrased the ground truth discrepancy descriptions with Gemini 3.1 Pro and GPT-5 (see §3.1). We analyze the similarity between these two descriptions in Fig. 6. We find the GPT descriptions to be slightly longer (181 ± 40 vs. 135 ± 34 tokens). We further find moderate lexical overlap between the two ground truths (average TF-IDF cosine similarity is 0.52), but high semantic similarity (average 0.83 computed using EmbeddingGemma), indicating that the models use different wording to describe the same discrepancy. Finally, we analyze how often the GPT-OSS 20B judge matches a prediction against none, either, or both ground truths. We find that in 97.7% of cases, the judge comes to the same verdict given the two different discrepancy descriptions, yielding a Cohen’s Kappa of $\kappa = 0.83$.

B.2 Synthetic Code Analysis

We conduct a quantitative analysis of the synthetic data to validate that our generation pipeline produces minimal changes, leaving the main code intact. Fig. 7 presents the analysis of the generated diffs.

Code Changes We find the generated discrepancies in the code to be highly targeted. The vast majority affect only a single file (1.08 ± 0.29 on average), with a maximum of three files. Similarly, the number of distinct generated diffs is low, with a mean of 1.24 ± 0.61 snippets per discrepancy. This confirms that the generated errors are specific deviations in logic rather than large-scale refactors. The number of line edits (i.e., added or removed lines) further supports the subtlety of the dataset. The distribution of line counts is heavily skewed towards small edits: On average, 2.41 ± 3.25 lines are added and 3.13 ± 4.47 lines are removed. The slightly higher count for removed lines suggests the discrepancies often simplify logic (e.g., removing a normalization step) rather than adding complex boilerplate.

Code Similarity To quantify the similarity between the original and modified code, we calculate the Jaccard similarity. We measure the similarity at two levels: over the entire generated code block (which often contains a few unchanged lines) and over only the code lines that were modified. Comparing the entire code block generated by the model against the original yields a high mean similarity of 0.74 ± 0.20 . This indicates that the surrounding context and structure remain largely iden-

Model	GPT-5	Gemini	GPT-OSS
False Positives	10	3	34
<i>Code Misunderstanding</i>	1	0	20
<i>Paper Misunderstanding</i>	2	0	9
<i>Paper Ambiguity</i>	2	1	0
<i>OCR Error</i>	1	1	1
<i>3rd Party Code</i>	2	0	1
<i>No Discrepancy</i>	1	0	1
<i>Minor (Config)</i>	1	0	2
<i>Other</i>	0	1	0
True Positives	60	48	69
<i>Unique</i>	23	19	34
<i>Minor</i>	7	7	12
Total	70	51	103
Precision	85.7	94.1	67.0

Table 4: Analysis of unlabeled discrepancy predictions of GPT-5, Gemini 2.5 Pro, and GPT-OSS 20B on 20 NLP and CV papers from the real subset of SciCOQA.

tical. Comparing only the specific lines that differ (added vs. removed) yields a mean similarity of 0.50 ± 0.29 . This moderate overlap suggests the changes are often modifications of existing statements (e.g., changing an operator or variable) rather than complete replacements. Overall, the data confirms that the synthetic discrepancies are precise and minimal, strictly adhering to the “small” and “conceptually meaningful” constraints of the synthetic data generation prompt (§F.5).

B.3 Programming Language Distribution

Fig. 8 shows the distribution of programming languages in the SciCOQA dataset. Since the real data originates from the CS/AI domain, we observe Python as the dominant language, along with a small portion of C/C++, MATLAB, and CUDA files. Besides Python, other languages are present in the synthetic data, including Java, Scala, Julia, and R, although their overall presence is relatively small.

C Extended Results

C.1 Error Analysis

Table 4 provides the detailed breakdown of our human analysis from §5.2 on the predictions of GPT-5, Gemini 2.5 Pro, and GPT-OSS 20B.

C.2 Performance by Programming Language

In the synthetic dataset, the known injection points allow us to analyze detection performance across different programming languages. Since the dataset is skewed towards Python, drawing conclusions for specific low-resource languages is difficult due to

Code File Prompt Position	# Samples
$\leq 32k$	253 (39.8%)
$\leq 64k$	463 (72.9%)
$\leq 131k$	568 (89.4%)
$\leq 262k$	631 (99.4%)
$\leq 524k$	635 (100.0%)

Table 5: Number of data samples by position of the last relevant code file in the prompt.

the small sample size. Therefore, we group all non-Python instances into a unified subset of 46 samples for analysis (“Not Python”). The resulting recall scores are presented in Table 6. Most models show only small, inconsistent differences between Python and non-Python recall (within ± 5 pp). The notable exceptions are Devstral 24B Small, which collapses from 15.1% on Python to 0% on non-Python, and Gemini 2.5 Flash (+9.7 pp); conversely, GPT-5 Mini (−5.9 pp), GPT-OSS 20B (−5.4 pp), and Nemotron Nano 9B v2 (−7.8 pp) actually perform slightly better on non-Python. Beyond these aggregate trends, two specific languages stand out. MATLAB is a positive outlier: GPT-5 and GPT-5 Mini both achieve 90.0% recall, substantially above their Python performance. With $n = 10$ this is tentative, but we hypothesize that MATLAB’s mathematical syntax aligns more closely with the equations presented in scientific papers, facilitating easier logical alignment for the models. Conversely, CUDA is a near-universal weak spot: 17 of 22 models score 0% recall on CUDA ($n = 4$), including GPT-5, Gemini 3.1 Pro, and GPT-OSS 120B. Small sample sizes limit definitive conclusions for both languages.

C.3 Truncation Analysis

In §5 we analyzed the model performance by input length, finding that longer inputs degrade performance. We extend this analysis by investigating the performance as a function of the position of the last relevant code file in the prompt. The relevant code files are obtained in the validation/rephrasing step for the real data, and during discrepancy generation for the synthetic data. Table 5 shows the distribution of the position of the last relevant code file in the prompt. We find that 89.4% of samples in our data contain all relevant code files within a context window of 131k tokens, and 99.4% within a 262k context window.

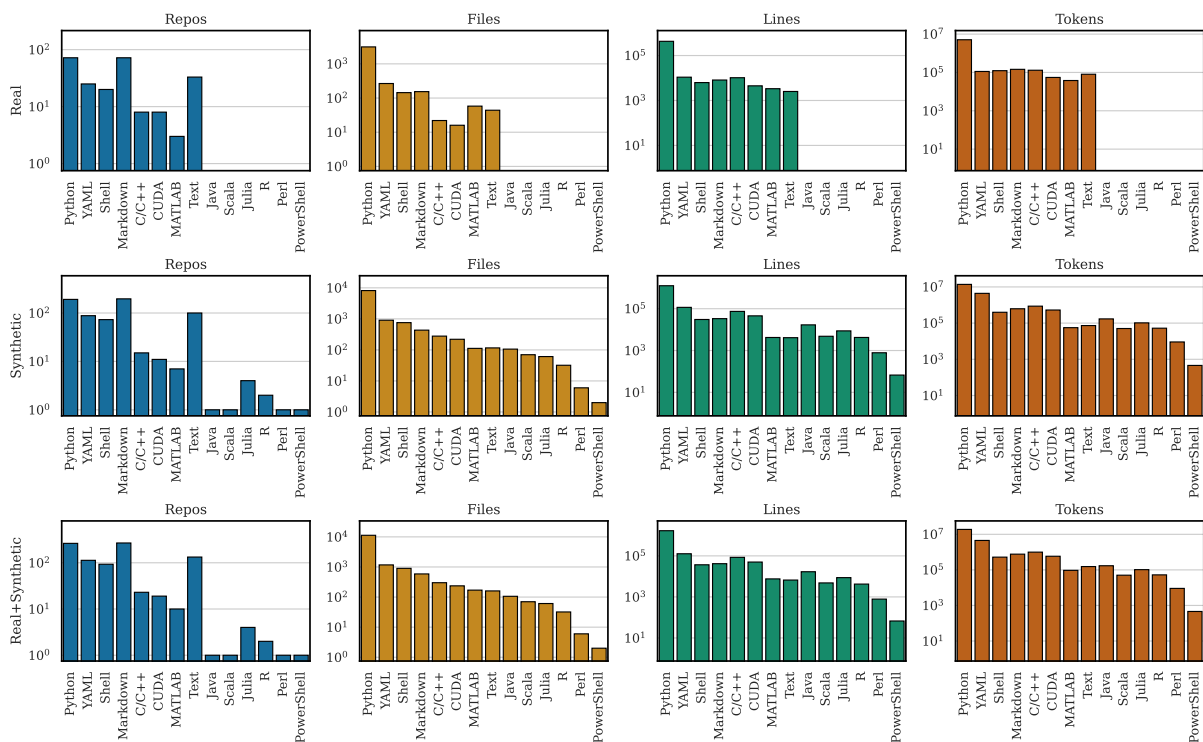


Figure 8: Distribution of programming languages in SCICOQA. The columns of the plot show the distribution by number of *Repos*, *Files*, *Lines*, and *Tokens* for each respective programming language. The rows show the distribution for the real, synthetic, and combined data.

Model	Python	Not Python	Julia	MATLAB	YAML	C/C++	CUDA	R	Shell	Py+YAML	AVG
# Discrepancies	496	46	12	10	9	5	4	3	3	1	543
GPT-5	70.4	65.2	66.7	90.0	75.0	60.0	0.0	100.0	33.3	100.0	70.0
GPT-5 Mini	63.7	69.6	75.0	90.0	66.7	60.0	25.0	100.0	33.3	100.0	64.3
Gemini 3.1 Pro	56.9	50.0	50.0	80.0	50.0	60.0	0.0	66.7	0.0	100.0	56.4
GPT-OSS 20B	46.8	52.2	50.0	60.0	66.7	60.0	50.0	33.3	0.0	0.0	47.1
Gemini 2.5 Pro	48.8	45.7	41.7	60.0	50.0	40.0	50.0	33.3	33.3	0.0	48.4
GPT-5 Codex	48.8	45.7	33.3	70.0	50.0	40.0	25.0	66.7	33.3	100.0	48.6
GPT-OSS 120B	44.6	39.1	33.3	80.0	41.7	40.0	0.0	33.3	0.0	0.0	44.0
Gemini 2.5 Flash	42.3	32.6	33.3	40.0	33.3	40.0	0.0	66.7	33.3	0.0	41.4
GPT-5 Nano	27.2	30.4	33.3	30.0	41.7	40.0	0.0	33.3	0.0	0.0	27.4
Nemotron Super 49B v1.5	24.0	23.9	16.7	20.0	16.7	40.0	50.0	33.3	33.3	0.0	23.9
Qwen3 30B Coder	24.0	21.7	25.0	30.0	41.7	0.0	0.0	33.3	0.0	0.0	23.8
Gemini 2.5 Flash Lite	24.0	19.6	16.7	20.0	50.0	20.0	0.0	0.0	0.0	100.0	23.8
Nemotron Nano 9B v2	16.1	23.9	25.0	40.0	50.0	0.0	0.0	0.0	0.0	0.0	16.8
Qwen3 30B Inst.	22.0	19.6	25.0	20.0	33.3	0.0	0.0	33.3	0.0	0.0	21.7
DeepSeek R1 32B	16.7	13.0	8.3	20.0	8.3	0.0	0.0	66.7	0.0	0.0	16.4
DeepSeek Coder 16B V2	10.1	4.3	8.3	10.0	0.0	0.0	0.0	0.0	0.0	0.0	9.6
Qwen3 4B Inst.	16.7	8.7	16.7	10.0	0.0	0.0	0.0	33.3	0.0	0.0	16.0
Magistral 24B Small	14.5	10.9	25.0	0.0	25.0	0.0	0.0	0.0	0.0	0.0	14.2
Devstral 24B Small	15.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	14.0
Qwen3 30B Think.	15.5	15.2	8.3	30.0	8.3	20.0	0.0	33.3	0.0	0.0	15.5
Qwen3 4B Think.	11.3	8.7	8.3	0.0	41.7	0.0	0.0	0.0	0.0	0.0	11.0
DeepSeek R1 8B	5.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.2

Table 6: Discrepancy recall by programming language on the synthetic SCICOQA data.

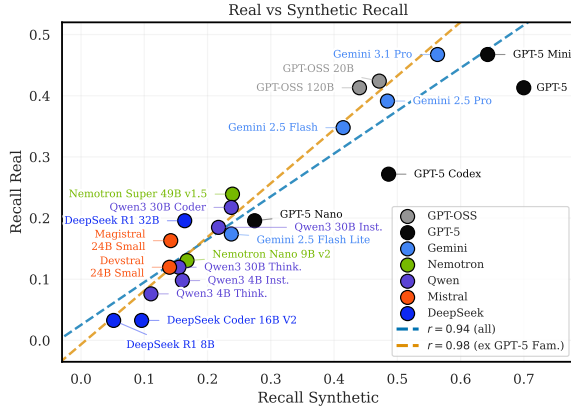


Figure 9: Correlation between model recall on the synthetic (x-axis) and real (y-axis) subsets of SCICOQA. Each point represents one of the 22 evaluated models. The dashed lines show linear fits across all models (blue) or excluding the GPT-5 family (orange). The r values denote Pearson correlation coefficients.

We further analyze recall by the position of the last relevant code file in the prompt, binning the samples and reporting results in Table 9. We observe substantial recall degradation even for models that experience virtually zero truncation. GPT-5 drops from 77.5% to 50.7% and Gemini 3.1 Pro from 64.4% to 37.3% as the relevant file appears later in the prompt.

Both the code file position distribution and the performance degradation demonstrate that the dominant factor is long-context reasoning, not truncation, consistent with our analysis in Fig. 4 and prior findings on long-context degradation (Liu et al., 2024b; Hsieh et al., 2024; Levy et al., 2024).

C.4 Real vs. Synthetic Recall

To assess the validity of our synthetic data generation pipeline, we analyze the relationship between model performance on real-world discrepancies versus synthetic discrepancies. Fig. 9 visualizes the recall and the relationship between the subsets. We observe a high positive Pearson correlation of $r = 0.94$. When excluding the GPT-5 family (i.e., GPT-5, GPT-5 Mini, GPT-5 Nano, and GPT-5 Codex) from the correlation analysis, the Pearson correlation rises to $r = 0.98$, indicating that the GPT-5 family sits slightly off the shared trend line that the other models follow closely. Apart from this caveat, the relative ranking of models remains consistent regardless of the data origin; models that demonstrate strong capabilities on our synthetic injections are reliably better at identifying real-world discrepancies. While this trend might

Model	Gemini			Δ
	+GPT	Gemini	GPT	
GPT-5	45.0	45.0	43.8	-1.3
GPT-5 Mini	48.8	47.5	43.8	-3.7
Gemini 3.1 Pro	50.0	50.0	48.8	-1.3
GPT-OSS 20B	47.5	45.0	42.5	-2.5
Gemini 2.5 Pro	41.2	40.0	40.0	0.0
GPT-5 Codex	27.5	27.5	27.5	0.0
GPT-OSS 120B	42.5	37.5	40.0	2.5
Gemini 2.5 Flash	36.2	32.5	33.8	1.3
GPT-5 Nano	21.2	18.8	18.8	0.0
Nemotron 49B v1.5	26.2	23.8	21.2	-2.5
Qwen3 30B Coder	23.8	17.5	20.0	2.5
Gemini 2.5 Flash Lite	18.8	16.2	16.2	0.0
Nemotron Nano 9B v2	13.8	8.8	11.2	2.5
Qwen3 30B Inst.	20.0	13.8	17.5	3.7
DeepSeek R1 32B	21.2	17.5	17.5	0.0
DeepSeek Coder 16B V2	2.5	2.5	1.2	-1.2
Qwen3 4B Inst.	10.0	7.5	8.8	1.2
Magistral 24B Small	16.2	15.0	13.8	-1.2
Devstral 24B Small	12.5	8.8	11.2	2.5
Qwen3 30B Think.	13.8	12.5	13.8	1.3
Qwen3 4B Think.	8.8	7.5	8.8	1.2
DeepSeek R1 8B	3.8	3.8	2.5	-1.2

Table 7: Model performance on the real-world SCICOQA data for the 80 instances where both a GPT-5 and Gemini 3.1 Pro ground truth discrepancy description is available. The columns show the recall performance when assessed against both (Gemini+GPT), Gemini-only, or GPT-only ground truth. A negative Δ shows a preference for Gemini’s ground truth descriptions, while a positive Δ shows a preference for GPT’s.

be influenced by the general capabilities of the models, the strong correlation confirms that the synthetic data preserves relative model rankings. Consequently, absolute performance should be assessed on the real subset, while the synthetic subset serves as a reliable proxy for relative model comparison, justifying its use for scaling the benchmark to scientific domains where real-world examples are scarce.

C.5 Gemini vs. GPT Ground Truth Recall

A potential concern with the LLM-rephrased ground truth discrepancy descriptions is that a model from the same family might have an advantage when evaluated against the ground truth (GT) that was generated by the same model or from the same family. To test this, we compare per-model recall when evaluating against only the GPT-generated versus Gemini-generated ground truth descriptions across the 80 discrepancies for which both are available. We report results in Table 7. We find no evidence of systematic provider bias. In fact, the closed OpenAI models perform the same or better against the Gemini-generated

Model	P	R	F1
GPT-5	84.1	63.6	72.4
GPT-5 Mini	70.7	62.8	66.5
GPT-OSS 20B	68.9	63.6	66.1
Gemini 2.5 Pro	85.7	48.8	62.2
Gemini 3.1 Pro	76.1	50.4	60.6
GPT-OSS 120B	61.1	52.7	56.6
Gemini 2.5 Flash	63.2	41.1	49.8
GPT-5 Codex	79.1	34.9	48.4
Gemini 2.5 Flash Lite	57.1	30.2	39.5
Nemotron Super 49B v1.5	56.1	30.2	39.3
GPT-5 Nano	44.3	34.1	38.5
Devstral 24B Small	45.6	27.9	34.6
Qwen3 30B Inst.	42.6	27.1	33.2
Qwen3 30B Coder	39.1	27.9	32.6
Qwen3 30B Think.	62.5	18.6	28.7
Qwen3 4B Inst.	38.9	22.5	28.5
DeepSeek R1 32B	32.8	22.5	26.7
Magistral 24B Small	28.7	24.0	26.2
Qwen3 4B Think.	61.5	14.7	23.8
Nemotron Nano 9B v2	23.5	17.8	20.3
DeepSeek Coder 16B V2	11.1	6.2	8.0
DeepSeek R1 8B	14.3	1.6	2.8

Table 8: Precision (P), Recall (R), and F1 scores on 20 NLP and CV papers with 129 discrepancies. Ground truths have been pooled from the predictions of GPT-5, Gemini 2.5 Pro, and GPT-OSS 20B. Additionally, discrepancies from the real-world data of SCICOQA have been added.

ground truth. Gemini 3.1 Pro performs slightly better against its own generations ($\Delta = -1.3$), but the other Gemini models do not exhibit any bias towards these descriptions. Among third-party model families, which serve as unbiased controls since they are not affiliated with either GT provider, results are similarly balanced: DeepSeek and NVIDIA models show no consistent preference, while the Qwen models slightly favor GPT GT (mean $\Delta = +2.0$ pp). Across all 22 models, the overall mean Δ is $+0.2$ pp with a maximum absolute difference of just 3.7 pp, and 9 models favor GPT GT versus 8 favoring Gemini GT (5 tied). These small, inconsistent differences suggest that the choice of ground truth source does not introduce meaningful evaluation bias.

C.6 Pooled Ground Truth Results

In §5.2 we created a pooled ground truth for 20 NLP and CV papers by manually verifying the pre-

dicted discrepancies from GPT-5, Gemini 2.5 Pro, and GPT-OSS 20B. Together with the existing discrepancies from the real-world data in SCICOQA on these papers, this yields 129 verified paper-code discrepancies. Following the pooling methodology from information retrieval (Buckley and Voorhees, 2004), where relevance judgments are aggregated from multiple systems to approximate a more complete ground truth, we evaluate all models against the pooled set. In case multiple ground truths exist for the same discrepancy, e.g., when both GPT-5 and Gemini 2.5 Pro surfaced the same unannotated discrepancy, we consider it detected if the prediction matches at least one description. Table 8 reports the results.

GPT-5 achieves the highest F1 (72.4%), balancing precision (84.1%) and recall (63.6%) the best, followed by GPT-5 Mini and GPT-OSS 20B (66.5% and 66.1%). As established in the pooling literature, systems contributing to the pool have a structural advantage, so the evaluations of GPT-5, Gemini 2.5 Pro, and GPT-OSS 20B should be viewed as upper bounds (Buckley and Voorhees, 2004). Nonetheless, clear patterns emerge: a precision–recall tradeoff is evident across models, with Gemini 2.5 Pro reaching the highest precision (85.7%) at the cost of recall (48.8%). GPT-OSS 20B and 120B are the strongest open-weight models, while the remaining open-weight models lag behind considerably (next best: Nemotron 49B, F1 of 39.3%). Despite the pool bias favoring contributing systems, recall remains the primary bottleneck: even under favorable conditions, the best model detects at most 63.6% of verified discrepancies.

Model	Bin # Samples	All 635	0-32k 253	32-64k 210	64-131k 105	$\geq 131k$ 67
GPT-5		65.8	77.5	62.9	53.3	50.7
GPT-5 Mini		61.7	75.1	59.5	44.8	44.8
Gemini 3.1 Pro		55.0	64.4	55.7	41.9	37.3
GPT-OSS 20B		46.5	60.1	41.9	35.2	26.9
Gemini 2.5 Pro		47.1	58.1	43.8	37.1	31.3
GPT-5 Codex		45.5	57.7	41.9	36.2	25.4
GPT-OSS 120B		43.6	59.3	38.6	28.6	23.9
Gemini 2.5 Flash		40.5	51.4	40.5	24.8	23.9
GPT-5 Nano		26.3	34.8	22.4	21.9	13.4
Nemotron Super 49B v1.5		23.9	29.6	20.5	20.0	19.4
Qwen3 30B Coder		23.5	27.3	20.0	24.8	17.9
Gemini 2.5 Flash Lite		22.8	33.6	18.1	14.3	10.4
Nemotron Nano 9B v2		16.2	21.7	13.8	14.3	6.0
Qwen3 30B Inst.		21.3	26.5	20.5	12.4	17.9
DeepSeek R1 32B		16.9	25.7	16.7	4.8	3.0
DeepSeek Coder 16B V2		8.7	13.8	4.3	5.7	7.5
Qwen3 4B Inst.		15.1	21.7	12.9	8.6	7.5
Magistral 24B Small		14.5	20.9	14.3	6.7	3.0
Devstral 24B Small		13.7	19.0	11.9	5.7	11.9
Qwen3 30B Think.		15.0	19.4	14.3	7.6	11.9
Qwen3 4B Think.		10.6	15.4	11.4	3.8	0.0
DeepSeek R1 8B		4.9	5.5	4.8	2.9	6.0

Table 9: Model performance by position of the last relevant code file in the prompt. The *All* column considers the full SCiCOQA dataset, the following columns partition the data based on the position of the last relevant code file in the prompt.

C.7 Detailed Results per Model

Model	Real+Synthetic		Real			Synthetic			Type		
	# Preds.	Recall	All	GH	RP	All	CS	Other	CO	PO	Diff.
# Real # Synthetic	92 543	92 543	92 10	42 10	50 10	0 543	0 286	0 257	12 55	29 64	51 424
GPT-5	4.6	65.8	41.3	54.8	30.0	70.0	65.7	74.7	70.1	51.6	68.0
GPT-5 Mini	5.3	61.7	46.7	59.5	36.0	64.3	58.7	70.4	70.1	45.2	63.8
Gemini 3.1 Pro	3.8	55.0	46.7	59.5	36.0	56.4	54.9	58.0	62.7	41.9	56.4
GPT-OSS 20B	5.8	46.5	42.4	47.6	38.0	47.1	41.6	53.3	61.2	40.9	45.5
Gemini 2.5 Pro	3.5	47.1	39.1	54.8	26.0	48.4	43.0	54.5	46.3	34.4	49.7
GPT-5 Codex	2.2	45.5	27.2	31.0	24.0	48.6	45.5	52.1	59.7	28.0	46.9
GPT-OSS 120B	5.0	43.6	41.3	50.0	34.0	44.0	40.6	47.9	52.2	40.9	42.9
Gemini 2.5 Flash	4.0	40.5	34.8	47.6	24.0	41.4	36.0	47.5	47.8	33.3	40.8
GPT-5 Nano	3.8	26.3	19.6	26.2	14.0	27.4	22.7	32.7	40.3	17.2	26.1
Nemotron Super 49B v1.5	5.1	23.9	23.9	28.6	20.0	23.9	23.8	24.1	37.3	15.1	23.8
Qwen3 30B Coder	4.3	23.5	21.7	26.2	18.0	23.8	23.4	24.1	38.8	11.8	23.6
Gemini 2.5 Flash Lite	2.8	22.8	17.4	31.0	6.0	23.8	19.2	28.8	32.8	11.8	23.6
Nemotron Nano 9B v2	6.7	16.2	13.0	14.3	12.0	16.8	15.0	18.7	13.4	16.1	16.6
Qwen3 30B Inst.	3.8	21.3	18.5	21.4	16.0	21.7	17.8	26.1	32.8	15.1	20.8
DeepSeek R1 32B	4.7	16.9	19.6	31.0	10.0	16.4	14.3	18.7	23.9	14.0	16.4
DeepSeek Coder 16B V2	3.5	8.7	3.3	2.4	4.0	9.6	9.1	10.1	19.4	3.2	8.2
Qwen3 4B Inst.	2.8	15.1	9.8	14.3	6.0	16.0	12.9	19.5	22.4	11.8	14.7
Magistral 24B Small	4.3	14.5	16.3	21.4	12.0	14.2	13.3	15.2	19.4	14.0	13.9
Devstral 24B Small	4.8	13.7	12.0	16.7	8.0	14.0	11.2	17.1	25.4	15.1	11.8
Qwen3 30B Think.	1.7	15.0	12.0	19.0	6.0	15.5	15.0	16.0	23.9	4.3	15.8
Qwen3 4B Think.	1.6	10.6	7.6	7.1	8.0	11.0	9.1	13.2	11.9	7.5	10.9
DeepSeek R1 8B	4.2	4.9	3.3	7.1	0.0	5.2	6.6	3.5	9.0	1.1	5.1

Table 10: Detailed recall performance of all evaluated models in the SCICoQA dataset. *# Preds.* refers to the average number of predictions the model makes per paper, *GH* stands for discrepancies originating from GitHub, *RP* originating from reproducibility papers. The synthetic data is split by the Computer Science (*CS*) domain and *Others*. The discrepancies are further split by *Type*, specifically Code Omissions (*CO*), Paper Omissions (*PO*), and Differences (*Diff*). The *# Real/Synthetic* row indicates how many samples per column are from the real or synthetic data.

C.8 Code Only Ablation Results

Model	Real			Synthetic			Real+Synthetic		
	P+C	C	Δ	P+C	C	Δ	P+C	C	Δ
GPT-5	41.3	29.3	-12.0	70.0	51.0	-19.0	65.8	47.9	-18.0
GPT-5 Mini	46.7	23.9	-22.8	64.3	39.6	-24.7	61.7	37.3	-24.4
Gemini 3.1 Pro	46.7	25.0	-21.7	56.4	42.0	-14.4	55.0	39.5	-15.4
GPT-OSS 20B	42.4	17.4	-25.0	47.1	33.3	-13.8	46.5	31.0	-15.4
Gemini 2.5 Pro	39.1	18.5	-20.7	48.4	33.1	-15.3	47.1	31.0	-16.1
GPT-5 Codex	27.2	12.0	-15.2	48.6	28.0	-20.6	45.5	25.7	-19.8
GPT-OSS 120B	41.3	22.8	-18.5	44.0	34.6	-9.4	43.6	32.9	-10.7
Gemini 2.5 Flash	34.8	17.4	-17.4	41.4	28.5	-12.9	40.5	26.9	-13.5
Average	39.9	20.8	-19.2	52.5	36.3	-16.3	50.7	34.0	-16.7

Table 11: Discrepancy recall of the top 8 performing models depending on the input: The *P+C* column refers to the performance when providing paper and code, *C* for only providing the code, and Δ shows the absolute difference between the two.

D Example Discrepancies in SCICOQA

Type	Category	Description
Paper Omission	Algorithm	<p><i>Gemini 3.1</i></p> <p>The paper discusses utilizing Feature-wise Linear Modulation (FiLM) to adjust convolutional model features based on the loss weights. For this modulation, it describes using multi-layer perceptrons to compute unconstrained scaling (σ) and shifting (μ) parameter bounds which are applied as affine transformations to the features. In contrast, the code implements a ‘FiLMBlock’ that processes the shift (‘mu’) and scale (‘sigma’) through a sigmoid activation function prior to applying them to the features. The difference is the undocumented use of an output sigmoid layer on the FiLM parameters in the implementation which restricts the scaling factor to solely shrink the features and the shift translation strictly to positive values compared to the unrestricted parameter bounds discussed in the text.</p>
		<p><i>GPT-5</i></p> <p>The paper conditions layers using FiLM by mapping the loss-parameter vector λ through two MLPs to obtain σ and μ and then applying the affine transform $\hat{f} = \sigma \odot f + \mu$; there is no mention of any output activation on σ or μ, and in the style transfer details it explicitly states that σ and μ are computed via affine maps. In the repository, the FiLMBlock applies a sigmoid activation to both μ and σ before using them, thereby constraining both to the (0,1) range. Thus, the code restricts the FiLM coefficients to be positive and bounded, whereas the paper’s description implies unconstrained affine coefficients produced by linear layers. This is a clear deviation between the implementation and the method described in the paper.</p>
Paper Omission	Model	<p><i>Gemini 3.1</i></p> <p>The paper states that a three-layer GNN is trained for the node-classification tasks before generating explanations, but omits any description of batch normalization layers being used as part of the architecture. However, the code implementation for node classification models includes batch normalization by default (‘args.bn = True’). Specifically, in the GCN model code, batch normalization is applied after the first and second GCN layers during the forward pass. This introduces an undocumented architectural choice present in the code but not mentioned in the paper.</p>
		<p><i>GPT-5</i></p> <p>The paper states that, following GNNExplainer’s setup, a three-layer GNN is first trained and then explained, but it does not mention any batch normalization in this backbone network. In the public code for node-classification tasks (synthetic datasets), the GCN model enables batch normalization by default (args.bn=True) and applies tf.keras.layers.BatchNormalization after all GCN layers except the last, i.e., after the first and second layers in a 3-layer GCN. This BatchNorm usage is not described in the paper’s model description or experimental setup, making it an undocumented architectural difference between the paper and the released code.</p>
Code Omission	Data	<p><i>Gemini 3.1</i></p> <p>The paper states that for the MNIST experiments, the "Original + CGN" baseline is trained on a combination of real and counterfactual images. However, the implemented code for training the MNIST classifiers does not combine these datasets. The data loading logic defined in ‘get_tensor_dataloaders’ exclusively loads the counterfactual tensor dataset when a counterfactual dataset name is evaluated, without appending or mixing in the original training dataset. Therefore, the code implementation trains solely on counterfactual data, lacking the original data indicated by the paper.</p>
		<p><i>GPT-5</i></p> <p>The paper describes training classifiers using a combination of original MNIST data and CGN-generated counterfactual images. This is explicitly stated in Section 3.2 and reinforced by the MNIST ablation (Appendix A.3) and Table 2 which presents an “Original + CGN” setting. In contrast, the released MNIST training code (mnists/train_classifier.py) only supports training on a single dataset per run; the provided datasets and README usage instruct training on the counterfactual dataset alone (e.g., wildlife_MNIST_counterfactual) without combining with original data. Thus, the code implements CF-only training for MNIST classifiers rather than the paper’s combined original+CF setup, constituting a mismatch in the MNIST training protocol.</p>

Table 12: Discrepancy ground truth examples from the SCICOQA dataset generated by both Gemini 3.1 Pro and GPT-5, including discrepancy type and category annotations. All samples are from real data.

Type	Category	Description
Code Omission	Training	<p><i>Gemini 3.1</i></p> <p>The paper algorithm outlines that within each global aggregation round, each worker performs τ local update steps on different minibatches and accumulates the computed gradients to represent the overall worker gradient. However, in the code, the ‘federated_worker_process’ loops over the dataloader, performs a single forward and backward pass, and then immediately unconditionally breaks out of the loop. Additionally, the lines of code previously intended for accumulating gradients (‘accum_grads = add_param_list(accum_grads, worker_grads’) are commented out, which means the code effectively enforces $\tau = 1$.</p> <p><i>GPT-5</i></p> <p>The paper’s FL formulation and Algorithm 1 state that each worker performs τ local updates per round and accumulates the resulting gradients $\mathbf{g}_k(t,b)$ into an accumulated stochastic gradient $\mathbf{g}_k(t) = \sum_{b=0}^{\tau-1} \mathbf{g}_k(\theta_k(t,b))$, which is then used by LBGm for projection and communication. In the code, the federated worker loop processes only one minibatch and immediately breaks, with the lines that would accumulate gradients across batches commented out; an assertion on τ is present but no actual multi-step accumulation is performed. Consequently, the implemented training uses only a single local update per round (effectively $\tau=1$) rather than accumulating τ local updates as described in the paper.</p>
Difference	Eval	<p><i>Gemini 3.1</i></p> <p>The paper describes the accuracy-based confidence score as the mean accuracy of the candidate responses, computed by taking the average of the indicator functions that check if each sampled response matches the ground-truth answer. In the provided code, the confidence score calculation is implemented by returning the maximum value of the sampled correctness scores (‘max(probabilities)’) instead of their average. This creates a discrepancy where the code leverages a binary indicator reflecting whether at least one correct response was generated (1 if yes, 0 otherwise), rather than the continuous proportion of correct responses specified by the formula in the paper.</p> <p><i>GPT-5</i></p> <p>The paper specifies “accuracy-based confidence” for a question as the mean accuracy across K sampled responses, i.e., the fraction of correct generations among the K samples. In the code, however, the function calculate_confidence takes the list of per-sample correctness indicators (greedy_scores) and returns max(greedy_scores), which reduces confidence to a binary value: 1 if any sample is correct, 0 otherwise. Moreover, while the code also computes greedy_scores_avg (the intended mean), it does not use that for the confidence fed into the downstream datasets; instead, the binary max value is stored as “confidence” and used for reward/align data. Thus, the implemented confidence deviates from the paper’s definition by using maximization rather than averaging.</p>
Difference	Loss	<p><i>Gemini 3.1</i></p> <p>The paper defines the P_2 penalty term as the sum of $\min(\xi(x,y), 0)^2$, which applies a squared penalty strictly to negative evaluations of $\xi(x,y)$ in order to enforce the provided optimal transport constraints. However, the code implements this term using ‘torch.clamp(full_xi_vals, min=0)**2’. The PyTorch function with ‘min=0’ imposes a lower bound of 0, mathematically returning $\max(\xi(x,y), 0)^2$, which zeroes out the negative values and squares the positive ones. Consequently, the codebase incorrectly penalizes positive values of $\xi(x,y)$ instead of negative values as stated in the methodology.</p> <p><i>GPT-5</i></p> <p>The paper defines the second penalty P_2 to enforce the admissibility constraints $(\phi, \phi^c) \in \text{ADM}(c)$ by penalizing only negative values of $\xi(x,y) = c(x,y) - \phi(x) - \phi^c(y)$, specifically using $\min(\xi(x,y), 0)^2$. In the code, P_2 is implemented in admissible_penalty as torch.mean(torch.clamp(full_xi_vals, min=0)**2), which penalizes positive ξ values (since negative values are clamped to zero). Therefore, the implementation reverses the intended sign: the paper penalizes violations where $\xi < 0$, while the code penalizes cases where $\xi > 0$. This directly contradicts the paper’s definition for P_2.</p>

Table 12: Discrepancy ground truth examples from the SCICOQA dataset generated by both Gemini 3.1 Pro and GPT-5, including discrepancy type and category annotations. All samples are from real data. (Continued)

E Example Synthetic Discrepancies in SCICoQA

Discrepancy	Code Change
<p><i>Paper:</i> https://arxiv.org/abs/1906.09436</p> <p><i>Domain:</i> Machine Learning (stat.ML)</p> <p>The paper emphasizes that classification in FDA is performed using Euclidean distances in the Fisher subspace (and shows its equivalence to LDA under Euclidean metrics). The code normalizes both projected samples and class centroids to unit norm prior to classification. This switches the effective decision metric from Euclidean distance to an angular/cosine-like similarity, diverging from the Euclidean geometry prescribed in the paper.</p>	<p><i>Code:</i> https://github.com/concavegit/kfda</p> <p><i>File:</i> kfda/kfda.py</p> <pre>centroids_ = m_classes @ self.weights_ + centroids_ = centroids_ / (np.linalg.norm(centroids_, axis=1, keepdims=True) + 1e-12) self.clf_ = NearestCentroid().fit(centroids_, self .classes_) # ... - return pairwise_kernels(X, self.X_, metric=self. kernel, **self.kwds) @ self.weights_ + Z = pairwise_kernels(X, self.X_, metric=self. kernel, **self.kwds) @ self.weights_ + Z_norm = np.linalg.norm(Z, axis=1, keepdims=True) + 1e-12 + return Z / Z_norm # ... new_centroids = y_onehot.T @ projections / y_onehot.T.sum(1) + new_centroids = new_centroids / (np.linalg.norm(new_centroids, axis=1, keepdims=True) + 1e-12) concatenated_classes = np.concatenate([self. classes_, new_classes]) concatenated_centroids = np.concatenate([self.clf_ .centroids_, new_centroids])</pre>
<p><i>Paper:</i> https://arxiv.org/abs/2409.13224</p> <p><i>Domain:</i> General Relativity and Quantum Cosmology (gr-qc)</p> <p>The paper defines squared coherence as the magnitude-squared of the complex cross-spectrum normalized by the product of the auto-spectra. In the code, coherence is computed using only the square of the real part of the cross-spectrum (co-spectrum) divided by the product of the auto-spectra. Thus, while the paper's definition uses $S_{xy} ^2$, the implementation effectively uses $Re(S_{xy})^2$, ignoring the quadrature (imaginary) component. This changes the coherence measure away from magnitude-squared coherence to co-coherence, which is not described in the paper.</p>	<p><i>Code:</i> https://github.com/nz-gravity/sgvb_psd</p> <p><i>File:</i> src/sgvb_psd/postproc/plot_coherence.py</p> <pre>def compute_coherence(pxx, pyy, pxy): - coh_q = np.real(np.abs(pxy) ** 2) / np.real(pxx) / np.real(pyy) + coh_q = (np.real(pxy) ** 2) / (np.real(pxx) * np.real(pyy)) coh_uniform = None</pre> <p><i>File:</i> src/sgvb_psd/postproc/plot_triangle_psd_coherence.py</p> <pre>def compute_coherence(pxx, pyy, pxy): - coh_q = np.real(np.abs(pxy) ** 2) / np.real(pxx) / np.real(pyy) + coh_q = (np.real(pxy) ** 2) / (np.real(pxx) * np.real(pyy)) coh_uniform = None</pre>
<p><i>Paper:</i> https://arxiv.org/abs/1911.10194</p> <p><i>Domain:</i> Computer Vision (cs.CV)</p> <p>The paper defines the offset vector in pixel space and groups pixels by moving each pixel by its predicted offset and assigning it to the nearest predicted center. The code upscales predictions to the input resolution but does not scale the offsets accordingly during this upsampling, meaning the offsets are left in a lower-resolution coordinate system when grouped at full resolution. This contradicts the paper's grouping formulation that assumes offsets and coordinates are in the same resolution.</p>	<p><i>Code:</i> https://github.com/bowenc0221/panoptic-deeplab</p> <p><i>File:</i> segmentation/model/meta_arch/panoptic_deeplab.py</p> <pre>result = OrderedDict() for key in pred.keys(): out = F.interpolate(pred[key], size= input_shape, mode='bilinear', align_corners= True) - if 'offset' in key: - scale = (input_shape[0] - 1) // (pred[key] .shape[2] - 1) - out *= scale result[key] = out return result</pre>

Table 13: Examples from the synthetic SCICoQA data. The *Code Change* column shows the code snippet that has been modified to create a paper-code discrepancy.

Discrepancy	Code Change
<p><i>Paper:</i> https://arxiv.org/abs/2302.12835</p> <p><i>Domain:</i> Image and Video Processing (eess.IV)</p> <p>The paper specifies that training uses only a data fidelity term plus enforcement of the no-slip condition on the wall, explicitly noting no physics residuals. In the code, an additional divergence-free penalty term is added by computing spatial derivatives of the velocity outputs with respect to the input coordinates and penalizing their sum (divergence), scaled by a configurable weight. This introduces an explicit incompressibility prior not described in the paper.</p>	<p><i>Code:</i> https://github.com/saitta-s/INRs-4DFlowMRI</p> <p><i>File:</i> fit_inr.py</p> <pre>def closure(): optimizer.zero_grad() + X.requires_grad_(True) outputs = model(X) obs_loss = loss_fn(outputs[:-n_wall, :], Y0) wall_loss = loss_fn(outputs[-n_wall:, :], wall_vel[wall_idx, :]) + if cfg.div_weight > 0: + u = outputs[:, 0] + v = outputs[:, 1] + w = outputs[:, 2] + grads_u = torch.autograd.grad(u.sum(), X, create_graph=True, retain_graph=True)[0] + grads_v = torch.autograd.grad(v.sum(), X, create_graph=True, retain_graph=True)[0] + grads_w = torch.autograd.grad(w.sum(), X, create_graph=True, retain_graph=True)[0] + div = grads_u[:, 0] + grads_v[:, 1] + grads_w[:, 2] + div_loss = (div ** 2).mean() + loss = obs_loss + wall_loss + cfg. div_weight * div_loss + else: + loss = obs_loss + wall_loss loss.backward() return loss</pre>
<p><i>Paper:</i> https://arxiv.org/abs/2412.04595</p> <p><i>Domain:</i> Numerical Analysis (math.NA)</p> <p>The paper derives a single Fourier-Chebyshev representation for the long-range part and does not duplicate the zero wavenumber contribution. In the code, the zero in-plane Fourier mode is already computed via a dedicated “zeroth-order” 1D path, but the spectral path is also modified to include the $k_x = k_y = 0$ contribution. This leads to a double counting of the zeroth Fourier mode in the long-range part, which the paper’s formulation does not include.</p>	<p><i>Code:</i> https://github.com/HPMolSim/FastSpecSoG.jl</p> <p><i>File:</i> src/FFCT/interpolate.jl</p> <pre>for j in 1:size(H_r, 2) k_yj = k_y[j] for i in 1:size(H_r, 1) k_xi = k_x[i] k2 = k_xi^2 + k_yj^2 - if !(k2 ≈ zero(T)) - phase = phase_x[i] * phase_y[j] - cheb_ij = cheb_mat[i, j] - for k in 1:size(H_r, 3) - r_zk = r_z[k] - H_r[i, j, k] += q * phase * cheb_ij(abs(z - r_zk)) - end + phase = phase_x[i] * phase_y[j] + cheb_ij = cheb_mat[i, j] + for k in 1:size(H_r, 3) + r_zk = r_z[k] + H_r[i, j, k] += q * phase * cheb_ij(abs(z - r_zk))</pre>
<p><i>Paper:</i> https://arxiv.org/abs/2402.18396</p> <p><i>Domain:</i> Biomolecules (q-bio.BM)</p> <p>The paper describes that the confidence model reasons about local interactions by feeding only residues whose $C\alpha$ lies within 20 Å of any predicted ligand atom. The code instead crops the receptor using a centroid-based distance (distance from $C\alpha$ to the ligand centroid), not the minimum distance to any ligand atom. This changes which residues are kept and contradicts the paper’s “any atom” criterion.</p>	<p><i>Code:</i> https://github.com/LDeng0205/confidence-bootstrapping</p> <p><i>File:</i> utils/utils.py</p> <pre>def crop_beyond(complex_graph, cutoff, all_atoms): ligand_pos = complex_graph['ligand'].pos + center = ligand_pos.mean(dim=0, keepdim=True) receptor_pos = complex_graph['receptor'].pos - residues_to_keep = torch.any(torch.sum((ligand_pos.unsqueeze(0) - receptor_pos. unsqueeze(1)) ** 2, -1) < cutoff ** 2, dim=1) + residues_to_keep = torch.sum((receptor_pos - center) ** 2, -1) < cutoff ** 2</pre>

Table 13: Examples from the synthetic SCICOQA data. The *Code Change* column shows the code snippet that has been modified to create a paper-code discrepancy. (Continued)

F Prompts

F.1 GitHub Issue Discrepancy Extraction

PROMPT | GitHub Issue Discrepancy Extraction

You are an assistant that analyzes GitHub issues of scientific codebases. Your primary goal is to determine if the code repository contains any inconsistency, discrepancy, or mismatch between what is described in the paper and implemented in the code. For this, you analyze a GitHub issue and determine whether it reports a concrete discrepancy.

A **concrete discrepancy** means the issue clearly describes a mismatch between what is stated in the paper (e.g., formulas, algorithms, hyperparameters, methods, logic, or processes) and what is implemented in the repository's code.

Important: Only label issues as a discrepancy if they point to a specific, concrete difference between paper and code. Do not label general reproducibility problems, missing details, or unrelated bugs as discrepancies.

Not a discrepancy issue

Label as **Not a discrepancy issue** if:

- The issue is about anything other than a difference between the paper and the code.
- The issue describes reproducibility problems (e.g., different results) but does **not** identify a concrete paper-code mismatch.
- The issue is about missing information needed to reproduce results without pointing to a mismatch.
- The issue is about bugs or errors unrelated to the paper's described methods or experiments.

Discrepancy issue

Label as **Discrepancy issue** if:

- The issue explicitly reports a mismatch between the paper and the code implementation.
- The mismatch can involve hyperparameters, formulas, algorithms, logic, processes, or other settings described in the paper.

Response Format

After analyzing the issue in detail and applying the definitions above, provide your final classification in the structure defined below:

Final Answer

```yaml

issue\_label: <the issue label: "Not a discrepancy issue" or "Discrepancy issue">

```

Issue

{issue}

F.2 GitHub Issue Discrepancy Verification

PROMPT | GitHub Issue Discrepancy Verification

Your task is to verify a paper-code discrepancy described in a GitHub issue. Your goal is to verify whether the discrepancy is valid or not.

Follow these steps to verify the discrepancy:

1. Analyze the issue and ensure you understand exactly the claimed discrepancy between the paper and the code.
2. Analyze the paper and the code and understand in detail the relevant paper sections and code files.
3. Using your understanding of paper, code, and the discrepancy in the issue, analyze whether the discrepancy is valid or not. A discrepancy is valid, if the claimed discrepancy actually exists and there is a difference between the paper description and the code.
4. Provide your final judgement whether the reported discrepancy is valid or not, and if so a summary and the relevant paper sections and code files in the format below:

```yaml

is\_valid\_discrepancy: <yes or no>

is\_valid\_discrepancy\_reason: <provide a short explanation for your judgement>

discrepancy\_summary: <if valid provide the following description, else this should be empty: a summary of the discrepancy between the paper and the code in 3-8 sentences. Your description should contain three parts focusing on the discrepancy: 1) summarize what is described in the paper, 2) summarize what is implemented in the code, and 3) summarize the difference. Do not speculate about the impact.>

relevant\_paper\_sections:

- <verbatim quote any parts from the paper that are relevant to the discrepancy>

- <if there are multiple relevant parts, paste each of them.>

relevant\_code\_files:

- <name any code files that are relevant to the discrepancy by providing the file name>

- <if there are multiple relevant code files, paste each of them.>

```

Issue

{issue}

Paper

{paper}

Code

{code}

F.3 Reproducibility Paper Discrepancy Extraction

PROMPT | Reproducibility Paper Discrepancy Extraction

You are an assistant that analyzes reproducibility reports of scientific papers. Your primary goal is to detect whether the report identifies any discrepancies between the original paper and the original code repository.

What counts as a discrepancy - A concrete paper-code discrepancy means the report clearly describes a mismatch between what is stated in the original paper (e.g., formulas, algorithms, logic, methods, processes, or other settings) and what is implemented in the original code repository. - Each distinct mismatch should be reported as a separate item. - Hyperparameter mismatches (e.g., learning rate, batch size, dropout rate) do not count as discrepancies, since these are typically configurable in code repositories. - If the report only describes differences between reproduced results and original results, without identifying a paper-code mismatch, do not list it. - If the report speculates about a possible mismatch (uncertain or ambiguous wording), still list it, but mark it with confidence: low.

What does not count as a discrepancy - General reproducibility problems (e.g., "we could not match the reported results").

- Missing information in the paper (e.g., "the authors did not specify X").
- Missing implementation in the original code repository (e.g., "the authors did not provide the code for X").
- Bugs or errors in the code that are unrelated to what the paper describes.
- Differences between reproduced implementation/results and the original paper.

Output format

Summarize all discrepancies found in the following structure, providing a description of the discrepancy, evidence supporting the discrepancy as verbatim quotes from the reproducibility report, and a confidence estimate of the authors on the reported discrepancy.

```
“yaml
concrete_paper_code_discrepancies:
- description: "<3-8 sentence descriptive summary of the first discrepancy>"
  evidence:
  - "<paste any evidence (e.g. a paragraph describing the discrepancy) from the reproducibility report that support the discrepancy.>"
  - "<If there are multiple evidence, paste each of them.>"
  confidence: <Estimate of the confidence the authors have on the reported discrepancy. One of: low, medium, high>- ... “
```

If no discrepancies are reported, return:

```
“yaml
concrete_paper_code_discrepancies: []
...”

Reproducibility Report:
{paper}
```

F.4 Reproducibility Paper Discrepancy Verification

PROMPT | Reproducibility Paper Discrepancy Verification

Your task is to verify a paper-code discrepancy described in a reproducibility report. You will be provided with a summary of the discrepancy according to the report, and potentially multiple quotes from the reproducibility report that support the discrepancy. Your goal is to verify whether the discrepancy is valid or not.

Follow these steps to verify the discrepancy:

1. Analyze the description and evidence from the reproducibility report and ensure you understand exactly the claimed discrepancy between the paper and the code.
2. Analyze the paper and the code and understand in detail the relevant paper sections and code files.
3. Using your understanding of paper, code, and the reported discrepancy, analyze whether the discrepancy is valid or not. A discrepancy is valid, if the claimed discrepancy actually exists, i.e. the described difference between paper and code exists.
4. Provide your final judgement whether the reported discrepancy is valid or not, and if so a summary and the relevant paper sections and code files in the format below:

```
“yaml
is_valid_discrepancy: <yes or no>
is_valid_discrepancy_reason: <provide a short explanation for your judgement>
discrepancy_summary: <if valid provide the following description, else this should be empty: a summary of the discrepancy between the paper and the code in 3-8 sentences. Your description should contain three parts focusing on the discrepancy: 1) summarize what is described in the paper, 2) summarize what is implemented in the code, and 3) summarize the difference. Do not speculate about the impact.>
relevant_paper_sections:
- <verbatim quote any parts from the paper that are relevant to the discrepancy>
- <if there are multiple relevant parts, paste each of them.>
relevant_code_files:
- <name any code files that are relevant to the discrepancy by providing the file name>
- <if there are multiple relevant code files, paste each of them.>
...”

## Reproducibility Report with Discrepancy
{discrepancy_in_report}
## Paper
{paper}
## Code
{code}
```

F.5 Synthetic Discrepancy Generation

Below the prompt for generating synthetic discrepancies for computer science papers. For non-CS paper, we remove the discrepancy type definition and prediction from the prompt.

PROMPT | Synthetic Discrepancy Generation

Your task is to generate 5 realistic paper–code discrepancies by introducing small, conceptually meaningful modifications to the codebase of the provided research paper.

Follow these steps to perform the edit:

1. Carefully read and understand both the research paper and the entire code repository provided below. Your goal is to identify the key ideas, methods, and components described in the paper and how they correspond to the implementation in the code.

2. Your changes must adhere to the following constraints:

- **Small**: The changes must affect a few lines of code or a short function. It may affect multiple files, but only to the extent necessary to create a coherent and realistic discrepancy.

- **Relevance**: The changes must relate directly to a core scientific or algorithmic idea of the paper and would likely have an impact on reproducibility of results and validity of claims.

- **Significance**: The changes must introduce a conceptual discrepancy — not a simple hyperparameter, or formatting change, or a change that could be fixed via a simple (command line) argument.

- **Scope**: The changes should not all rely on already implemented features of the code base, but also implement new features or introduce modifications. Balance between relying on existing features and implementing new features.

- **No Comments**: Do not add comments to the changed code which would easily allow to identify the discrepancy.

- **No Bugs**: The introduced discrepancies should not be bugs, i.e., code that could be detected as erroneous by inspecting the code itself. The discrepancy must be related to both the paper and the code to be identified.

- **No Implications**: The discrepancies must not rely on anything the paper only implies or assumes, but the paper must clearly conflict with the code after the change, or omit an important concept that is implemented in the changed code.

3. Your changes can be one of the following types. You can create multiple discrepancies of the same or different types.

- **Paper Omission**: Modify the code such that it implements a concept or idea that is not described in the paper.

- **Code Omission**: Modify the code such that it drops a specific concept or idea that is described in the paper.

- **Difference**: Modify the code such that there is a difference between the paper and the code, i.e., the paper describes one thing and the code implements another, e.g., by changing the order of steps, operations, or core logic.

4. Decide which paper-code discrepancies are most appropriate for the given paper, choosing from the following categories. Note, you can create multiple discrepancies of the same or different types.

- **Loss**: changes to loss definition or terms

- **Algorithm**: changes in the order of steps, operations, or core logic

- **Training**: changes to the learning process, schedule, or optimization

- **Evaluation**: changes to evaluation logic, metrics, or scripts

- **Model**: architectural or initialization changes

- **Data**: dataset usage, preprocessing, augmentation, or filtering

- **Other**: other types of discrepancies that are not covered by the above categories but appropriate for the given paper.

5. Generate 5 discrepancies in the following strict format:

```
"md
```

```
# Discrepancy 1
```

```
- Type: <choose one from: Paper Omission, Code Omission, Difference>
```

```
- Category: <choose one from: Loss, Algorithm, Training, Evaluation, Model, Data, Other>
```

```
- Description: <a summary of the discrepancy between the paper and the code in 3-8 sentences. When referring to the code, do not mention 'original' or 'modified', but assume the code is published with the modification and discrepancy. Your description should contain three parts focusing on the discrepancy: 1) summarize what is described in the paper, 2) summarize what is implemented in the modified code, and 3) summarize the difference. Do not speculate about the impact.>
```

```
## Code Changes
```

```
««««ORIGINAL CODE: <relative/path/to/file.py>
```

```
(paste the relevant lines of the original code exactly as they appear)
```

```
=====
```

```
(paste the lines of the modified code containing your change that replace the original code)
```

```
»»»»DISCREPANCY
```

```
If multiple files are affected, repeat the diff block for each.
```

```
## Relevant Paper Sections
```

```
- <verbatim quote any parts from the paper that are relevant to the discrepancy.>
```

```
- <if there are multiple relevant parts, paste each of them.>
```

```
# Discrepancy 2
```

```
...
```

```
# Discrepancy 3
```

```
...
```

```
...
```

```
## Paper
```

```
{paper}
```

```
## Code
```

```
{code}
```

F.6 Discrepancy Prediction

PROMPT | Discrepancy Prediction

You are an expert in analyzing scientific papers and their code implementations.
Your task is to carefully identify concrete discrepancies between what is described in a paper and what is actually implemented in the code.

What counts as a discrepancy

- A concrete paper-code discrepancy means a mismatch between what is stated in the original paper (e.g., formulas, algorithms, logic, methods, processes, or other settings) and what is implemented in the original code repository.
- Each distinct mismatch should be reported as a separate item.

What does not count as a discrepancy

- Missing information in the paper like hyperparameters (e.g., "the authors did not specify X").
- Hyperparameter mismatches (e.g., learning rate, batch size, dropout rate), since these are typically configurable in code repository.
- Missing implementation in the original code repository (e.g., "the authors did not provide the code for X").
- Bugs or errors in the code that are unrelated to what the paper describes.

Output format

Provide your findings in the following YAML structure:

```
""yaml
discrepancies:
- <a summary of the discrepancy between the paper and the code in 3-8 sentences. Your description should contain three parts focusing on the discrepancy: 1) summarize what is described in the paper, 2) summarize what is implemented in the code, and 3) summarize the difference. Do not speculate about the impact.>
- <if there are multiple discrepancies, put each of them in a separate item.>
""

## Paper
{paper}
## Code
{code}
```

F.7 Discrepancy Evaluation

PROMPT | Discrepancy Evaluation

Your task is to evaluate whether a reference paper - code discrepancies matches a predicted paper - code discrepancy. Follow these steps:

1. Analyze which part of the paper or code each discrepancy is describing. Extract the core claims and issues from the reference and predicted discrepancies.
2. Analyze whether the core claims are about the same issue, i.e. if they describe the same or different paper-code discrepancies. The two discrepancies might use different wording or one might be more detailed than the other. Focus on whether the issue is the same, even if minor details are different. However, if they describe different issues (even about the same topic or part of the paper or code) they do not match.
3. Provide a brief explanation of your reasoning.

Reference Paper-Code Discrepancy
{reference discrepancy}

Predicted Paper-Code Discrepancy
{predicted discrepancy}

Answer Format

Provide your answer in the following format:

```
""yaml
core_claim_reference: <core claim from reference discrepancy >
core_claim_predicted: <core claim from predicted discrepancy >
reasoning: <explanation of why the core claims concern the same issue >
match: <yes | no >
""
```