Think Like a Researcher: A Dataset for Scientific Ideation with Large Language Models

Julia Moska[0009-0003-8581-1098], Maciej Piasecki[0000-0003-1503-0993], and Arkadiusz Janz[0000-0002-9203-5520]

Wrocław University of Science and Technology, Poland
{julia.moska}@pwr.edu.pl

Abstract. Research hypothesis generation using Large Language Models (LLMs) remains largely prompt-based, with limited work on model alignment. We present one of the few systematic reviews of existing datasets for hypothesis generation, analyzing their structure and suitability for alignment. Based on this analysis, we introduce a dataset designed for training and aligning LLMs that encodes literature-derived background knowledge as structured concept connections, along with evaluation metrics grounded in references and their content. Using this dataset, we study how alignment with structured knowledge affects the novelty and grounding of generated hypotheses.

Keywords: scientific ideation · hypothesis generation · LLMs · preference dataset · alignment · LLM fine-tuning

1 Introduction

Advances in LLMs enabled new NLP applications including creative generation, mathematical proof synthesis, and drug discovery. However, autonomous research, focused on LLM-based scientific hypothesis generation, designing experiments, and supporting publishable research is still challenging due to the challenges of literature review, identification of knowledge gaps, ideation and planning, and execution of experiments. *Idea generation* remains underexplored in several aspects. Existing solutions [6] often rely on API access, which limits reproducibility and adaptation. The field still lacks robust datasets and benchmarks. Scientific ideas must be novel, grounded in prior work, and methodologically feasible. Current evaluation methods, such as semantic similarity or LLM-as-a-judge [7], only partially address this challenge, often oversimplifying novelty and ignoring the motivational and structural context of idea generation. Moreover, existing datasets often represent references as flat collections of related papers [17], rather than as structured and evolving knowledge.

Following [17, 6, 22] we introduce a dataset for scientific ideation built from structured, pre-processed background knowledge. It pairs ground-truth work with semantically labeled citation graphs and structured summaries (Fig. 1), enabling fine-tuning, alignment, and evaluation for scientific ideation. Compared with [6], we construct an offline semantically filtered multi-hop citation graph,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_35

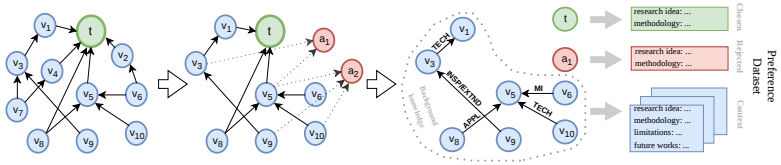

Fig. 1: A citation network is mined and processed to create chosen-rejected research idea–method pairs, forming DPO preference samples [18].

preserving fine-grained relations between articles. Unlike [17], our approach retains a broader citation context and represents each paper with structured summaries of its ideas, methods, limitations, and future directions. In contrast to [22], our representations are deterministic, reusable, and independent of a specific inference-time pipeline. Valid research ideas may diverge from the ground truth while still addressing meaningful gaps. We introduce metrics beyond surface-level semantic similarity to assess conceptual relevance and logical coherence with the citation context. We use them to compare preference-aligned models against base models and reasoning models evaluated under the same conditions.

2 Related work

Scientific idea generation is framed as inference-time retrieval over scientific knowledge. Systems start from a seed paper or query and traverse citation graphs to collect context for generation [6, 19], typically through public APIs such as arXiv[1] or Semantic Scholar [2]. Training-based alternatives use SFT [15], [5] apply preference learning over topic-aligned abstracts, and [8] explore reinforcement learning in a domain-specific multi-agent setting. Such approaches may overlook prior problems with correct grounding essential for relevant idea generation. Most prior work relies on inference-time API access, limiting task-specific fine-tuning, and controlled experimentation. Although resources such as ArXiv Dataset [4], Open Academic Graph [20], SciPIP [22], and BigSurvey [12] provide metadata and citations, but often reduced to titles and abstracts. Additional structures may come from IE [21, 15], but such methods lack validation (Tab. 1).

The evaluation of scientific ideas, a non-trivial task for humans, is a challenge to automate, leading to proxy-based metrics. The novelty assessment is carried out by cosine similarity to existing work [19, 7], embedding-based distance [5], and comparing similarity to both historical and contemporary literature [17]. Although novelty remains a central focus, other dimensions such as feasibility, diversity, clarity, and potential impact [7, 19, 9, 6] are considered in a holistic assessment. The use of the LLM-as-a-Judge framework is widespread [11, 17], but the evaluation of scientific ideation remains a challenge due to the lack of sufficient human annotation. Sections such as future work and limitations are only rarely used as explicit signals in the evaluation process.

[1] `https://arxiv.org/help/api/`
[2] `https://api.semanticscholar.org/`

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_35

Table 1: Existing datasets for scientific ideation along key dimensions: domain, data format, availability of k-hop citation graph (K-REF), readiness for SFT (SFT-R) and alignment training (AL-R), validation of the paper summaries and extracted metadata (HV), availability of the data in open access (OA).

Dataset	Domain	Data format	K-REF	SFT-R	AL-R	OA	HV
[4]	multi-domain	abstracts/full texts	✗/ ✓	✗	✗	✗/ ✓	N/A
[12]	multi-domain	abstracts	✗	✗	✓	✓	N/A
[7]	LLMs	abstracts	✗	✗	✗	✗	N/A
[20]	multi-domain	abstracts	✓	✗	✗	✓	N/A
[5]	multi-domain	abstracts	✗	✓	✓	✗	N/A
[21, 17]	biomedical	abstracts	✗	✓	✗	✓	N/A
[10]	AI	full text*	✗	✓	✗	✗	N/A
[1]	biomedical	knowledge graph	✗	✗	✗	✓	N/A
[15, 16]	biomedical	summaries	✗	✓	✗	✗	✗
[13]	computer-science	summaries	✗	✓	✗	✓	✗
[22]	AI	summaries	✗	✗	✗	✓	✗
ours	NLP/AI	summaries	✓	✓	✓	✓	✓

3 Alignment Data for Scientific Ideation

We propose a scientific idea generation dataset for both training and evaluation. It combines structured knowledge with textual summaries of key paper content[3], and represents the literature as a semantically annotated graph (Fig. 1), enabling a richer view of the research context evolution. To avoid potential data leakage, the dataset construction began with the test set. The research papers were retrieved via Semantic Scholar, using keywords covering several NLP terms within the core area of expertise of the authors. 223 target articles T were identified, which form the basis of the evaluation. The publication cut-off date was chosen to ensure that the content was not seen by the selected open-source models. Full-text content was extracted using the OLMOCR[4].

For each target article $t \in T$, citation data was retrieved up to three hops from the global citation graph G to build a local literature subgraph $G_t = (V_t, E_t)$, where V_t are articles and E_t citation edges. We applied the personalised PageRank algorithm [14] to identify the most relevant nodes in G_t. UsingQwen2 1.5B GTE[5], each node $v_j \in V_t$ received a relevance score $w_j = \text{cos-sim}(v_j, t)$, capturing semantic similarity to t. After normalizing these scores into a valid personalization vector, we pruned nodes below the dynamic threshold $\mu + \sigma$, resulting in a reduced and semantically focused subgraph $\tilde{G}_t \subseteq G_t$. Additionally, we define proximate articles as those with substantial structural overlap with the target article's local citation subgraph G_t. For each node $v_j \in V_t$, we searched the global citation graph G for external articles $a_j \notin V_t$ that cite v_j, producing candidate neighbours that reference a subset of nodes in V_t (see Fig.1). For each a_j, we defined its reference set V_{a_j} and measured overlap with V_t using Jaccard set similarity $Jaccard(V_{a_j}, V_t) = \frac{|V_{a_j} \cap V_t|}{|V_{a_j} \cup V_t|}$. Articles with scores exceeding experimen-

[3] Dataset: https://github.com/mskaa3/think-like-a-researcher
[4] https://olmocr.allenai.org
[5] https://huggingface.co/Alibaba-NLP/gte-Qwen2-1.5B-instruct

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_35

tally established thresholds $\Theta = \langle 0.3, 0.4, 0.5 \rangle$ were retained in neighbourhood sets $\mathcal{A}_t(\Theta)$, encoding increasing levels of structural similarity.

Inter-paper Relation Recognition. Building on [6], we expand the typology of inter-paper relations (see Fig. 1), by annotating each citation edge $(v_j, v_k) \in \tilde{G}_t$ with a semantic citation label $\ell_{jk} \in \mathcal{L}$ using LLaMA 3 70B[6]. Given that the semantic similarity used to filter the initial graph does not capture nuanced dependencies, these labels aim to capture how one work relates conceptually or methodologically. Relation types include Inspiration, Extension, Critique, Foundation, Replication, Comparison, Application, Supplement, Historical background, Technical, and Marginal Influence. We developed detailed annotation guidelines and had a subset of citation links manually annotated by two domain experts. Agreement was measured using Gwet's AC1 (Table 2), chosen for its robustness against the high-agreement–low-kappa problem [3], particularly relevant in imbalanced data. To evaluate relation recognition, we computed the agreement score between domain experts and LLaMA 3 70B (Table 2).

Table 2: Inter-annotator agreement (IAA) and model–annotator agreement (AI–human) scores per citation category, measured using Gwet's AC1.

	FOUND	EXTND	SUPPL	REPL	INSP	CRIT	COMPR	APPL	MENT
$AC1_{\text{Human-AI}}$	0.747	0.413	0.241	0.842	0.906	0.837	0.368	0.820	0.928
$AC1_{\text{IAA}}$	0.714	0.631	0.453	0.964	0.926	0.855	0.309	0.850	0.947

Paper Summary Extraction. To obtain concise but informative article representations, we used GPT-4o mini[7] to generate structured summaries covering four dimensions: *research idea*, *methodology*, *limitations*, and *future directions*. They serve both as elements of learning examples and as an evaluation basis. A domain expert assessed summary quality by comparing the extracted descriptions with the original papers on a five-point scale.The results indicate good extraction quality, with average scores of 3.95±0.47 for research idea, 3.36±0.63 for method details, 3.54±0.85 for limitations, and 3.40±0.77 for future work. We retrieved an additional 1,000 articles published before Oct. 2024 using the same keyword set as training set, following the same processing pipeline, with any overlap with the test set removed. Each data instance is a tuple $(\tilde{G}_t, t, A_t)$, where $\tilde{G}_t$ is a directed citation graph, with each node $v_j \in \tilde{G}_t$ represented as $v_j = (r_j, m_j, l_j, f_j)$, and each edge (v_j, v_k) annotated with semantic label $\ell_{jk} \in \mathcal{L}$. The target article t is abstracted as $t = (r_t, m_t)$. Proximate article sets $\mathcal{A}_t(\theta)$ for thresholds $\theta \in \{0.3, 0.4, 0.5\}$ where $a_j \in \mathcal{A}_t(\theta)$ represented as $a_j = (r_{a_j}, m_{a_j})$.

4 Evaluation of Model-Driven Scientific Ideas

Scientifically valid ideas may diverge from the gold outputs by extending, reinterpreting, or building on prior work. To capture contextual grounding, we introduce measures based on limitations and future directions identified in prior

[6] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
[7] https://openai.com

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_35

works, revealing scientific motivations and constraints. Let $(\tilde{G}_t, t, \mathcal{A}_t)$ denote the representation of a test instance as defined in Sec. 3. The k generated ideas for t are embedded as vectors $\mathbf{g}_1, \ldots, \mathbf{g}_k$, using QWEN2 1.5B GTE, while the target idea is represented by $\mathbf{t}$. The context embeddings set is $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, where each $\mathbf{v}_j$ corresponds to a referenced article $v_j \in \tilde{G}_{t_i}$. The proximate article embeddings is $\mathcal{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$, where $\mathbf{a}_j$ corresponds to $a_j \in \mathcal{A}_t$.

(1) Inter-Generation Dissimilarity. To assess idea diversity, we compute pairwise cosine distances among the k generations for each test example. The diversity score is $D = \frac{2}{k(k-1)} \sum_{1 \le i < j \le k} (1 - \cos(\mathbf{g}_i, \mathbf{g}_j))$, where $\cos(\mathbf{g}_i, \mathbf{g}_j)$ is the cosine similarity. The final score is the average D over all test examples.

(2) Gold Similarity. We compute the cosine similarity between the embeddings of each generated idea $\mathbf{g}_i$ and the reference answer $\mathbf{t}$ as $GoldSim. = \frac{1}{k} \sum_{i=1}^{k} \cos(\mathbf{g}_i, \mathbf{t})$. Scores are averaged across all examples to provide a clear measure of how closely the model adheres to the target response, as in [19, 7].

(3) Relative Novelty. Inspired by [17], we define a relative novelty score as the ratio of an idea's semantic similarity to proximate articles ($\mathcal{A}$) to its similarity to background references ($\mathcal{V}$): $N = \frac{1+s(\mathcal{A})}{1+s(\mathcal{V})}$. Our approach utilizes references as historical context and all existing proximate articles as current context. The normalized cosine similarity between the generated ideas and the reference set $\mathcal{X}$ is $s(\mathcal{X}) = \frac{1}{k|\mathcal{X}|} \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathcal{X}} \frac{\cos(\mathbf{g}_i, \mathbf{x})+1}{2}$.

(4) Limitations and Future Directions Alignment Scores. For each test article, we extract limitations $L = \{l_1, \ldots, l_n\}$ and future directions $F = \{f_1, \ldots, f_n\}$ from reference papers in V. To assess whether a generated idea addresses a limitation or aligns with a future direction, we use an LLM-as-a-judge function $M(g_i, s_j)$, where $s_j \in L$ or $s_j \in F$, returning 1 if the relation holds. For $S \in \{L, F\}$, where $S = \{s_1, \ldots, s_{|S|}\}$, we define

$$A_{\max}(S) = \max_{i=1}^{k} \left(\frac{1}{|S|} \sum_{j=1}^{|S|} M(g_i, s_j) \right), A_{\text{avg}}(S) = \frac{1}{k} \sum_{i=1}^{k} \left(\frac{1}{|S|} \sum_{j=1}^{|S|} M(g_i, s_j) \right) \quad (1)$$

where $M(g_i, s_j) = 1$ if generated idea g_i addresses s_j, and 0 otherwise. For $S = L$, this yields $LA_{\max}$ and LA_{avg}; for $S = F$, $FDA_{\max}$ and FDA_{avg}.

(5) Topic Consistency Score. It extends the thematic coherence of generated ideas [5] and idea-to-topic matching [17] by measuring the alignment of generated ideas with keyword-based topics derived from the set of reference papers (V).

5 Results and Conclusions

We fine-tuned QWEN 7B INSTRUCT[8] with DPO [18]. The prompting-only base model serves as the main baseline, and DEEPSEEK-R1-DISTILL-QWEN-7B[9] as an additional reasoning baseline. Rejected samples are drawn either from proximate articles $\mathcal{A}_t$ or randomly from the corpus, defining proximate and unrelated negative settings. Inputs follow [2] and consist of an instruction with a textual

[8] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[9] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_35

citation-graph representation with node summaries and semantic edge labels (Fig. 1); MI, TECH, and HIST edges are removed. We compare full summaries (research idea, methodology, limitations, future directions) with reduced summaries (research idea, methodology only). Training uses AdamW for 10 epochs with learning rate 10^{-6}, weight decay 0.001, 5% warm-up, linear scheduling, and $\beta = 0.01$. The evaluation is performed at $T \in \{0.1, 0.5\}$ with 10 generations per instance. Tab. 3 demonstrates the impact of alignment strategies and contextual information on the quality of generated research ideas.

Effect of DPO Alignment. The alignment improves core semantic metrics but reduces inter-generation diversity. The drop in limitations alignment suggests that increased consistency comes at the cost of overfitting to surface-level patterns rather than engaging with the reference literature. Compared with the base model, which draws on a broader range of background material and addresses limitations more variably, fine-tuning narrows behavior toward a smaller set of preferred patterns, reducing perspective and creativity in a manner reminiscent of catastrophic forgetting.

Table 3: Evaluation results of baseline and DPO-aligned models under different contextual configurations, reported at temperatures (T) 0.1 and 0.5. Baselines are evaluated in a prompting-only setup, while aligned models are fine-tuned with different rejected-sample strategies and context keys derived from structured summaries.

Model	Gold Sim.	Rel. Nov.	Dissim.	LA_{avg}	LA_{max}	FDA_{avg}	FDA_{max}	TCS	T
keys in context: research idea, methodology, limitations, future works									
Q_{base}	0.585	0.878	6.386	**0.239**	0.472	0.277	0.459	0.177	0.1
$Q_{prox.}$	0.591	**0.891**	5.155	0.184	0.275	0.219	0.374	0.203	
$Q_{unrel.}$	**0.602**	**0.892**	4.432	0.096	0.171	0.204	0.371	**0.217**	
$Q_{reason.}$	0.554	0.884	13.621	0.228	**0.482**	**0.322**	**0.577**	0.168	
Q_{base}	0.584	0.879	10.687	0.207	**0.461**	0.288	0.515	0.176	0.5
$Q_{prox.}$	0.586	**0.891**	9.426	0.132	0.323	0.216	0.393	0.202	
$Q_{unrel.}$	**0.597**	**0.893**	8.406	0.106	0.284	0.201	0.374	**0.219**	
$Q_{reason.}$	0.559	0.880	13.002	**0.214**	0.429	**0.311**	**0.565**	0.168	
keys in context: research idea, methodology									
Q_{base}	0.587	0.878	6.231	**0.236**	**0.497**	0.296	0.460	0.177	0.1
$Q_{prox.}$	**0.599**	0.881	4.733	0.123	0.282	0.198	0.343	0.203	
$Q_{unrel.}$	**0.601**	**0.892**	4.180	0.095	0.272	0.194	0.354	**0.222**	
$Q_{reason.}$	0.565	0.885	13.726	**0.236**	0.490	**0.334**	**0.573**	0.167	
Q_{base}	0.593	0.878	10.443	0.201	0.438	0.291	0.468	0.179	0.5
$Q_{prox.}$	**0.597**	0.881	9.348	0.120	0.310	0.212	0.379	0.202	
$Q_{unrel.}$	**0.597**	**0.893**	7.646	0.094	0.253	0.204	0.366	**0.220**	
$Q_{reason.}$	0.572	0.882	13.083	**0.247**	**0.469**	**0.322**	**0.577**	0.167	

Influence of Rejected Answer Sampling Strategy. Models trained with unrelated rejected samples consistently outperform those trained with adjacent rejected samples in semantic similarity and topic consistency. This suggests that more diverse chosen-rejected pairs guide the model more effectively toward the ground truth. However, this also reduces output diversity and weakens alignment with intended limitations, indicating a shift toward topical relevance over the specific methodological or conceptual gaps emphasized in the reference context.

The Impact of Structured Summary Content. Across aligned models, differences in *LA* and *FDA* are small, suggesting that fine-tuning mainly

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_35

increases semantic similarity to the ground truth rather than explicit use of limitations or future work. In contrast, restricting the context to only research ideas and methodology consistently reduces output diversity, indicating that access to limitations and future directions supports more exploratory ideation. Reasoning models tend to utilize contextual information more effectively, suggesting that their fine-tuning is a promising direction for exploiting structured summaries.

Influence of Decoding Temperature. Higher decoding temperature increases inter-generation dissimilarity, indicating greater output diversity. This increase does not reduce topic relevance or semantic similarity, which remain relatively stable across decoding conditions. Therefore, higher temperature provides a viable mechanism for encouraging idea variation without sacrificing quality. Reasoning models show lower sensitivity to decoding temperature without alignment, suggesting that their diversity is less driven by sampling effects.

LA and FDA metrics. These metrics are demanding yet informative measures of conceptual grounding and relevance. Ground-truth ideas achieve modest scores, which is expected, as citation subgraphs capture diverse problems and multi-hop relations, while individual contributions derive from a subset of closely related works. Even so, LA and FDA provide a more discriminative signal than surface-level semantic metrics. Reasoning models achieve higher scores, indicating a stronger tendency to ground their ideas in limitations and future directions of the literature. This suggests that they leverage background knowledge more explicitly, though at the cost of dissimilarity, reflecting broader exploration rather than tight semantic alignment.

Our results highlight both the promise and the limits of alignment for scientific ideation. Fine-tuning improves consistency and semantic alignment, but reduces diversity and weakens grounding in limitations and future directions, indicating that semantic alignment alone is insufficient to capture deeper reasoning and creativity. Citation graphs and extracted key summaries provide rich context within practical length constraints, while access to limitations and future directions supports more diverse and exploratory ideation. Reasoning models emerge as a promising direction for balancing broader exploration with stronger grounding in prior work. Future work may benefit from citation-graph based reasoning, fine-tuning reasoning models and designing objectives that jointly optimize relevance and explanatory depth.

Acknowledgements Financed by: CLARIN-PL project financed as part of the investment: "CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (2024-2026), funded by the Polish Ministry of Science and Higher Education (2024/WK/01).

Disclosure of Interests The authors declare that they have no competing interests.

References

[1] Buehler, M.J.: Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning. Mach. Learn.: Sci. Technol. **5** (2024)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_35

[2] Chen, N., Li, Y., Tang, J., Li, J.: GraphWiz: An instruction-following language model for graph computational problems. In: Proc. of KDD. pp. 353–364 (2024)
[3] Cicchetti, D.V., Feinstein, A.R.: High agreement but low kappa: Ii. resolving the paradoxes. Journal of Clinical Epidemiology **43**(6), 551–558 (1990)
[4] Clement, C.B., Bierbaum, M., O'Keeffe, K.P., Alemi, A.A.: On the use of ArXiv as a dataset (2019), arXiv:1905.00075
[5] Dasgupta, D., Mondal, A., Chakrabarti, P.P.: Empowering AI as autonomous researchers: Evaluating LLMs in generating novel research ideas through automated metrics. In: AI for Research and Scalable, Efficient Systems. pp. 108–141 (2025)
[6] Gao, X., Zhang, Z., Xie, M., Liu, T., Fu, Y.: Graph of AI ideas: Leveraging knowledge graphs and LLMs for AI research idea generation (2025), arXiv:2503.08549
[7] Hu, X., Fu, H., Wang, J., Wang, Y., Li, Z., Xu, R., Lu, Y., Jin, Y., Pan, L., Lan, Z.: Nova: An iterative planning and search approach to enhance novelty and diversity of LLM generated ideas (2024), arXiv:2410.14255
[8] Hu, X., Liu, G., Zhao, Y., Zhang, H.: De novo drug design using reinforcement learning with multiple GPT agents. In: Proc. of the 37th NeurIPS (2023)
[9] Li, L., Xu, W., Guo, J., Zhao, R., Li, X., Yuan, Y., Zhang, B., Jiang, Y., Xin, Y., Dang, R., Zhao, D., Rong, Y., Feng, T., Bing, L.: Chain of ideas: Revolutionizing research via novel idea development with LLM agents (2024), arXiv:2410.13185
[10] Li, R., Jing, L., Han, C., Zhou, J., Du, X.: LDC: Learning to generate research idea with dynamic control (2025), arXiv:2412.14626
[11] Lin, E., Peng, Z., Fang, Y.: Evaluating and enhancing Large Language Models for novelty assessment in scholarly publications. In: AISD 2025. pp. 46–57 (2025)
[12] Liu, S., Cao, J., Yang, R., Wen, Z.: Generating a structured summary of numerous academic papers: Dataset and method. In: Proc. of IJCAI. pp. 4259–4265 (2022)
[13] O'Neill, C., Ghosal, T., Răileanu, R., Walmsley, M., Bui, T., Schawinski, K., Ciucă, I.: Sparks of science: Hypothesis generation using structured paper data (2025), arXiv:2504.12976
[14] Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Tech. Rep. 1999-66, Stanford InfoLab (1999)
[15] Qi, B., Zhang, K., Li, H., Tian, K., Zeng, S., Chen, Z.R., Zhou, B.: Large Language Models are zero shot hypothesis proposers (2023), arXiv:2311.05965
[16] Qi, B., Zhang, K., Tian, K., Li, H., Chen, Z.R., Zeng, S., Hua, E., Hu, J., Zhou, B.: Large Language Models as biomedical hypothesis generators: A comprehensive evaluation (2024), arXiv:2407.08940
[17] Qiu, Y., Zhang, H., Xu, Z., Li, M., Song, D., Wang, Z., Zhang, K.: Ai idea bench 2025: Ai research idea generation benchmark (2025), arXiv:2504.14191
[18] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct preference optimization: your language model is secretly a reward model. In: Proc. of NeurIPS 2023 (2023)
[19] Si, C., Yang, D., Hashimoto, T.: Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. In: ICLR 2025 (2025)
[20] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: Extraction and mining of academic social networks. In: ACM SIGKDD 2008. pp. 990–998 (2008)
[21] Wang, Q., Downey, D., Ji, H., Hope, T.: SciMON: Scientific inspiration machines optimized for novelty. In: Proc. of the ACL 2024. pp. 279–299 (2024)
[22] Wang, W., Gu, L., Zhang, L., Luo, Y., Dai, Y., Shen, C., Xie, L., Lin, B., He, X., Ye, J.: SciPIP: An LLM-based scientific paper idea proposer (2025), arXiv:2410.23166

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_35