Understanding Large Language Models' Ability on Interdisciplinary Research

Yuanhao Shen [*1] Daniel Xavier de Sousa [*1 2] Ricardo Marçal [2] Ali Asad [1] Hongyu Guo [3] Xiaodan Zhu [1]

Abstract

This work introduces IDRBench — a pioneering benchmark featuring an expert-annotated dataset and a suite of tasks tailored to evaluate LLMs' capabilities in proposing valuable research ideas for Interdisciplinary Research (IDR). To ensure a reliable evaluation, our dataset consists of scientific publications sourced from the ArXiv platform covering six distinct disciplines and is annotated by domain experts with diverse academic backgrounds. The design of evaluation tasks in IDRBench follows a progressive, real-world perspective, reflecting the natural stages of interdisciplinary research development, including 1) IDR Paper Identification, 2) IDR Idea Integration, and 3) IDR Idea Recommendation. Using IDRBench, we construct baselines across 10 LLMs and observe that despite fostering some level of IDR awareness, LLMs still struggle to produce quality IDR ideas. These findings could not only spark new research directions, but also help to develop next-generation LLMs that excel in interdisciplinary research.

1. Introduction

Aristotle, one of history's greatest polymaths, made enduring contributions across distinct areas, shaping centuries of intellectual development. However, as the body of scientific knowledge has grown considerably, modern science has fragmented into ever more specialized fields, making it increasingly difficult to bridge interdisciplinary scientific domains. At the same time, recent advancements in Large Language Models (LLMs) (et al., 2024b; OpenAI, 2024; et al., 2024a; Anthropic, 2024; et al., 2025), with access to massive corpora spanning diverse scientific fields and improved reasoning abilities, have shown enormous potential in scientific discovery.

While some frameworks (Shi, 2024; Lu et al., 2024) and benchmarks measuring LLM's ideation quality (Guo et al., 2024; Lin et al., 2024; Liu et al., 2025; Yang et al., 2025) have provide solid evaluations on the outcome quality of the proposed ideas in generic research, there is a notable gap in evaluating LLM's ability to perform ideation when the research context is Interdisciplinary Research - IDR. Existing works that explicitly explore the potential of using LLMs to tackle IDR problems (Liu et al., 2024a; Xu et al., 2024; Liu et al., 2024b; Zheng et al., 2024) often conduct investigations that provide different approaches and frameworks, yet lack a thorough evaluation of the inherent LLM capability to perform IDR tasks. As a result, the excessive focus on outcome-driven analyses of LLMs tends to overlook IDR's core characteristic of integration of knowledge across distinct disciplines, leaving scarce evidence to answer this key research question: Are LLMs capable of conducting interdisciplinary research?

Answering this question is not easy. Specifically, we need to evaluate whether an LLM can capture the "Eureka" moment in interdisciplinary research – a flash of sudden enlightenment arising from different disciplines during which a new scientific discovery is made (Knoblich & Oellinger, 2006). However, to the best of our knowledge, none of the existing works have assessed such capacity of LLMs. One major issue that hinders such evaluation is the lack of a dedicated benchmark specifically designed for interdisciplinary research.

To evaluate LLMs' ability in IDR, we introduce IDRBench, a pioneering benchmark for IDR assessment. IDRBench adopts the perspective of a progressive assessment of LLMs conducting IDR by introducing three tasks: (i) IDR Paper Identification (IPI), evaluating LLM's ability to classify interdisciplinary papers; (ii) IDR Idea Integration (I3), testing LLM's ability to combine papers from distinct disciplines into feasible and novel IDR; and (iii) IDR Idea Recommendation (I2R), evaluating LLM's ability to recommend the IDR

[*]Equal contribution [1]Department of Electrical and Computer Engineering & Ingenuity Labs Research Institute, Queen's University, Canada [2]Instituto Federal de Goiás, Brazil [3]National Research Council Canada. Correspondence to: Yuanhao Shen <yuanhao.shen@queensu.ca>.

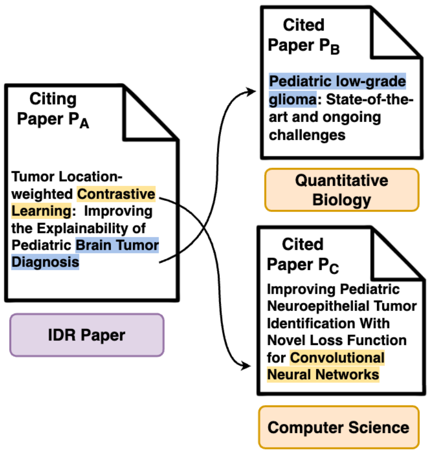

Figure 1: Triplet data format in IDRBench - Showing that Papers P_B and P_C are integrated (more than merely referenced) to generate the IDR Paper P_A.

paper given a list of unrelated papers.

To gauge LLMs' performance on these tasks, the IDR-Bench dataset is structured as a knowledge triplet, Figure 1. For instance, an IDR Paper serves as "Citing Paper P_A", accompanied by two "Cited Papers" P_B and P_C from the disciplines of quantitative biology and computer science, respectively. In this way, an IDR triplet is represented in the form $[P_A; (P_B, P_C)]$.

Our work establishes the first quantified baseline that accurately measures the inherent LLM's capability in conducting IDR. Our results show divergent performance on each of the three IDR tasks, with LLMs describing an advantage over both I3 and I2R tasks, while leaving room for improvement in the IPI task. We summarize our main contributions as follows:

- We release IDRBench, the first ever benchmark with human annotation to comprehensively evaluate LLMs' ability in the IDR setting, including a human-annotated dataset and a suite of dedicated IDR tasks: paper identification (IPI), idea integration (I3), and idea recommendation (I2R).

- Our analysis results reveal divergent performance across LLMs, showcasing their advantages in both I3 and I2R tasks, while leaving notable room for improvement in the IPI task. Further analysis also suggests that LLMs are still struggling to provide high-quality IDR proposals despite showing early signs of IDR awareness.

2. Related Works

Motivated by existing LLM-assisted research ideation frameworks (Radensky et al., 2025; Si et al., 2025), several studies propose novel benchmarks that aim to evaluate the quality of research ideas. For example, (Lin et al., 2024) introduces SchNovel dataset along with a retrieval-augmented framework to assess the novelty of proposed research ideas. (Guo et al., 2024; Liu et al., 2025; Yang et al., 2025; 2024) take the perspective of hypothesis generation to curate large-scale datasets that benchmark the performance of LLMs in formulating novel yet feasible hypotheses. To the best of our knowledge, IDRBench is the first to explicitly provide an evaluation on interdisciplinary research.

3. The IDRBench

We adopt the widely used definition of "interdisciplinary research" (Cantone, 2024; on Key Challenge Areas for Convergence et al., 2014; Nakhoda et al., 2023; Sell et al., 2022), which accounts for the capability to integrate papers from distinct disciplines into a novel research paper whose scope goes beyond existing ones. Following this definition in our annotation user interface, our proposed IDRBench provides a novel benchmark of two key components: (i) a dataset based on paper triplets, and (ii) a suite of three progressive evaluation tasks.

3.1. The IDRBench Dataset

Our dataset in IDRBench employs the knowledge triplet $[P_A; (P_B, P_C)]$ (Figure 1), which was used in our annotation process and is explored in different ways in our proposed tasks (in Section 3.2). To form the triplet, we derive a subset from the ArXiv Platform[1] with over 58,444 papers between November 2024 and January 2025. For positive instances, our annotators selected 360 papers, including the field and subfield following the ArXiv taxonomy, and forming 120 instances of structured triplets $[P_A; (P_B, P_C)]$. For negative instances, we perform a random selection of 9,125 papers from the same ArXiv subset (out of 58,444 papers), with stratified sampling strategies tailored to the goal of each task. The rationale behind the random approach is that interdisciplinary connections are less frequent and more sparse compared to non-IDR papers(Boyack et al., 2005). Further details regarding the annotation process are provided in Appendix A.1.

To ensure a more accurate evaluation of LLMs' IDR capabilities, we curate our data from ArXiv, which

[1] https://arxiv.org/

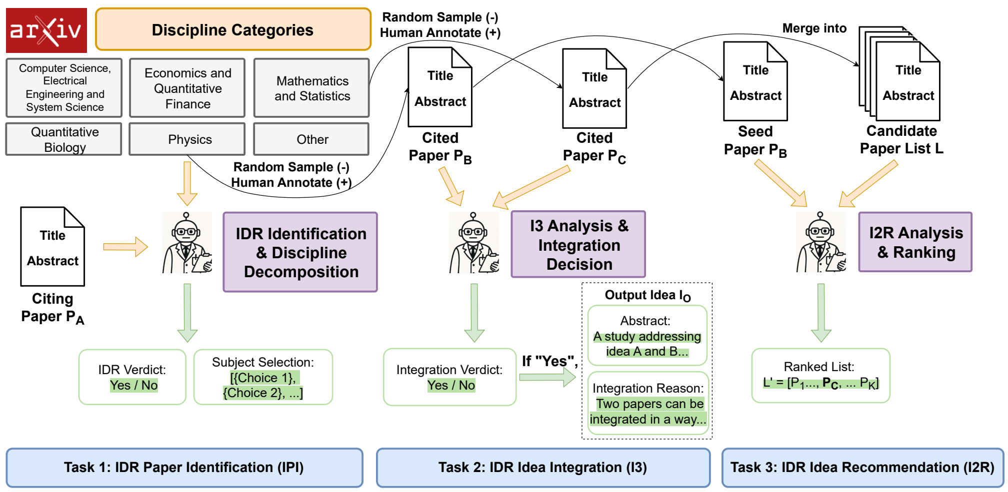

Figure 2: Tasks IPI, I3, and I2R. Orange (green) stands for input (output) flow. Purple shows LLM's functions.

is well-suited for two main reasons: (i) many of the recent papers uploaded have not yet been published in any formal venue, which reduces the risk of information leakage during LLM pretraining (also noted by (Liu et al., 2025)), and (ii) ArXiv requires authors to select specific category taxonomies (referred to as disciplines in our work) when submitting papers. Moreover, we select specific combinations of disciplines that are conceptually more distant from one another, e.g., Computer Science + Quantitative Biology. We also merge some closely related disciplines into single categories to reflect their conceptual proximity, e.g., Computer Science and Electrical Engineering.

3.2. The IDRBench Tasks

IDR Paper Identification (IPI) - Task: IDR Paper Identification (IPI) (first column in Figure 2) measures the ability of LLM to identify whether an academic work is an IDR. Given the title and abstract of a citing paper P_A, the LLM performs a classification task to determine whether P_A is an IDR. When positive, it is further instructed to decompose the IDR idea in P_A into a combination of ArXiv taxonomies (Figure 2).

IPI - Dataset. Negative and positive instances are described as a paper P_A (paper+title), a label (yes or no), and (for positive cases) two disciplines integrated in P_A. The positive instances are derived from human annotation, and negative selected from ArXiv papers with only one defined discipline upon submission by the authors, based on the absence of IDR.

IDR Idea Integration (I3) - Task. IDR Idea Integration (I3), second column in Figure 2, takes a step further to investigate LLMs IDR integration ability. We provide LLMs with two cited papers P_B and P_C–with title + abstract. The LLM is thereafter instructed to perform an idea integration analysis under an IDR setting, predicting whether the integration between P_B and P_C is a promising IDR. The integration follows three criteria: (1) whether the outcome idea aligns with IDR definition, (2) whether the integration yields a feasible IDR, and (3) whether the output IDR idea bears sufficient novelty. We assume that the combination of P_B and P_C is only viable when all three criteria are met. If the answer is yes, we further prompt the model to generate an IDR idea I_o represented by a working paper abstract and the integration reasoning using P_B and P_C.

I3 - Dataset. For positive pairs, we use the cited papers (P_B, P_C) from our annotation process, considering that both were used to be integrated in the IDR P_A paper. For the negative pairs, two papers are randomly selected from two distinct disciplines. This approach is based on the observation that valid interdisciplinary ideas are extremely sparse within the overall space of possible paper combinations.

IDR Idea Recommendation (I2R) - Task: IDR Idea Recommendation (I2R) tests the LLMs' holistic IDR ability in a multiple-step recommendation setting – third column in Figure 2. Given a positive cited paper pair P_B and P_C, the LLM receives P_B as the seed paper, while the other paper P_C is merged into a candidate paper list $L = \{P_1, P_2, ...P_k\}$, that contains unrelated papers. The model is instructed to summarize ideas from seed P_B, search through the ideas in L, and produce a re-ranked paper list L', following criteria

Task	Metric	ChatGPT Mini			Gemini		Llama 3		Claude	DeepSeek	
		4o	o1	o3	1.5-pro	2.0	3.1	3.3	3.7	v3	r1
IPI	F1	0.396	0.412	0.379	0.277	0.337	0.416	0.358	0.347	0.387	0.401
I3	F1	0.667	0.505	0.239	0.509	0.608	0.609	0.723	0.384	0.676	0.582
I2R	MRR	0.682*	0.589	0.640	0.703	0.687*	0.622	0.619	0.679*	0.674*	0.588

Table 1: Classification accuracy of LLMs on IPI and I3 task in IDRBench.

as novelty, feasibility, and IDR. The model is asked to make a comparison to investigate the feasibility of conducting promising IDR when each paper $P_i \in L$ is paired with P_B, considering a pairwise comparison(Si et al., 2025). Better results mean P_C at the top of L' after LLM's re-ranking.

I2R - Dataset: An instance of the dataset has a seed paper P_B, a list L with K negative papers, and one positive paper P_C. The list L contains a set of random negative papers and P_C.

4. Evaluation with IDRBench

IPI Task Performance. The first rows in Table 11 report the F1 scores of various LLMs on the IPI task, considering the dataset in IPI. Compared to a random-guessing baseline F1 score of 0.176, we can see all models achieve higher performance, with llama-3.1-70B model achieving the highest performance of 0.416.

I3 Task Performance. The second row of Table 11 is the I3 performance in F1 scores, with Dataset in I3. Compared to a random guessing of 0.176, there is a higher average F1 score of 0.723.

I2R Task Performance. The third row in Table 11 presents the I2R recommendation results in IDRBench. We apply the Mean Reciprocal Rank (MRR) with Wilcoxon signed-rank test (Hsieh et al., 2008). There is a statistical best result (with asterisks) for ChatGPT, gemini-1.5-pro, Claude, and DeepSeek-v3.

4.1. Detailed Analysis

Are LLMs capable of conducting interdisciplinary research? Our results indicate that there are certain tasks where LLMs show promising performance, such as I3 and I2, but in IPI, LLMs still require improvement. Detailed results can be found in Appendix A.4.

LLMs still struggle to generate quality IDR ideas. Considering the I3 task, we perform two comparisons using the Output Idea I_O in Figure 2: i) Integration reason, comparing LLM output with annotated gold labels (from annotation process), and ii) Abstract, comparing the actual abstract of IDR paper P_A with the abstract generated by LLM. We use semantic similarity (measured by SciBERT Similarity (Beltagy et al., 2019)) and content similarity (measured by BLEU(Papineni et al., 2002)). As result, both comparisons exhibit a strong alignment with the annotated samples in semantic similarity – with a similarity score around 0.75 over the models. However, in terms of content similarity (with BLEU score around 0.011), no model demonstrates a high match with the original text. This divergence suggests that LLMs are focusing more on the semantics of IDR ideation rather than the content itself, indicating substantial potential for improvement in automating the idea generation process in IDR.

LLMs are self-conscious IDR idea recommenders. Given the recommendation results in the I2R task, we also investigate whether the ranking results are influenced by the semantic similarity between cited paper P_B and each candidate in the paper list L, rather than the IDR idea feasibility. Considering the Kendall's τ correlation between the rankings produced by the LLM and those based on contextual similarity (using SciBERT similarity), we note that the correlation ranges from approximately 0.18 to 0.25, indicating that contextual similarity does not strongly influence the LLM's re-ranking decisions, further suggesting that LLMs are capable of handling basic IDR recommendation tasks without reliance on paper similarity.

5. Conclusion

We introduce IDRBench, a novel benchmark to probe LLMs' capacity for interdisciplinary research. IDRBench includes a high quality dataset collected from human annotation. We also design a set of three progressive tasks: (i) IDR Paper Identification (IPI), (ii) IDR Idea Integration (I3), and (iii) IDR Idea Recommendation (I2R) — aiming to answer our research question: Are LLMs capable of conducting interdisciplinary research? Throughout experiment results and further analysis, we uncover that LLMs have already fostered some level of IDR awareness, but they still struggle to produce quality IDR ideas in a plug-and-play fashion. To the best of our knowledge, IDRBench is the first to rigorously investigate LLMs' IDR ability and to provide insights that encourage follow-up studies. We hope IDRBench can become the "Eureka" moment for upcoming studies that inspire the development of more benchmarks to evaluate LLM and IDR.

References

Anthropic. Claude 3.7 sonnet. https://www.anthropic.com/claude/sonnet, 2024. Accessed: 2025-05-15.

Beltagy, I., Lo, K., and Cohan, A. SciBERT: A pretrained language model for scientific text. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371/.

Boyack, K. W., Klavans, R., and Börner, K. Mapping the backbone of science. Scientometrics, 64 (3):351–374, 2005. ISSN 1588-2861. doi: 10.1007/s11192-005-0255-6. URL https://doi.org/10.1007/s11192-005-0255-6.

Cantone, G. G. How to measure interdisciplinary research? a systemic design for the model of measurement. JO - Scientometrics, 2024. doi: 10.1007/s11192-024-05085-1.

et al., A. G. The llama 3 herd of models, 2024a. URL https://arxiv.org/abs/2407.21783.

et al., D. G. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

et al., J. A. Gpt-4 technical report, 2024b. URL https://arxiv.org/abs/2303.08774.

Guo, S., Shariatmadari, A. H., Xiong, G., Huang, A., Xie, E., Bekiranov, S., and Zhang, A. Ideabench: Benchmarking large language models for research idea generation, 2024. URL https://arxiv.org/abs/2411.02429.

Hsieh, J. G., Lin, Y. L., and Jeng, J. H. Preliminary study on Wilcoxon learning machines. Journal of IEEE Transactions on Neural Networks and Learning Systems, 19(2):201–211, 2008.

Knoblich, G. and Oellinger, M. The eureka moment. Scientific American Mind, 17(5):38–43, 2006. ISSN 15552284, 2331379X. URL http://www.jstor.org/stable/24921587.

Lin, E., Peng, Z., and Fang, Y. Evaluating and enhancing large language models for novelty assessment in scholarly publications, 2024. URL https://arxiv.org/abs/2409.16605.

Liu, J., Wang, X., Wang, D., and Pei, L. Aigc-enabled interdisciplinary science measurement. In Sserwanga, I., Joho, H., Ma, J., Hansen, P., Wu, D., Koizumi, M., and Gilliland, A. J. (eds.), Wisdom, Well-Being, Win-Win, pp. 161–170, Cham, 2024a. Springer Nature Switzerland. ISBN 978-3-031-57850-2.

Liu, Y., Sharma, P., Oswal, M. J., Xia, H., and Huang, Y. Personaflow: Boosting research ideation with llm-simulated expert personas. arXiv preprint arXiv:2409.12538, 2024b.

Liu, Y., Yang, Z., Xie, T., Ni, J., Gao, B., Li, Y., Tang, S., Ouyang, W., Cambria, E., and Zhou, D. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition, 2025. URL https://arxiv.org/abs/2503.21248.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The ai scientist: Towards fully automated open-ended scientific discovery, 2024. URL https://arxiv.org/abs/2408.06292.

Nakhoda, M., Whigham, P., and Zwanenburg, S. Quantifying and addressing uncertainty in the measurement of interdisciplinarity. Scientometrics, 128(11): 6107–6127, September 2023. ISSN 0138-9130. doi: 10.1007/s11192-023-04822-2. URL https://doi.org/10.1007/s11192-023-04822-2.

on Key Challenge Areas for Convergence, C., on Life Sciences; Division on Earth, H. B., and Council, L. S. N. R. Convergence: Facilitating Transdisciplinary Integration of Life Sciences. The National Academies Press, Washington, DC, 2014. ISBN 978-0-309-30151-0. doi: 10.17226/18722. URL https://pubmed.ncbi.nlm.nih.gov/24830066/.

OpenAI. Introducing openai o1-preview. https://openai.com/index/introducing-openai-o1-preview/, 2024. Accessed: 2025-05-15.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pp. 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Radensky, M., Shahid, S., Fok, R., Siangliulue, P., Hope, T., and Weld, D. S. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination, 2025. URL https://arxiv.org/abs/2409.14634.

Sell, K., Hommes, F., Fischer, F., and Arnold, L. Multi-, inter-, and transdisciplinarity within the public health workforce: A scoping review to assess definitions and applications of concepts. International Journal of Environmental Research and Public Health, 19(17), 2022. ISSN 1660-4601. doi: 10.3390/ijerph191710902. URL https://www.mdpi.com/1660-4601/19/17/10902.

Shi, Feng ; Evans, J. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. Conference on Human Factors in Computing Systems | CHI Workshop 2024, 2024. doi: 10.1038/s41467-023-36741-4.

Si, C., Yang, D., and Hashimoto, T. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=M23dTGWCZy.

Xu, Z., Han, H., Li, L., Zhang, J., and Zhou, Z. Identifying multidisciplinary problems from scientific publications based on a text generation method. Journal of Data and Information Science, 9(3): 213–237, 2024. doi: 10.2478/jdis-2024-0021. URL https://doi.org/10.2478/jdis-2024-0021.

Yang, Z., Du, X., Li, J., Zheng, J., Poria, S., and Cambria, E. Large language models for automated open-domain scientific hypotheses discovery. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), Findings of the Association for Computational Linguistics: ACL 2024, pp. 13545–13565, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.804. URL https://aclanthology.org/2024.findings-acl.804/.

Yang, Z., Liu, W., Gao, B., Xie, T., Li, Y., Ouyang, W., Poria, S., Cambria, E., and Zhou, D. MOOSE-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=X9OfMNNepI.

Zheng, C., Zhang, Y., Huang, Z., Shi, C., Xu, M., and Ma, X. Disciplink: Unfolding interdisciplinary information seeking process via human-ai co-exploration. In Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, UIST '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706288. doi: 10.1145/3654777.3676366. URL https://doi.org/10.1145/3654777.3676366.

A. Appendix

A.1. Dataset Details

This section describes the details of our IDRBench dataset, including the details of our custom web platform and annotator background details.

A.1.1. Annotation Platform

To collect the positive samples in IDRBench, we designed a custom Web platform that includes four tasks for the annotators to complete. Specifically, the annotators are first provided with a panel of papers shown in Figure 3 when logged in, where they can choose the paper that they feel comfortable annotating. In task one, shown in Figure 4, they are instructed to first read the title and abstract of a given paper and decide whether it is an IDR paper. If so, they are directed to task two (Figure 5) to indicate the research type (i.e., whether it is basic research or applied research) of the IDR paper. In the third task, the annotators are asked to specify the exact papers that contribute to the IDR idea in this paper. They can add up to 4 papers that describe the IDR idea, as shown in the screenshot of Figure 6. Finally, they are asked to annotate the specific sentence(s) in this IDR paper that specifically describe such integration, as shown in Figure 7. The Web platform is also designed in a way where the annotators can revert any progress they have made so far in case they changed their minds, shown in Figure 8.

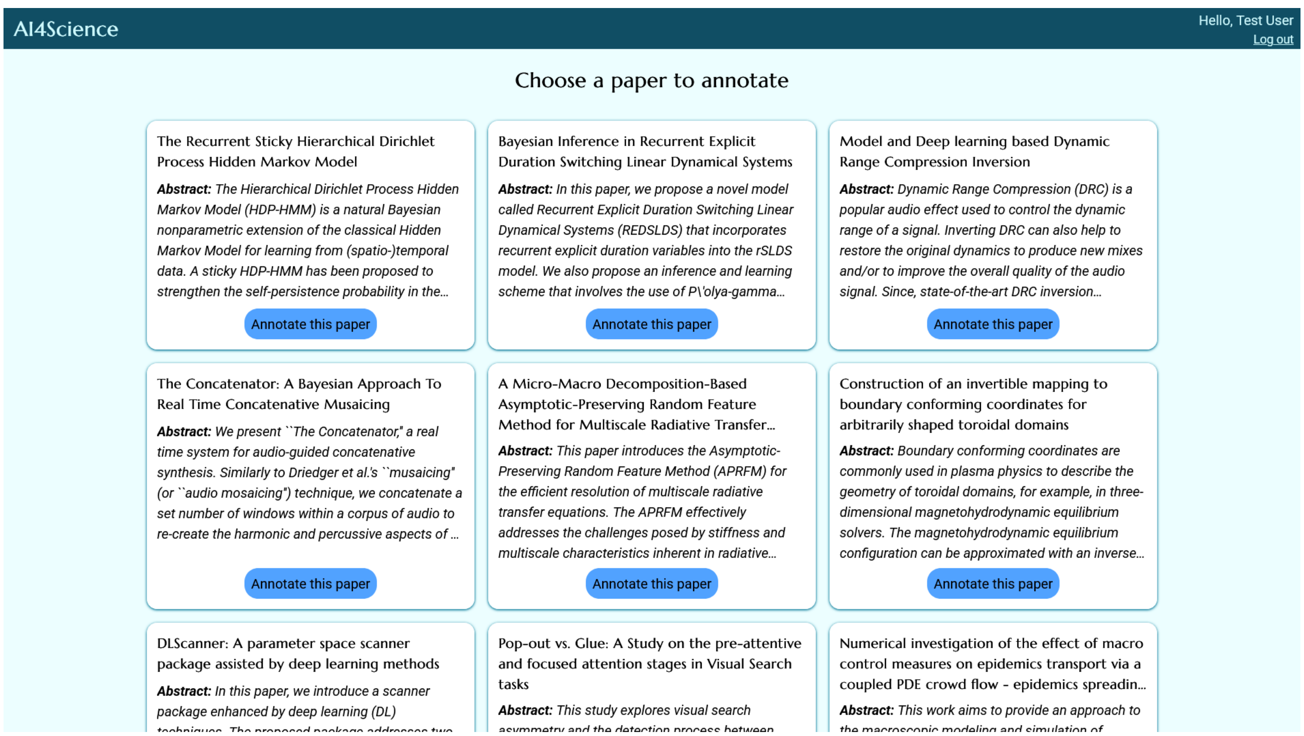

Figure 3: List of papers available for the annotator to choose from.

A.1.2. Human Annotation

We recruit 8 students from diverse academic backgrounds to perform the human annotation. Each student received about \$18 for each work hour. The academic level of the annotators ranges from the fourth year of undergraduate studies to the second year of PhD. Table 2 showcases a summary of the annotators' background, including their level of study and their area of expertise.

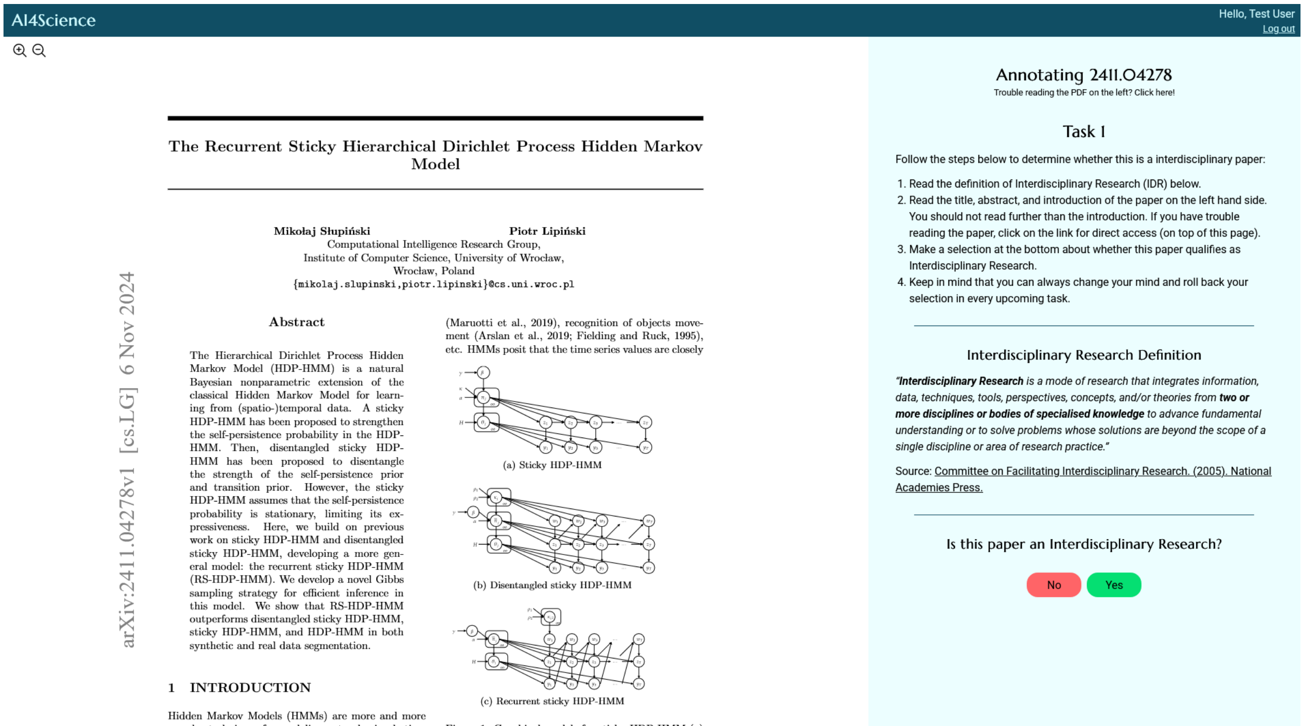

Figure 4: Task 1, displayed after the annotator selects a paper.

Level of study	Count	Expertise Areas
Undergraduate	3	Robotics, Control, AI, Natural Language Processing, Robotics, Rehabilitation, AI, Medical science, Mechanical and Materials Engineering, Mechatronics and Robotics Engineering, Brain-computer interfaces, Machine Learning
Master's	2	
PhD	3	

Table 2: Academic background summary of annotators.

A.2. Experiment Details

This section introduces the details of our experiment setup using IDRBench. Section A.2.1 introduces the models that we use in our evaluation, Section A.2.2 lists the evaluation metrics used to quantify the performance of different models, and Section A.2.3 provides the prompts that we used in our experiments.

A.2.1. Models

ChatGPT. First released in November 2022, ChatGPT is a series of models that is trained by OpenAI to conduct complex tasks using natural language prompts (et al., 2024b; OpenAI, 2024). In our paper, we select two types of GPT models from the GPT family, where gpt-4o-mini focuses on instruction following abilities, and gpt-o1-mini and gpt-o3-mini have enhanced reasoning capabilities in complex tasks.

Claude-3.7. This is the latest model released by Anthropic that incorporates an adaptive thinking process when prompted with a task. Specifically, it simulates an adaptive switch between the system 1 thinking and system 2 thinking, akin to human (Anthropic, 2024). We use the latest model claude-3.7-sonnet to run our benchmark.

Llama-3. As an updated version from llama-2 (et al., 2024a), llama-3 is trained with more recent corpora from various sources and achieves a better performance in various benchmarks. Different from the GPT family, Llama

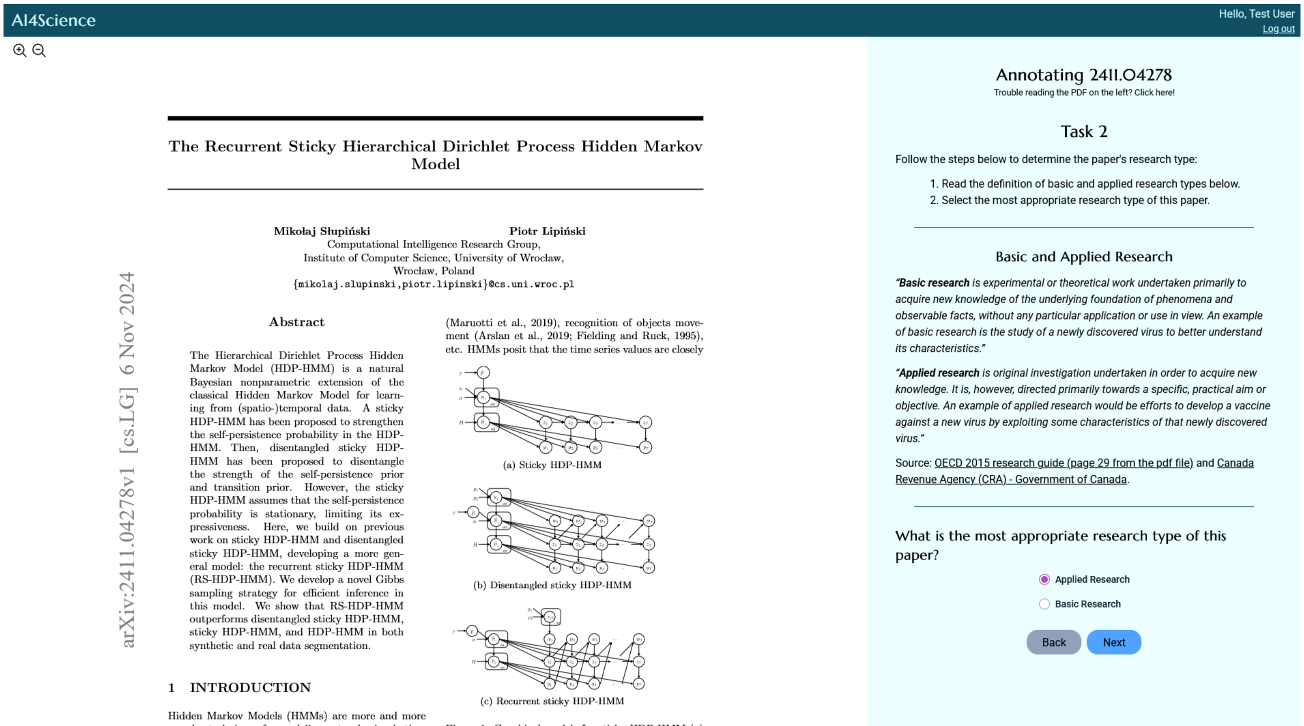

Figure 5: Task 2, displayed if the annotator answers "Yes" in Task 1.

models are completely open-source. In our evaluation, we use llama-3.1-70b and llama-3.3-70b, the two most powerful models in the family so far.

Deepseek. Deepseek is a suite of models trained under a schema that has higher efficiency yet with a comparable performance to those closed-source models (et al., 2025). The Deepseek family includes a basic chat model called deepseek-v3 and an advanced reasoning model deepseek-r1. We include both the chat model and the reasoning model in the results.

Model	Cutoff Date	Release Date
gpt-4o-mini	Oct 2023	Jul 2024
gpt-o1-mini	Oct 2023	Jul 2024
gpt-o3-mini	Oct 2023	Jul 2024
llama-3.1-70B	Dec 2023	Jul 2024
llama-3.3-70B	Dec 2023	Jul 2024
gemini 1.5-pro	May 2024	May 2024
gemini-2.0-flash	Jun 2024	Dec 2024
claude-3.7-Sonnet	Nov 2024	Feb 2025
deepSeek-v3	Jul 2024	Dec 2024
deepSeek-r1	Oct 2023	Dec 2024
Earliest paper in IDRBench	Nov 2024	

Table 3: Knowledge cutoff dates of LLMs' pretraining data.

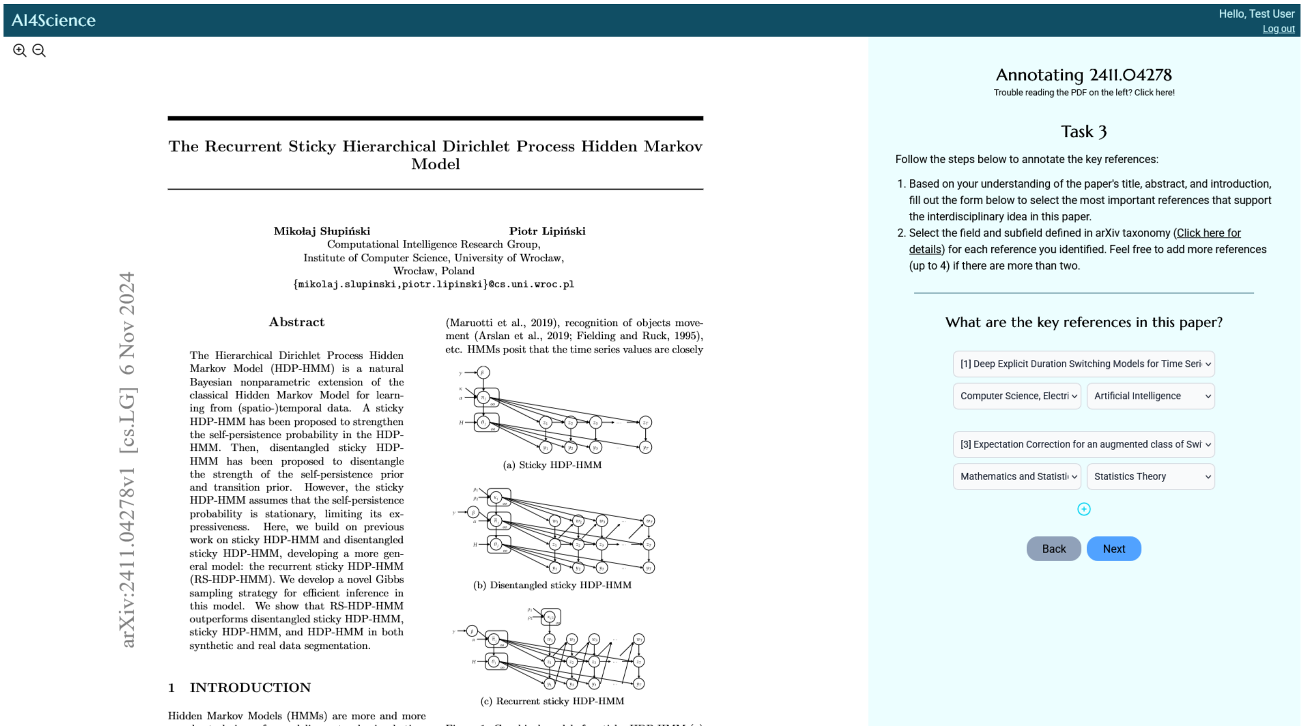

Figure 6: Task 3, displayed after completing Task 2.

A.2.2. Evaluation Metrics

We use a set of mainstream metrics for both our classification tasks and recommendation tasks. For the classification tasks, we report F1 and Macro F1 scores, and for the recommendation tasks, we report MRR. MRR measures, on average, how far down the ranked list the first relevant item appears. Specifically, for a set of queries Q, let rank_q be the position of the first relevant item for query $q \in Q$. The Reciprocal Rank (RR) for q is

$$\text{RR}_q = \begin{cases} \dfrac{1}{\text{rank}_q}, & \text{if a relevant item is found,} \\ 0, & \text{otherwise.} \end{cases}$$

The Mean Reciprocal Rank is then

$$\text{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \text{RR}_q = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_q}.$$

A.2.3. Prompts

In this section, we list the prompts that we used in our evaluation, including 1) IDR Paper Identification (IPI), 2) IDR Idea Integration (I3), and 3) IDR Idea Recommendation (I2R).

A.2.4. Experiment Cost

We use a variety of API-based platforms that host the models in our evaluation. The cost of the experiment was calculated based on the per-million input/output token rates listed by each host platform.

A.3. Detailed Experiment Results

A.3.1. IPI Results

We provide detailed IPI results with Precision, Recall, and F1 Scores included in Table ??.

IPI prompt
Read the title and abstract of a given academic paper and identify whether this is an interdisciplinary research paper. Also, select one or more subjects from the list below to indicate which subject(s) does this paper belong to. After you provide your verdict and your choice, provide a score from 0 to 100 to indicate your confidence level in the correctness of the verdict. The official definition of a typical interdisciplinary paper can be found below: "Interdisciplinary Research is a mode of research that integrates information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or bodies of specialised knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or area of research practice." Think carefully to make your verdict, answer "Yes" when this is a valid IDR paper. Otherwise, answer "No". Note: The confidence level indicates the degree of certainty you have about your verdict and is represented as a percentage. For instance, if your confidence level is 80, it means you are 80 percent certain that your answer is correct and there is a 20 percent chance that it may be incorrect. Paper title: %s; Paper abstract: %s; Subject list: ["Computer Science, Electrical Engineering and System Science", "Economics and Quantitative Finance", "Mathematics and Statistics", "Physics", "Quantitative Biology", "Other"] Use the template (in this format, with no markdown and lines separated by '\n') below to provide your answer. Your verdict: A simple answer containing either "Yes" or "No". Confidence score: A numeric score ranging from 0 to 100 Subject: Your choice of subjects from the list above. Use a list with square brackets "[]" separated by comma and remember to use "" to wrap your answer.

Table 4: Prompt used for the IPI task.

I3 prompt
Read the title and abstract of papers from two disciplines and decide whether you can extract concepts from both disciplines to create a novel multidisciplinary research idea. After you provide your verdict, provide a score from 0 to 100 to indicate your confidence level in the correctness of the verdict. Keep in mind a good Interdisciplinary Research idea includes the following standards: * This research idea should be Interdisciplinary, whereas the idea stems from the combination of ideas from the two papers introduced above. * The Interdisciplinary Research ideas should follow this definition: “Interdisciplinary Research is a mode of research that integrates information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or bodies of specialised knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or area of research practice.” * This research idea should be feasible, whereas the hypothesis is not purely theoretical and can be validated by experiments. * This research idea should be novel, whereas it is not only rare but also ingenious, imaginative, or surprising. * This research idea should be useful, whereas it applies to the stated problem and is effective at solving the problem. Think carefully to make your decision, and you should only answer ”Yes” when this multidisciplinary idea meets ALL of the standards above. Otherwise, you should answer ”No”. Note: The confidence level indicates the degree of certainty you have about your verdict and is represented as a percentage. For instance, if your confidence level is 80, it means you are 80 percent certain that your answer is correct and there is a 20 percent chance that it may be incorrect. Paper in Discipline 1: %s Paper in Discipline 2: %s Use the template (in this format, with no markdown and lines separated by ’\n’) to provide your answer. Your verdict: A simple answer containing either ”Yes” or ”No”. Your reason: A short paragraph less than 50 words briefly describes your reasons that you made the verdict above. Confidence score: A numeric score ranging from 0 to 100

Table 5: Prompt used for the I3 task.

I2R prompt (0-shot)
In this task, you are given a main paper introducing the key concepts that provides certain parts in a Interdisciplinary idea as well as two candidate papers that forms the remaining parts of a Interdisciplinary idea. Compare them and select which one is better to pair with the main paper in forming a multidisciplinary idea. After you provide your selection, provide a score from 0 to 100 to indicate your confidence level in the correctness of making this choice. Keep in mind a good Interdisciplinary Research idea includes the following standards: * This research idea should be Interdisciplinary, whereas the idea stems from the combination of ideas from the two papers introduced above. * The Interdisciplinary Research ideas should follow this definition: “Interdisciplinary Research is a mode of research that integrates information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or bodies of specialised knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or area of research practice.” * This research idea should be feasible, whereas the hypothesis is not purely theoretical and can be validated by experiments. * This research idea should be novel, whereas it is not only rare but also ingenious, imaginative, or surprising. * This research idea should be useful, whereas it applies to the stated problem and is effective at solving the problem. Note: The confidence level indicates the degree of certainty you have about your verdict and is represented as a percentage. For instance, if your confidence level is 80, it means you are 80 percent certain that your answer is correct and there is a 20 percent chance that it may be incorrect. Main paper title: %s; Main paper abstract: %s; Paper 1 title: %s; Paper 1 abstract: %s; Paper 2 title: %s; Paper 2 abstract: %s; Use the template (in this format, with no markdown and lines separated by '\n') to provide your answer. Your choice: A simple answer containing either ”Paper 1” or ”Paper 2”. Confidence score: A numeric score ranging from 0 to 100

Table 6: Prompt used for the I2R task.

IPI Few-shot Examples
Example 1: Paper title: Designing a Light-based Communication System with a Biomolecular Receiver; Paper abstract: Biological systems transduce signals from their surroundings in numerous ways. This paper introduces a communication system using the light-gated ion channel Channelrhodopsin-2 (ChR2), which causes an ion current to flow in response to light. Our design includes a ChR2-based receiver along with encoding, modulation techniques and detection. Analyzing the resulting communication system, we discuss the effect of different parameters on the performance of the system. Finally, we discuss its potential design in the context of bio-engineering and light-based communication and show that the data rate scales up with the number of receptors, indicating that high-speed communication may be possible. Your verdict: Yes Example 2: Paper title: BarcodeMamba: State Space Models for Biodiversity Analysis; Paper abstract: DNA barcodes are crucial in biodiversity analysis for building automatic identification systems that recognize known species and discover unseen species. Unlike human genome modeling, barcode-based invertebrate identification poses challenges in the vast diversity of species and taxonomic complexity. Among Transformer-based foundation models, BarcodeBERT excelled in species-level identification of invertebrates, highlighting the effectiveness of self-supervised pretraining on barcode-specific datasets. Recently, structured state space models (SSMs) have emerged, with a time complexity that scales sub-quadratically with the context length. SSMs provide an efficient parameterization of sequence modeling relative to attention-based architectures. Given the success of Mamba and Mamba-2 in natural language, we designed BarcodeMamba, a performant and efficient foundation model for DNA barcodes in biodiversity analysis. We conducted a comprehensive ablation study on the impacts of self-supervised training and tokenization methods, and compared both versions of Mamba layers in terms of expressiveness and their capacity to identify ”unseen” species held back from training. Our study shows that BarcodeMamba has better performance than BarcodeBERT even when using only 8.3% as many parameters, and improves accuracy to 99.2% on species-level accuracy in linear probing without fine-tuning for ”seen” species. In our scaling study, BarcodeMamba with 63.6% of BarcodeBERT’s parameters achieved 70.2% genus-level accuracy in 1-nearest neighbor (1-NN) probing for unseen species.; Your verdict: Yes Example 3: Paper title: An ADHD Diagnostic Interface Based on EEG Spectrograms and Deep Learning Techniques; Paper abstract: This paper introduces an innovative approach to Attention-deficit/hyperactivity disorder (ADHD) diagnosis by employing deep learning (DL) techniques on electroencephalography (EEG) signals. This method addresses the limitations of current behavior-based diagnostic methods, which often lead to misdiagnosis and gender bias. By utilizing a publicly available EEG dataset and converting the signals into spectrograms, a Resnet-18 convolutional neural network (CNN) architecture was used to extract features for ADHD classification. The model achieved a high precision, recall, and an overall F1 score of 0.9. Feature extraction highlighted significant brain regions (frontopolar, parietal, and occipital lobes) associated with ADHD. These insights guided the creation of a three-part digital diagnostic system, facilitating cost-effective and accessible ADHD screening, especially in school environments. This system enables earlier and more accurate identification of students at risk for ADHD, providing timely support to enhance their developmental outcomes. This study showcases the potential of integrating EEG analysis with DL to enhance ADHD diagnostics, presenting a viable alternative to traditional methods.; Your verdict: Yes

Table 7: Few-shot learning samples in IPI task.

IPI Few-shot Examples (Cont.)
Example 4: Paper title: Graph Neural Controlled Differential Equations For Collaborative Filtering; Paper abstract: Graph Convolution Networks (GCNs) are widely considered state-of-the-art for recommendation systems. Several studies in the field of recommendation systems have attempted to apply collaborative filtering (CF) into the Neural ODE framework. These studies follow the same idea as LightGCN, which removes the weight matrix or with a discrete weight matrix. However, we argue that weight control is critical for neural ODE-based methods. The importance of weight in creating tailored graph convolution for each node is crucial, and employing a fixed/discrete weight means it cannot adjust over time within the ODE function. This rigidity in the graph convolution reduces its adaptability, consequently hindering the performance of recommendations. In this study, to create an optimal control for Neural ODE-based recommendation, we introduce a new method called Graph Neural Controlled Differential Equations for Collaborative Filtering (CDE-CF). Our method improves the performance of the Graph ODE-based method by incorporating weight control in a continuous manner. To evaluate our approach, we conducted experiments on various datasets. The results show that our method surpasses competing baselines, including GCNs-based models and state-of-the-art Graph ODE-based methods.; Your verdict: No Example 5: Paper title: Mechano-Bactericidal Surfaces Achieved by Epitaxial Growth of Metal-Organic Frameworks; Paper abstract: Mechano-bactericidal (MB) surfaces have been proposed as an emerging strategy for preventing biofilm formation. Unlike antibiotics and metal ions that chemically interfere with cellular processes, MB nanostructures cause physical damage to the bacteria. The antibacterial performance of artificial MB surfaces relies on rational control of surface features, which is difficult to achieve for large surfaces in real-life applications. Herein, we report a facile and scalable method for fabricating MB surfaces based on metal-organic frameworks (MOFs) using epitaxial MOF-on-MOF hybrids as building blocks with nanopillars of less than 5 nm tip diameter, 200 nm base diameter, and 300 nm length. Two methods of MOF surface assembly, in-situ growth and ex-situ dropcasting, result in surfaces with nanopillars in different orientations, both presenting MB actions (bactericidal efficiency of 83% for E. coli). Distinct MB mechanisms, including stretching, impaling, and apoptosis-like death induced by mechanical injury are discussed with the observed bacterial morphology on the obtained MOF surfaces.; Your verdict: No

Table 8: Few-shot learning samples in IPI task (Cont.).

I3 Few-shot Examples
Example 1: Paper in Discipline 1: - title: ”Relation Between Retinal Vasculature and Retinal Thickness in Macular Edema” - abstract: ”This study has investigated the relationship of retinal vasculature and thickness for Macular Edema (ME) subjects. Ninety sets Fluorescein Angiograph (FA) Optical Coherence Tomography (OCT) 54 participants were analyzed. Multivariate analysis using binary logistic regression model was used to association between vessel parameters thickness. The results reveal feature i.e. fractal dimension (FD) as most sensitive parameter changes in associated with ME. Thus, indicating a direct which is caused due neovascular causing exudates, leakages hemorrhages, applications alternate modality detection” Paper in Discipline 2: - title: ”An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale” - abstract: ”While the Transformer architecture has become de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used replace certain components of networks while keeping their overall structure place. We show that this reliance on CNNs not necessary and a pure transformer directly sequences image patches can perform very well classification tasks. When pre-trained large amounts data transferred multiple mid-sized small recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision (ViT) attains excellent results compared state-of-the-art requiring substantially fewer computational resources train.” Your verdict: Yes Your reason: A novel work can combine transformers with two distinct methods that evaluate the quality of retinopathy”, Confidence score: 92 Example 2: Paper in Discipline 1: - title: ”Channelrhodopsin-2, a directly light-gated cation-selective membrane channel” - abstract: ”Microbial-type rhodopsins are found in archaea, prokaryotes, and eukaryotes. Some of them represent membrane ion transport proteins such as bacteriorhodopsin, a light-driven proton pump, or channelrhodopsin-1 (ChR1), recently identified light-gated channel from the green alga Chlamydomonas reinhardtii . ChR1 ChR2, related microbial-type rhodopsin C. , were shown to be involved generation photocurrents this alga. We demonstrate by functional expression, both oocytes Xenopus laevis mammalian cells, that ChR2 is directly light-switched cation-selective channel. This opens rapidly after absorption photon generate large permeability for monovalent divalent cations. desensitizes continuous light smaller steady-state conductance. Recovery desensitization accelerated extracellular H + negative potential, whereas closing decelerated intracellular expressed mainly under low-light conditions, suggesting involvement photoreception dark-adapted cells. The predicted seven-transmembrane helices characteristic G protein-coupled receptors but reflect different motif Finally, we may used depolarize small simply illumination.” Paper in Discipline 2: - title: ”Shannon capacity of signal transduction for multiple independent receptors, DESIGN AND IMPLEMENTATION OF VISIBLE LIGHT COMMUNICATION SYSTEM IN INDOOR ENVIRONMENT” - abstract: ”Cyclic adenosine monophosphate (cAMP) is considered a model system for signal transduction, the mechanism by which cells exchange chemical messages. Our previous work calculated Shannon capacity of single cAMP receptor; however, typical cell may have thousands receptors operating in parallel. In this paper, we calculate transduction with an arbitrary number independent, indistinguishable receptors. By leveraging prior results on feedback receptor, show (somewhat unexpectedly) that achieved IID input distribution, and n times receptor. Visible Light communication (VLC) using White Light Emitting Diode (LED) is a promising technology for next generation communication for short range, high speed wireless data transmission.”

Table 9: Few-shot learning samples in I3 task.

I3 Few-shot Examples (Cont.)
Your verdict: Yes Your reason: An interdisciplinary paper can aim to use channelrhodopsin-2 (ChR2), a biomolecule, as a receiver to design a light-based communication system, which is a work related to engineering. Confidence score: 85 Example 3: Paper in Discipline 1: - title: "A General Adaptive Dual-level Weighting Mechanism for Remote Sensing Pansharpening" - abstract: "Currently, deep learning-based methods for remote sensing pansharpening have advanced rapidly. However, many existing methods struggle to fully leverage feature heterogeneity and redundancy, thereby limiting their effectiveness. We use the covariance matrix to model the feature heterogeneity and redundancy and propose Correlation-Aware Covariance Weighting (CACW) to adjust them. CACW captures these correlations through the covariance matrix, which is then processed by a nonlinear function to generate weights for adjustment. Building upon CACW, we introduce a general adaptive dual-level weighting mechanism (ADWM) to address these challenges from two key perspectives, enhancing a wide range of existing deep-learning methods. First, Intra-Feature Weighting (IFW) evaluates correlations among channels within each feature to reduce redundancy and enhance unique information. Second, Cross-Feature Weighting (CFW) adjusts contributions across layers based on inter-layer correlations, refining the final output. Extensive experiments demonstrate the superior performance of ADWM compared to recent state-of-the-art (SOTA) methods. Furthermore, we validate the effectiveness of our approach through generality experiments, redundancy visualization, comparison experiments, key variables and complexity analysis, and ablation studies. Our code is available at https:github.comJie-1203ADWM." Paper in Discipline 2: - title: "Secure Semantic Communication With Homomorphic Encryption" - abstract: "In recent years, Semantic Communication (SemCom), which aims to achieve efficient and reliable transmission of meaning between agents, has garnered significant attention from both academia and industry. To ensure the security of communication systems, encryption techniques are employed to safeguard confidentiality and integrity. However, traditional cryptography-based encryption algorithms encounter obstacles when applied to SemCom. Motivated by this, this paper explores the feasibility of applying homomorphic encryption to SemCom. Initially, we review the encryption algorithms utilized in mobile communication systems and analyze the challenges associated with their application to SemCom. Subsequently, we employ scale-invariant feature transform to demonstrate that semantic features can be preserved in homomorphic encrypted ciphertext. Based on this finding, we propose a task-oriented SemCom scheme secured through homomorphic encryption. We design the privacy preserved deep joint source-channel coding (JSCC) encoder and decoder, and the frequency of key updates can be adjusted according to service requirements without compromising transmission performance. Simulation results validate that, when compared to plaintext images, the proposed scheme can achieve almost the same classification accuracy performance when dealing with homomorphic ciphertext images. Furthermore, we provide potential future research directions for homomorphic encrypted SemCom." Your verdict: No Your reason: The two papers are not related to each other. The first paper focuses on remote sensing pansharpening, while the second paper discusses secure semantic communication with homomorphic encryption. There is no clear interdisciplinary connection between them. Confidence score: 90

Table 10: Few-shot learning samples in I3 task (Cont.).

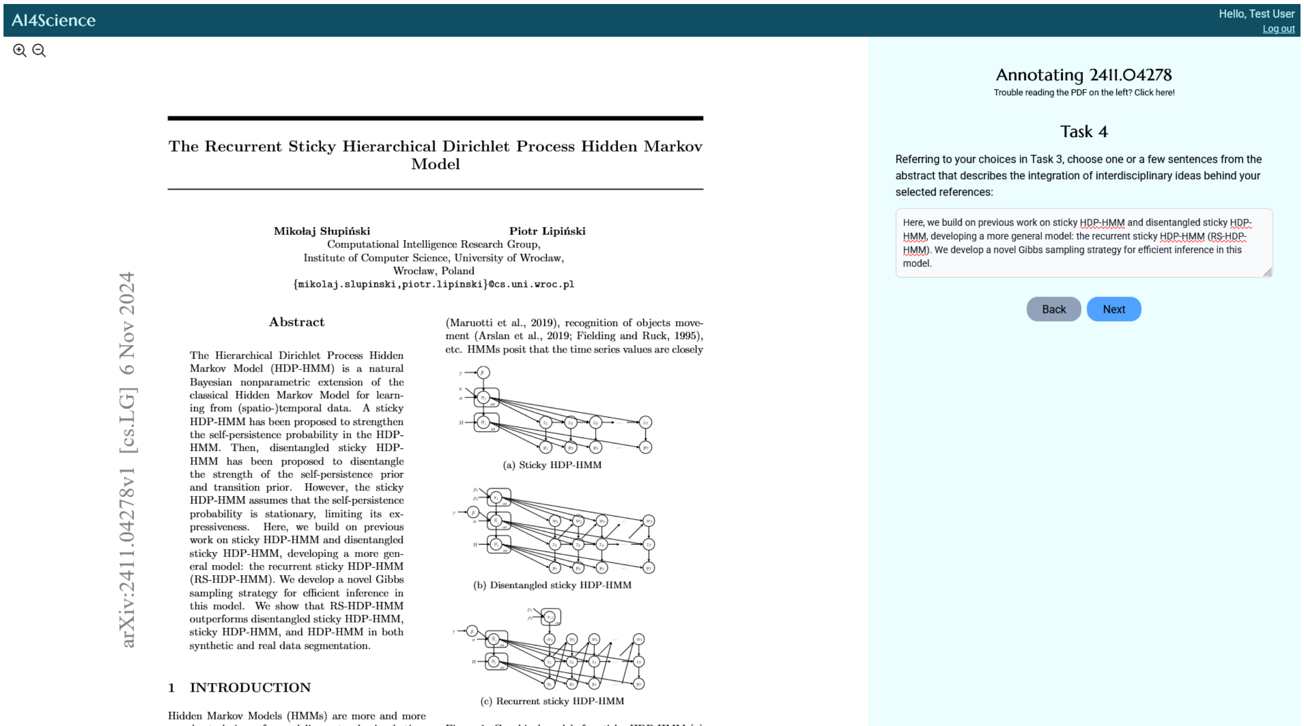

Figure 7: Task 4, displayed after completing Task 3.

A.3.2. I3 Results

We provide detailed I3 results in Table 12.

A.3.3. I2R Results

We provide detailed I3 results in Table 13.

A.4. Analysis

In this section, we provide the results in the analysis introduced in Section 4: Table 14 presents the performance of different models on discipline classification accuracy under the zero-shot setting. Table 15 presents the comparison results of IDR idea quality across different LLMs. Table 16 reports the Kendall's τ correlation between the rankings produced by the LLM and those based on contextual similarity (using SciBERT similarity).

A.4.1. Additional Analysis on I3

In addition to the reporting of similarity scores in Section 4, we also conduct a qualitative analysis on the LLM output reasons against human researchers' original writing on formulating a potential IDR research. Our analysis is twofold: on one hand, we compare the key integration reasoning from LLM output with the annotated sentences marked by the annotators; on the other hand, we compare the full abstract generated by LLM with the abstract of the targeted IDR paper. Figure 9 provides a comparison between the annotation and the LLM-generated response. As we can see, although LLMs like llama-3.3-70b can generate topic-related content in both integration reasoning and abstract generation, they still lack detailed methods and a solid experiment schedule in formulating an IDR project.

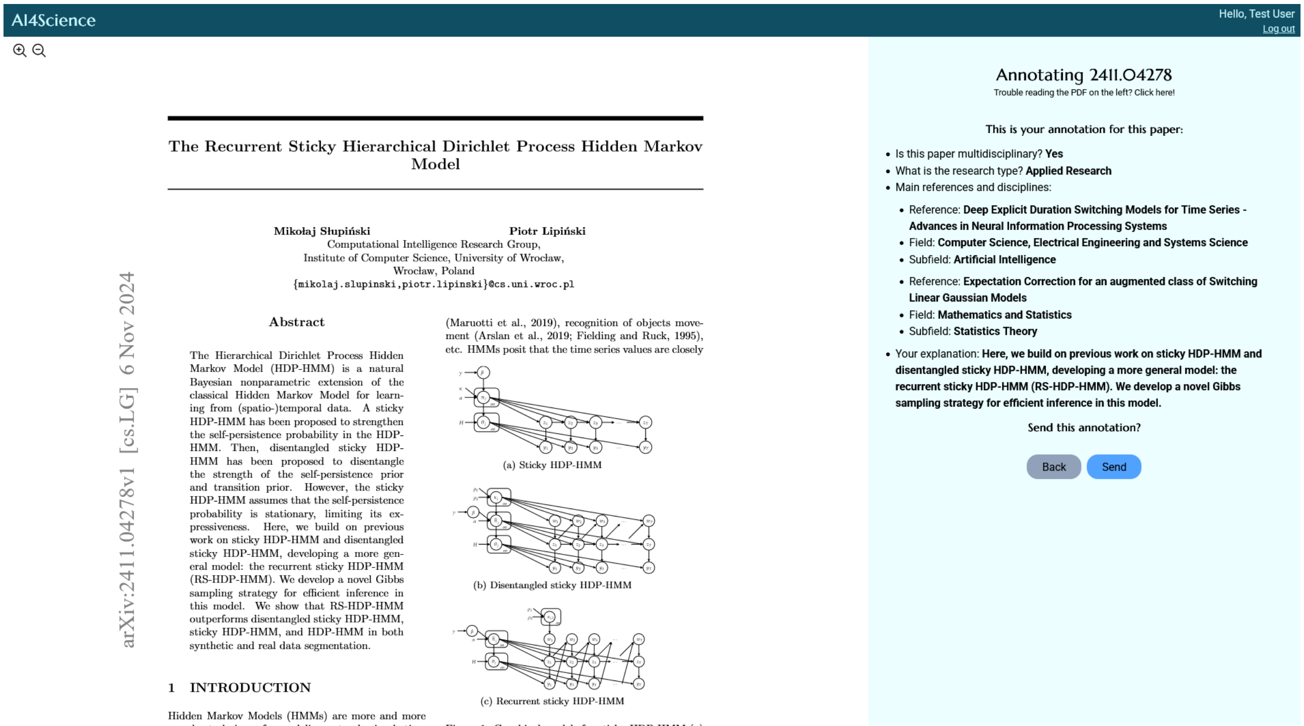

Figure 8: Review page, where the annotator can review and optionally go back and edit their answers.

Prompting	Metric	ChatGPT Mini			Gemini		Llama 3		Claude	DeepSeek	
		4o	o1	o3	1.5-pro	2.0	3.1	3.3	3.7	v3	r1
0-shot	Precision	0.268	0.264	0.235	0.161	0.205	0.283	0.227	0.211	0.274	0.254
	Recall	0.758	0.942	0.975	1.000	0.950	0.778	0.850	0.967	0.658	0.950
	F1	0.396	0.412	0.379	0.277	0.337	0.416	0.358	0.347	0.387	0.401
5-shot	Precision	0.266	0.283	0.275	0.189	0.232	0.177	0.193	0.174	0.235	0.310
	Recall	0.758	0.958	0.975	0.908	0.767	0.933	0.958	0.983	0.642	0.875
	F1	0.394	0.436	0.429	0.313	0.357	0.298	0.321	0.295	0.344	0.458

Table 11: Classification accuracy of LLMs on IPI task in IDRBench.

	Prompting	Metric	ChatGPT Mini			Gemini		Llama 3		Claude	DeepSeek	
			4o	o1	o3	1.5-pro	2.0	3.1	3.3	3.7	v3	r1
Level 1	0-shot	Precision	0.867	0.549	0.138	0.560	0.511	0.612	0.828	0.369	0.805	0.486
		Recall	0.542	0.467	0.875	0.467	0.750	0.607	0.642	0.400	0.583	0.725
		F1	0.667	0.505	0.239	0.509	0.608	0.609	0.723	0.384	0.676	0.582
	3-shot	Precision	0.820	0.478	0.135	0.210	0.282	0.382	0.528	0.547	0.642	0.258
		Recall	0.683	0.367	0.675	0.750	0.975	0.892	0.858	0.775	0.733	0.767
		F1	0.745	0.415	0.226	0.328	0.437	0.535	0.654	0.641	0.685	0.386
Level 2	0-shot	Precision	0.336	0.101	0.048	0.165	0.103	0.158	0.264	0.073	0.192	0.115
		Recall	0.544	0.367	0.844	0.467	0.733	0.663	0.678	0.367	0.556	0.722
		F1	0.415	0.158	0.091	0.243	0.181	0.255	0.380	0.122	0.286	0.198
	3-shot	Precision	0.220	0.090	0.049	0.054	0.073	0.092	0.109	0.131	0.153	0.071
		Recall	0.700	0.333	0.733	0.700	0.989	0.833	0.867	0.733	0.722	0.789
		F1	0.334	0.142	0.092	0.100	0.135	0.166	0.193	0.223	0.253	0.130

Table 12: Classification accuracy of LLMs on I3 task in IDRBench.

Difficulty	Metric	ChatGPT Mini			Gemini		Llama 3		Claude	DeepSeek	
		4o	o1	o3	1.5-pro	2.0	3.1	3.3	3.7	v3	r1
Level 1	MRR	0.682*	0.589	0.640	0.703	0.687*	0.622	0.619	0.679*	0.674*	0.588
Level 2	MRR	0.645*	0.500	0.571	0.662	0.621*	0.576	0.574	0.636*	0.650*	0.526

Table 13: Recommendation precision of I2R task. Numbers with "$*$" indicate that there is no statistical significance against the bold value.

ChatGPT Mini			Gemini		Llama 3		Claude	DeepSeek	
4o	o1	o3	1.5-pro	2.0	3.1	3.3	3.7	v3	r1
0.845	0.877	0.874	0.906	0.903	0.795	0.898	0.876	0.895	0.884

Table 14: 0-shot discipline identification accuracy among positive samples, reported in F1-Score.

	Metric	ChatGPT Mini			Gemini		Llama 3		Claude	DeepSeek	
		4o	o1	o3	1.5-pro	2.0	3.1	3.3	3.7	v3	r1
Integration Reason	Similarity	0.747	0.759	0.785	0.744	0.745	0.756	0.763	0.747	0.760	0.757
	BLEU	0.013	0.010	0.011	0.012	0.012	0.014	0.006	0.011	0.011	0.011
	ROUGE-1	0.218	0.160	0.182	0.186	0.206	0.215	0.145	0.195	0.192	0.161
Abstract	Similarity	0.800	0.830	0.818	0.802	0.798	0.793	0.785	0.829	0.831	0.837
	BLEU	0.015	0.014	0.014	0.016	0.016	0.017	0.018	0.014	0.014	0.012
	ROUGE-1	0.328	0.321	0.321	0.325	0.331	0.328	0.338	0.314	0.318	0.307

Table 15: Similarity comparison of integration reason and abstract between LLM-generated responses and the actual IDR paper. We use I3-Level 1 dataset and 0-shot setting.

Difficulty	ChatGPT Mini			Gemini		Llama 3		Claude	DeepSeek	
	4o	o1	o3	1.5-pro	2.0	3.1	3.3	3.7	v3	r1
Level 1	0.196	0.199	0.180	0.233	0.220	0.221	0.211	0.249	0.232	0.194
Level 2	0.180	0.158	0.162	0.196	0.184	0.184	0.187	0.200	0.194	0.169

Table 16: Kendall correlation among I2R ranking and similarity scores of seed and candidate papers.

Key Idea Integration Comparison

Paper Title: A lightweight Convolutional Neural Network based on U shape structure and Attention Mechanism for Anterior Mediastinum Segmentation

LLM I3 Reasoning: Combining AI-powered lesion segmentation with lung cancer screening guidelines can improve diagnosis accuracy.

Annotated I3 Reasoning: In this paper, we introduce a U-shaped structure network to segment AM. Two attention mechanisms were used for maintaining long-range dependencies and localization. In order to have the potential of Multi-Head Self-Attention (MHSA) and a lightweight network, we designed a parallel MHSA named Wide-MHSA (W-MHSA).

Full Abstract Generation Comparison

LLM Abstract: This study proposes an interdisciplinary approach to investigating incidental anterior mediastinal nodular lesions on chest computed tomography (CT) scans in asymptomatic subjects, with a focus on early detection of lung cancer and thymic epithelial tumors. By integrating a UNet-based neural network with attention mechanisms for automatic detection of anterior mediastinal nodular lesions, and guidelines for lung cancer screening from the American Cancer Society, this research aims to develop a novel computer-aided detection system. The system will analyze low-dose chest CT scans to identify high-risk individuals and detect incidental nodular lesions, which can be potential thymic epithelial tumors or lung cancer. The expected outcome of this study is to improve the accuracy of early detection and diagnosis of lung cancer and thymic epithelial tumors, and to provide valuable insights into the prevalence and characteristics of incidental nodular lesions in asymptomatic subjects.

True Abstract: To automatically detect Anterior Mediastinum Lesions (AMLs) in the Anterior Mediastinum (AM), the primary requirement will be an automatic segmentation model specifically designed for the AM. The prevalence of AML is extremely low, making it challenging to conduct screening research similar to lung cancer screening. Retrospectively reviewing chest CT scans over a specific period to investigate the prevalence of AML requires substantial time. Therefore, developing an Artificial Intelligence (AI) model to find location of AM helps radiologist to enhance their ability to manage workloads and improve diagnostic accuracy for AMLs. In this paper, we introduce a U-shaped structure network to segment AM. Two attention mechanisms were used for maintaining long-range dependencies and localization. In order to have the potential of Multi-Head Self-Attention (MHSA) and a lightweight network, we designed a parallel MHSA named Wide-MHSA (W-MHSA). Maintaining long-range dependencies is crucial for segmentation when we upsample feature maps. Therefore, we designed a Dilated Depth-Wise Parallel Path connection (DDWPP) for this purpose. In order to design a lightweight architecture, we introduced an expanding convolution block and combine it with the proposed W-MHSA for feature extraction in the encoder part of the proposed U-shaped network. The proposed network was trained on 2775 AM cases, which obtained an average Dice Similarity Coefficient (DSC) of 87.83%, mean Intersection over Union (IoU) of 79.16%, and Sensitivity of 89.60%. Our proposed architecture exhibited superior segmentation performance compared to the most advanced segmentation networks, such as Trans Unet, Attention Unet, Res Unet, and Res Unet++.

Figure 9: Qualitative comparison sample on both idea integration reasoning and full abstract generation using llama-3.3-70b, both under zero-shot setting.