The Ramón Llull's Thinking Machine for Automated Ideation

Xinran Zhao[1] **Boyuan Zheng**[†2] **Chenglei Si**[†3] **Haofei Yu**[†4] **Ken Ziyu Liu**[†3]
Runlong Zhou[†6] **Ruochen Li**[†5] **Tong Chen**[†6] **Xiang Li**[†4] **Yiming Zhang**[†1]
Tongshuang Wu[1] *
[1]CMU, [2]OSU, [3]Stanford, [4]UIUC, [5]UT Dallas [6]UW

Abstract

This paper revisits Ramón Llull's *Ars combinatoria*—a medieval framework for generating knowledge through symbolic recombination—as a conceptual foundation for building a modern Llull's "*thinking machine*" for research ideation. Our approach defines three compositional axes: Theme (*e.g.*, efficiency, adaptivity), Domain (*e.g.*, question answering, machine translation), and Method (*e.g.*, adversarial training, linear attention). These elements represent high-level abstractions common in scientific work—motivations, problem settings, and technical approaches—and serve as building blocks for LLM-driven exploration. We mine elements from human experts or conference papers and show that prompting LLMs with curated combinations produces research ideas that are diverse, relevant, and grounded in current literature. This modern thinking machine offers a lightweight, interpretable tool for augmenting scientific creativity and suggests a path toward collaborative ideation between humans and AI.

1 Introduction

There is a growing interest in the machine learning community in leveraging large language models (LLMs) to accelerate scientific discovery (Si et al., 2024; AI4Science & Quantum, 2023; Collins et al., 2024; Singh et al., 2025; Jansen et al., 2025; Si et al., 2025). Among these prominent directions, one challenging topic is to use LLMs to conduct or assist the *ideation* process. Despite recent success in ideation with state-of-the-art language models (Si et al., 2024), community simulation (Yu et al., 2024), and reinforcement learning (Li et al., 2024), model-generated ideas can lack diversity (Si et al., 2024). In response, in this paper, we ask: *Does conditioning on explicit concept combinations help build a minimalist pipeline for diverse and grounded research ideas?*

An ideal pipeline shall be simple, scalable, and it can generate a diverse set of ideas. In this work, we propose to create such a pipeline through revisiting one of the first human explorations of artificial intelligence invented at the end of the thirteenth century, which aims at creating new knowledge from logical combinations of concepts (Borges, 1937). Llull's machine includes multiple rotary disks of concepts, *e.g.*, goodness, power, glory, etc, where Llull believed studying all combinations of the elementary concepts would help understand a field of knowledge that can be covered by them. In light of the thinking, we revisit the idea of element combination to create a modern version for LLM ideation.

Specifically, we design three *disks* of elements *theme*, *domain*, and *method*. Corresponding research ideas are then synthesized with a finite set of rules combining all elements[1]. For example, with *less is more* as a *theme*, *confidence calibration* as a *domain*, *Mamba* (Gu & Dao, 2024) as a *method*, and a simplest *A+B+C* template, after rewriting the raw idea with Claude 3.7 (Anthropic, 2024), a candidate idea can be: *Less Parameters, Better Calibration: Confidence-Aware Training for Mamba Architectures*. We conduct a pilot study validating the pipeline with

* † denotes equal contribution in alphabetic order. Corresponding contact email addresses: {xinranz3,sherryw}@andrew.cmu.edu.

[1]Such a categorization is not exhaustive. We discuss this in the limitations section in the appendix.

human-written elements and then scale the ideation with elements mined automatically from top-tier conferences, *e.g.*, ICLR, ACL, etc.

To study the characteristics of the ideas, we first compare the statistics and elements (themes, domains, and methods) extracted from different conferences across years, which sheds light on the taste and preferences of different machine learning communities, *e.g.*, from the same number of papers, our pipeline extracts similar numbers of domain elements from ACL and more method elements from ICLR. Next, with the raw ideas combined through automatically extracted templates in the same pipeline, we further use LLMs to rewrite them into research ideas. We compare these output ideas from Ramón Llull's Thinking Machine with idea titles from previous work (Si et al., 2024; Yu et al., 2024), which suggests good diversity and coverage of the ideas generated from our minimalist method[2].

In this paper, we explore the potential of LLM ideation through reconstructing the thirteenth-century Ramón Llull's thinking machine with modern data mining and automatic evaluation techniques. We anticipate the proposed pipeline and resources to serve as (1) a simple but strong baseline for LLM ideation; (2) an interesting view that motivates human researchers to find or review their ideas. The authors acknowledge the core contribution of our work as an investigation into quantifying how much research ideation can be mechanically automated. We will open-source our code, data, and generated research ideas at `https://github.com/colinzhaoust/ramon_llull_public`.

2 Related Work

Symbolic Reasoning Ramón Llull's thinking machine (Borges, 1937) is one of the earliest attempts at formalizing reasoning, laying the foundation for symbolic AI. It motivates later developments such as mathematized logic (Uckelman, 2010), the universal Turing machine (Turing, 1936), ontologies (Goerss, 2024) and knowledge graphs (Ji et al., 2021). While it shares with knowledge graphs the goal of representing structured information, the key difference lies in their operational principles. Knowledge graphs capture large-scale relational structures among extracted entities. In contrast, Llull's thinking machine starts with a small and fixed set of core concepts and systematically explores their combinatorial possibilities using a rotating mechanism. This generative, combinatorial focus distinguishes it from the more static and structural nature of knowledge graphs.

Automatic Ideation Recent advancements have explored the use of LLMs to automate and enhance scientific and creative ideation. The most direct approach involves prompting LLMs to generate ideas in a single pass (Si et al., 2024). Building on this, other works incorporate more structured techniques such as iterative boosting (Wang et al., 2024), knowledge augmentation (Baek et al., 2025), multi-agent collaboration (Yu et al., 2024), reinforcement learning (Li et al., 2024), to refine ideation quality. A further step involves analogical reasoning (Hope et al., 2017), which mines high-quality ideas from structured knowledge by drawing connections between similar concepts. Our approach moves one step beyond analogy: we identify high-quality core concepts and systematically explore their combinatorial space—inspired by Llull's thinking machine—to generate novel and diverse ideas grounded in specific research communities. We further discuss related work on data mining from academic papers in Appendix A.2.

3 The Ramón Llull's Thinking Machine

From the historical context, Ramón Llull designed the machine. to provide answers to arbitrary questions with a combination of elements selected through spinning three concentric and revolving wood or metal disks[3]. Through patient manipulation of the multiplication and elimination, the machine will eventually produce a seemingly good answer.

[2]The authors note that the diversity and coverage do not necessarily suggest the novelty and utility of the ideas, which require extensive human experimentation and evaluation to validate.

[3]We present a figurative illustration in Appendix 3

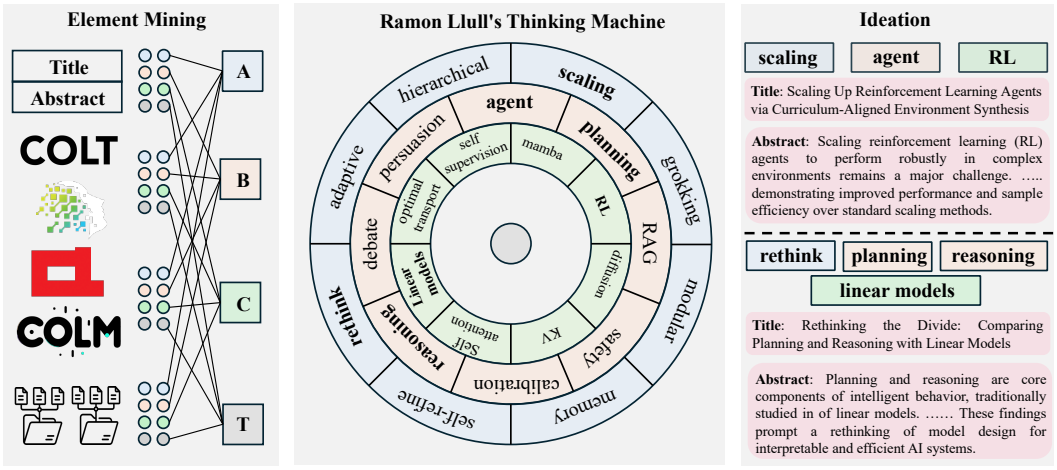

Figure 1: The overall pipeline of using the concept of Ramón Llull's thinking machine for research ideation. It includes three main steps (1) *element mining*: mining and merge elements (like keywords representing themes, domains, and methods) from papers in top conferences; (2) *combinational thinking*: combining extracted elements through symbolic recombination similar with Ramón Llull's thinking machine; (3) *idea generation*: generating abstract-style research ideas based on templates and elements.

We consider this process a simulation of one kind of human ideation process: a researcher may see a good paper and decide to apply it to their own domains, *e.g.*, the recent success in introducing teacher forcing to diffusion (Chen et al., 2024)[4]. There is no guarantee on if this ideation process is the best in terms of novelty, but it shall be considered as a common practice in various communities.

In a formal way, given three disks of elements $A = \{a_1, a_2, ...\}, B = \{b_1, b_2, ...\}, C = \{c_1, c_2, ...\}$ and a template T [5], Ramón Llull's Thinking Machine ϕ outputs the raw idea $x = \phi(A, B, C, T)$, where T can require ≥ 1 elements from each disk. Then, given a large language model m, following the setup of (Si et al., 2024), we denote the ideation as the generation of a title and a corresponding abstract, $(t, abs) \sim m(x)$. We show our pipeline of element mining and ideation in Figure 1, with details in the following sections. At this stage, we leave sampling execution plans from the raw idea to future work.

3.1 Building the Idea Generator

Elements. Similar to the original thinking machine, we design three disks to capture the minimum description of the ideation context:

- Theme (A): the theme of the work, which highlights a particular scene or purpose to conduct the study, *e.g.*, *less is more*, *few-shot*, *adaptive*, *aggregation*, *in-the-wild*, *is all you need*, etc. The theme elements might be revisited with different names in the literature, where "trendy" themes can also be different across communities.
- Domain (B): the domain of the work, which indicates a potential set of tasks to solve and previous work to follow, *e.g.*, *argument mining*, *question answering*, etc.
- Method (C): the method of the work, which shows how certain problems are addressed with specific adoption or adaptation of a model, data, or training framework, *e.g.*, *transformer*, *state-space models*, *preference optimization*, etc.

[4]We also note other processes, *e.g.*, see an abnormal phenomenon (Goyal et al., 2025), answer a question of the community (Wu et al., 2024), and push to the extreme condition (Shao et al., 2024).

[5]For example, $a + b + c$ or compare c_1 and c_2 in b_1 under a_1

Community	A (Theme)	B (Domain)	C (Method)
NLP	adaptive, less is more, hierarchical, in-the-wild, self-refine, hindsight, rethink, grokking, long-tail, compositional, multi-hop	agent, planning, retrieval, safety, calibration, reasoning, memorization, persuasion, debate	Mamba, RL, Linear Models, KV Cache, Quantization, Diffusion, Self-attention, Self-supervision
CV	test-time Training, meta learning, active learning, open-set calibration, open-vocab grounding, continual learning	image classification, detection, segmentation, optical flow estimation, action recognition	ViT, NeRF, ConvNext, point-transformer, Perceiver, Instant-NGP, Yolo, UNet, LoRA
RL Theory	parametric policy optimization, online learning, offline learning, adversarial, corruption, linear policy, general function approximation	multi-armed / contextual bandits, Markov decision processes, Markov games, stochastic shortest path	ε-greedy, Thompson sampling, upper confidence bound, optimism, pessimism

Table 1: Lists of Theme (A), Domain (B), and Method (C) written by researchers from different communities. NLP, CV, and RL Theory denote natural language processing, computer vision, and reinforcement learning theory, respectively. We present the full table in Appendix A.3.

Stats.	ICLR 24	COLM 24	COLT 24	ACL 24	ACL 23	ACL 22	All
# Papers	2000	299	170	1931	2052	1031	7483
# Theme (A)	391	118	91	307	359	224	682
# Domain (B)	330	87	62	300	272	208	633
# Method (C)	392	35	53	54	117	153	866
#Template (T)	278	71	75	121	277	165	925

Table 2: Statistics of the papers processed, themes, domains, method elements, and templates mined from various top-tier conferences. *All* denotes the cumulative elements after merging and filtering. # X denotes the number of X.

We select the current disks to form the minimum description of a research idea: we did a in b with c, as described in the IMRaD format of academic writing (Nair & Nair, 2014). However, elements in the disks can be non-exclusive, *e.g.*, *retrieval* can be considered as a domain with various tasks, as well as a set of methods for other domains. We discuss other potential axes of the machine in Section 4.4. We further discuss the relations of the intra/inter-disk elements in Appendix A.1.

Ideation. With the disks in hand, to capture the diverse relations and combinations among elements, we extend the original Ramón Llull's Thinking Machine with templates for generation, which is widely used in the knowledge graph construction (Zhang et al., 2020). A template serves as a way to combine the elements, with potential additional descriptive words on their relations. Besides the aforementioned "we did a in b with c", other rudimentary templates can be "compare c_1 and c_2 in b_1 under a_1" or "c_1 is all you need".

3.2 Mining the elements and templates

Pilot: Human Annotation. To validate our design of the disks, we first seek elements of the disks from PhD students from different communities[6]. Table 1 presents the theme, domain, and method elements written by human experts. With LLM rewriting, these elements can already lead to interesting research ideas, *e.g.*, from *hindsight*, *debate*, *RL*, Claude 3.7

[6]Each volunteer has published 5+ papers in the conferences of the corresponding community.

can output a title: *Learning to Argue in Hindsight: Multi-Agent Debate with Retrospective Reinforcement Learning* with a reasonable abstract.

Among different communities, there are similarities, *e.g.*, *transformer* and *diffusion*, and differences, *e.g.*, *multi-hop* vs. *inverse rendering*. In recent years, certain ideas from one community have motivated the novel directions in other communities, *e.g.*, transformer for vision (Dosovitskiy et al., 2021) and diffusion for text (Li et al., 2022), and vice versa, which inspires us to study and auto-extraction of these elements from conferences acknowledged in different communities.

Mining from the Literature. To automate and scale the element harvest, we propose to automatically mine the elements and templates from different top-tier conferences, which allows for extendability to our pipeline. Specifically, we use Gemini 2.0 Flash (Gemini Team et al., 2024) to process each paper title and abstract into lists of A, B, C, and a template with a carefully designed element extraction prompt. Upon acquiring the elements from different papers, since we observe duplications among the elements, we then leverage Gemini 2.0 Flash again to merge the elements based on the semantic similarities. The detailed prompts are presented in Appendix A.6.

We collected the papers from Paper Copilot (Yang, 2025), with ICLR 24, COLM 24, COLT 24, ACL 24, ACL 23, and ACL 22 as the selected conferences to cover a diverse set of topics. We randomly sampled 2,000 papers for ICLR 24. Table 2 presents the statistics of the acquired elements. In total, we collect more than 600 elements for each category from the analysis of 7,483 papers. From the same pipeline, for ACL, the number of method elements that can be extracted from Gemini decreases over the years, *e.g.*, *noise sensitivity*, *multi-criteria optimization*, and *early exit* that appear in ACL 23 no longer appear in ACL 22. With a similar number of papers analyzed, with a similar number of domain elements, ICLR 24 also covers more themes and method elements compared to ACL 2024. Example elements for other conferences are in the appendix (Table 6)

We note that these elements are merged from the raw elements processed from the papers, which can potentially lead to more elements at a finer granular view, *e.g.*, the element *generalization* is merged from *generalizability*, *domain generalizability*, and *temporal generalization*, which can lead to subtle but crucial changes in the paper story and experiment design. We will open-source the finer-granular elements as well as their visited counts.

3.3 Discussion

With all the elements in hand, we can then generate the raw ideas by combining them with the templates. We propose two uses for the resources: (1) randomly sample a template, *e.g.*, *Compare* c_1 *and* c_2 *in* b_1 *under* a_1, and randomly fill in the elements; (2) enumerate the top-visited elements and templates to construct a diverse set of raw ideas to fuel the studies on downstream execution or quality evaluation. We further study the characteristics of the generated ideas before and after LLM rewriting in Section 4.

In our current design, we treat each equally in sampling after ranking with their visit counts from the papers. We also note that the statistical features, such as the popularity of an element or selectional preference (Zhang et al., 2019) among elements, can potentially suggest a better sampling process for the raw ideas or disk categorization. For example, we can build a Viterbi-like sampling process considering the selectional preference as the transitive scores. On the other hand, a fine-grained element sampling process can lead to a controllable ideation process, *e.g.*, sampling the frequent elements can potentially increase the relevance to specific conferences, while sampling randomly can potentially increase the diversity of the ideas, which leads to a trade-off between the relevance and diversity.

At the current stage, we build the pipeline with pre-defined disk types and leave the extension of the data mining and ideation pipeline for future work.

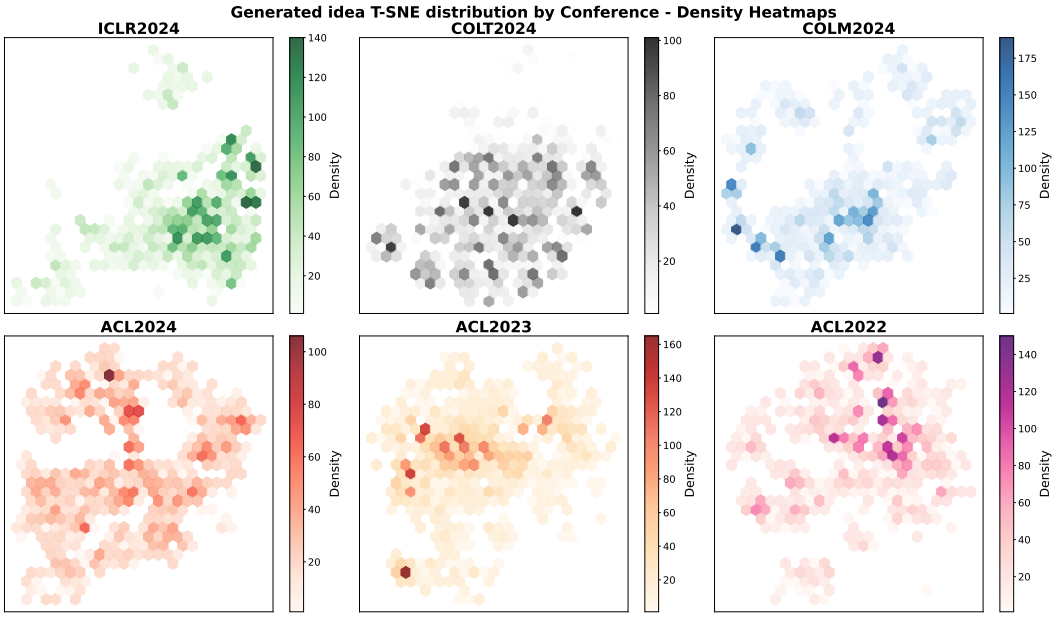

Figure 2: The density heatmap visualization of ideas generated from the basic *A*, *B*, *C* template with top-20 most visited elements from each disk for each conference. All sub-figures are aligned in the same distribution with t-SNE. We can observe different relations among the ideas from specific conferences, e.g., taking up different parts of the space.

Conference	Top A (Theme)	Top B (Domain)	Top C (Method)
ACL 24	in-context reasoning, in-the-wild, zero-shot, alignment, benchmarking	reasoning, question answering, calibration, safety, machine translation, natural language inference	LLMs, transformers, Self-attention, LoRA, retrieval-augmented generation
ICLR 24	generalization, efficiency, robustness, scalability, self-supervised	reasoning, federated learning, safety, reinforcement learning, planning	LLMs, deep learning, transformers, diffusion, vision-language models

Table 3: Qualitative comparison of the most frequent extracted elements from different conferences. We can observe both shared and different keywords.

4 Experiment and Analysis

In this section, we take a closer look at the characteristics of the raw ideas from the template combination of the elements from the perspective of (1) differences across conferences; (2) comparison with the generated research ideas from previous work.

4.1 Differences across conferences

We first compare ideas from different conferences with the extracted element. To avoid the noise from redundant words in the templates, we use the basic *A*, *B*, *C* template with the top 20 most visited elements from each disk to generate the raw research ideas. For each conference, we will have 4,000 raw research ideas. Then we convert the research ideas to TF-IDF vectors and apply t-SNE for the visualization.

As presented in Figure 2, we can observe different relations among the conferences: (1) COLT 2024 ideas (green dots) are comparatively standalone, with limited coverage with

Ideation Methods	# Ideas	# Words	Diversity	Similarity	Relevance
Si et al. (2024)	93	1,063	0.29	0.22	0.28
Yu et al. (2024)	100	2,379	0.29	0.19	0.18
Ramón Llull (Top)	100	1,014	0.21	0.26	0.11
Ramón Llull (Random)	100	1,105	0.41	0.23	0.05

Table 4: Statistics and metric results of different automatic ideation methods. The computation of the similarity and relevance uses ACL 2025 main paper titles as references.

other conferences. ICLR 2024 ideas (blue dots) have an overlap with ACL 2024 ideas, but still have a standalone area of clusters; (2) COLM 2024 ideas (red cross) lie in the intersection of ACL 2024, ICLR 2024, and COLT 2024, which shows the joint interests of language modeling from different communities, as the full name of COLM is Conference on language modeling; (3) As a sanity check, ACL 2024 and ACL 2023 ideas are largely overlapping, although shifts in interests still can lead to standalone clusters. We present an extended study on the differences of elements of ACL over the years in Appendix A.4.

We further qualitatively compare the elements for ACL 2024 and ICLR 2024 in Table 3, which indicates the causes of the geometrical relations in the t-SNE visualizations: how researchers submit to different conferences show shared (*e.g.*, LLMs and Transformers) and divergent interests (natural language inference vs. federated learning).

4.2 Comparing different ideation methods

With the elements in hand, we can then generate the research ideas with LLMs rewriting. Specifically, we use Gemini-1.5 Pro to rewrite the sampled combination, for example, with element *emergent*, *theory of mind*, and *variational inference*, the rewritten generated idea is: *Emergent Theory of Mind in Disentangled Latent Spaces via Variational Inference*. We consider two variants of our thinking machine based on the sampling methods: (1) Ramón Llull (Top): we select the most visited elements from the disks and enumerate all the combinations; (2) Ramón Llull (Random): we randomly sample elements from all disks and ensure that each element only appear once at most.

We compare these rewritten ideas from previous work on ideation: (1) Si et al. (2024) carefully sample and filter AI-generated ideas and list 93 high-quality ideas on 7 NLP topics, including Bias, Coding, Safety, Multilingual, Factuality, Math, and Uncertainty. These ideas are then used for novelty evaluation (Si et al., 2024) and execution study (Si et al., 2025); (2) Yu et al. (2024) simulate the diverse discussion in the research community and generate ideas in a question-and-answer format. To allow fair comparison. We select 100 ideas from Yu et al. (2024) from the batch where discussions are based on certain previous papers. To allow fair comparison, we select 100 ideas from each of our thinking machine variants with elements extracted from ACL 2024[7]. We only compare the sampled research idea titles for all the methods.

We consider various metrics to compare the ideation methods, including diversity, similarity, and relevance to certain conference papers.

For diversity, we follow (Li et al., 2015) to use *distinct-1* as a metric, the number of distinct unigram count normalized by the total number of words to capture the diversity of concise titles. For future work involving abstract or sections, the metric of diversity can be extended to entropy-based metrics as described in Zhang et al. (2025).

For relevance and similarity, we consider using paper titles from the main track accepted papers from ACL 2025 to ground the comparison, where the accepted papers are released

[7]The authors note that the comparative study is mainly set up to compare the characteristics of these ideation methods. Since the previous methods are designed to sample ideas for their own purposes: novelty evaluation (Si et al., 2024) and research community simulation (Yu et al., 2024). The results from our metrics do not suggest the superior quality of any method. Rigorous idea quality evaluation may involve extensive expert annotation and insights from execution (Si et al., 2025).

Metric	ACL 2024	ACL 2023	ACL 2022	ICLR 2024	COLM 2024	COLT 2024	Overall
Decomp. %	99.5%	99.2%	99.2%	99.6%	100.0%	100.0%	**99.5%**
Recon. %	17.6%	17.8%	19.9%	15.6%	11.0%	8.2%	**16.4%**

Table 5: Bijective coverage results across conferences.

after June 2025. To reduce the chance of LLMs seeing the paper titles in their training data (Gemini 1.5 Pro was released Feb 2024). Specifically, for relevance, we compute the average BLEU score (Papineni et al., 2002) between each generated idea and each conference paper title pair to measure how likely a generated title is relevant to the conference. For similarity, we measure how likely a research paper title in ACL 2025 is similar to a generated idea. Similar to our experiments in Appendix A.4, we use token-level Jaccard similarity to capture the similarity of a pair of titles. We report the similarity as the average across ACL 2025 paper titles that scored the top-K highest similarities, where K equals the number of model-generated ideas.

Table 4 presents the different characteristics of different automated ideation methods: Ramón Llull (Top) that enumerates combinations of the most trending elements achieved the highest similarity and Ramón Llull (Random) that samples random elements achieve the highest diversity, with a decrease in relevance - which demonstrates a trade-off between diversity and similarity/relevance. Our Ramón Llull thinking machines also show lower relevance compared to human filter ideas (Si et al., 2024) or ideas from simulated discussions grounded on certain papers (Yu et al., 2024). One potential reason can be that although the random sampling of elements leads to a diverse set of ideas, they are not necessarily the direction of research acknowledged by the community. Future fine-grained sampling identifying the relations among the elements can be a future direction to improve the ideation process of our Ramón Llull thinking machine.

4.3 Analysis: How much of research ideation is combinatorial?

In this section, we test to what degree the ideation of machine learning research can be explained by our proposed Ramón Llull system. To this end, we conduct a *coverage analysis*. This evaluation tests two complementary aspects:

- **Decomposition:** Given a research paper title, can it be decomposed into constituent A, B, C elements from our extracted disks? We consider a research idea decomposable if Gemini 2.0 Flash successfully converts the paper title into theme, domain and method elements that our method already extracted.
- **Reconstruction:** Given the theme, domain and method elements alone, can they be combined to approximately reconstruct the title of the original paper? We consider the research idea reconstructible if Gemini 2.0 can propose a title that is highly similar ($\geq$ 30% Jaccard similarity) given the extracted keywords.[8]

Table 5 presents our bijective coverage results across 7,483 papers from six major conferences. We find **near-universal decomposability (99.5%):** almost all research papers can be decomposed into our A+B+C framework. This validates that the three-disk design captures fundamental aspects of research ideation across machine learning communities.

On the other hand, **reconstructibility is limited (16.4%)**. That said, a non-trivial fraction of real paper ideas can already be faithfully reconstructed by combining ideas in our proposed Ramón Llull framework. While the framework successfully captures structural building blocks that are nearly universal across machine learning research, we note that the specific instantiation of a research idea may still require creativity and insights that go beyond mere combination of past ideas, and researchers' prior and taste may remain essential to effectively navigate this combinatorial space.

[8]Additional evaluation details (e.g., prompts) are reported in A.8.

4.4 Analysis: What is not covered by A+B+C?

Beyond the disk view of automated ideation, we identify other dimensions of the problem as follows, to serve as potential motivations for the community:

- **Perturbation.** Given the same set of A, B, C, the final paper can be drastically different. Akin to the trivial and non-trivial perturbation ($x^2 \rightarrow x^4$vs.$x^2 \rightarrow x^{-1}$) discussed in Huang et al. (2025), certain perturbations can potentially change the problem fundamentally. Studying pairs of ideas that have identical elements discovered in our pipeline can potentially allow a fine-grained study on the sparks of non-trivial perturbations that largely reshape the problem.
- **The 4th Axis.** In Section 3, we build our ideation pipeline with Theme (A), Domain (B), and Method (C) as the disks. Another axis of the machine can be "algorithms" vs "analysis". Axis C currently says that the research idea should *use* the specified method, but a method can be both *used* and *studied/analyzed*. With this fourth disk, the machine can cover further analysis-based work, such as adversarial examples (Szegedy et al., 2013) and edge-of-stability (Cohen et al., 2021). Such analysis work often leads to the discovery of new/revived phenomena, such as Agreement-on-the-line (Baek et al., 2022), Grokking (Power et al., 2022), and Model Collapse (Shumailov et al., 2024).
- **Negation.** Another dimension of non-A+B+C ideas is the negation of commonly believed A+B+C, which often leads to wide community discussion, rethinking of the directions, and improved evaluation, such as the mirage of a phenomenon (Schaeffer et al., 2023), the gap between automatic and human evaluation (Durmus et al., 2022; Gehrmann et al., 2022), and the misuse of certain tools (Grusky, 2023)

5 Discussion and Future Work

In this paper, we propose to create a modern Ramón Llull thinking machine for automated ideation, which serves as a lightweight and interpretable tool to create a diverse set of LLM-generated ideas, as well as a perspective to study the commonality and differences in the human ideation process across different communities. We discuss the intended usage and future work as follows:

Intended Usage. The proposed Ramón Llull thinking machine is **NOT** intended to (1) conduct DDOS (Distributed Denial of Service) on the current brittle reviewing system (Kim et al., 2025); (2) evaluate or attack certain human-generated ideas. Our Ramón Llull's Thinking Machine is intended to serve as (1) a baseline for future study on ideation with a filtered set of components, i.e., theme, topics, and domains; (2) motivation for human researchers to track the field status quo and their own ideas. We plan to open-source 1,000 high-quality human-filtered ideas to conduct a stealth human study on the execution of the ideas in a human-AI collaborative manner with real research labs.

Future Work. We expect to extend our pipeline through (1) evaluating the quality of the generated ideas; (2) studying the human ideation and polishing process through the lens of Ramón Llull's Thinking Machine; (3) studying how the execution process can serve as elements or factors of sampling.

Acknowledgments

The authors thank Hongming Zhang, Sihao Chen, Zhiyuan Zeng, Wenhao Yang, Fengyu Cai, and Ben Zhou, as well as other fellow AI2 interns (including but not limited to Hita, Yapei, Fede, Anej, Amanda, Michael, Nishant, Peiling, Alexiss, and etc) and UW students, for their insights into design and evaluation choices. The authors also thank the constructive discussions with colleagues from CMU WInE Lab. Xinran Zhao is supported by the ONR Award N000142312840. This work is supported by the OpenAI Research Credit program, the Amazon AI Research Gift Fund, and the Gemma Academic Program GCP Credit Award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

Ethics Statement

We foresee no ethical concerns or potential risks in our work. All of the datasets are open-sourced and from peer-reviewed research papers, as shown in Section 3.2. The LLMs we applied in the experiments are also publicly available. Given our context, the outputs of LLMs are unlikely to contain harmful and dangerous information. The experiments in our paper are mainly on English.

References

Microsoft Research AI4Science and Microsoft Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *ArXiv*, abs/2311.07361, 2023. URL https://api.semanticscholar.org/CorpusID:265150648.

AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1:1, 2024.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=EZZsnke1kt.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *NAACL 2025*, 2025.

Jorge Luis Borges. Ramon llull's thinking machine. 1937. URL https://gwern.net/doc/borges/1937-borges-raymondllullsthinkingmachine.pdf. Accessed June 20, 2025.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=yDo1ynArjj.

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

Katherine M Collins, Albert Q Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B Tenenbaum, William Hart, et al. Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, 121(24):e2318124121, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Esin Durmus, Faisal Ladhak, and Tatsunori Hashimoto. Spurious correlations in reference-free evaluation of text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1443–1454, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.102. URL https://aclanthology.org/2022.acl-long.102/.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text, 2022. URL https://arxiv.org/abs/2202.06935.

Gemini Team et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv*, 2024. URL https://arxiv.org/abs/2403.05530.

Eleanor Goerss. The mirror and the knot: The soul's recursive action in early lullian figures. *Res: Anthropology and Aesthetics*, 81(1):61–77, 2024.

Sachin Goyal, Christina Baek, J Zico Kolter, and Aditi Raghunathan. Context-parametric inversion: Why instruction finetuning may not actually improve context reliance. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SPS6HzVzyt.

Max Grusky. Rogue scores. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1914–1934, 2023.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 235–243, 2017.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, Yue Wu, Ming Yin, Shange Tang, Yangsibo Huang, Chi Jin, Xinyun Chen, Chiyuan Zhang, and Mengdi Wang. MATH-Perturb: Benchmarking LLMs' math reasoning abilities against hard perturbations. *arXiv preprint arXiv:2502.06453*, 2025.

Peter Jansen, Oyvind Tafjord, Marissa Radensky, Pao Siangliulue, Tom Hope, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Daniel S. Weld, and Peter Clark. Codescientist: End-to-end semi-automated scientific discovery with code-based experimentation, 2025. URL https://arxiv.org/abs/2503.22708.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514, 2021.

SeongKu Kang, Yunyi Zhang, Pengcheng Jiang, Dongha Lee, Jiawei Han, and Hwanjo Yu. Taxonomy-guided semantic indexing for academic paper search. *arXiv preprint arXiv:2410.19218*, 2024.

Jaeho Kim, Yunseok Lee, and Seulki Lee. Position: The ai conference peer review crisis demands author feedback and reviewer rewards. 2025. URL https://api.semanticscholar.org/CorpusID:278394195.

Dongha Lee, Jiaming Shen, SeongKu Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. Taxocom: Topic taxonomy completion with hierarchical discovery of novel topic clusters. In *Proceedings of the ACM Web Conference 2022*, pp. 2819–2829, 2022.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

Ruochen Li, Liqiang Jing, and Xinya Du. Learning to generate research idea with dynamic control. In *2nd AI4Research Workshop: Towards a Knowledge-grounded Scientific Research Lifecycle*, 2024. URL https://openreview.net/forum?id=zCb0dPvGYN.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. Diffusion-lm improves controllable text generation. *ArXiv*, abs/2205.14217, 2022.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.

Hans-Michael Müller, Eimear E Kenny, and Paul W Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS biology*, 2 (11):e309, 2004.

Pk Nair and Vimala Nair. *Scientific Writing and Communication in Agriculture and Natural Resources*. 01 2014. ISBN 978-3-319-03100-2. doi: 10.1007/978-3-319-03101-9.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ITw9edRDlD.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei Koh. Scaling retrieval-based language models with a trillion-token datastore. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=iAkhPz7Qt3.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022): 755–759, 2024.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv*, 2024.

Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The ideation-execution gap: Execution outcomes of llm-generated versus human research ideas, 2025. URL https://arxiv.org/abs/2506.20803.

Amanpreet Singh, Joseph Chee Chang, Chloe Anastasiades, Dany Haddad, Aakanksha Naik, Amber Tanaka, Angele Zamarron, Cecile Nguyen, Jena D. Hwang, Jason Dunkleberger, Matt Latzke, Smita Rao, Jaron Lochner, Rob Evans, Rodney Kinney, Daniel S. Weld, Doug Downey, and Sergey Feldman. Ai2 scholar qa: Organized literature synthesis with attribution, 2025. URL https://arxiv.org/abs/2504.10861.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, and Tom Goldstein. An open review of openreview: A critical analysis of the machine learning conference review process, 2020. URL https://arxiv.org/abs/2010.05137.

Alan Turing. Universal turing machine. *Informatika*, 1(3073):2k, 1936.

Sara L Uckelman. Computing with concepts, computing with numbers: Llull, leibniz, and boole. In *Conference on Computability in Europe*, pp. 427–437. Springer, 2010.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific Inspiration Machines Optimized for Novelty. In *ACL*, 2024.

Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D. Goodman. A reply to makelov et al. (2023)'s "interpretability illusion" arguments, 2024. URL https://arxiv.org/abs/2401.12631.

Jing Yang. Paper copilot: The artificial intelligence and machine learning community should adopt a more transparent and regulated peer review process. *ArXiv*, abs/2502.00874, 2025. URL https://api.semanticscholar.org/CorpusID:276094819.

Haofei Yu, Zhaochen Hong, Zirui Cheng, Kunlun Zhu, Keyang Xuan, Jinwei Yao, Tao Feng, and Jiaxuan You. Researchtown: Simulator of human research community. *arXiv preprint arXiv:2412.17767*, 2024.

Hongming Zhang, Hantian Ding, and Yangqiu Song. SP-10K: A large-scale evaluation set for selectional preference acquisition. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 722–731, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1071. URL https://aclanthology.org/P19-1071/.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. ASER: A large-scale eventuality knowledge graph. In *WWW*, pp. 201–211, 2020.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. Noveltybench: Evaluating language models for humanlike diversity. *arXiv preprint arXiv:2504.05228*, 2025.

Yunyi Zhang, Ming Zhong, Siru Ouyang, Yizhu Jiao, Sizhe Zhou, Linyi Ding, and Jiawei Han. Automated mining of structured knowledge from text in the era of large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6644–6654, 2024.

A Appendix

A.1 Limitations

Evaluating idea quality and novelty. Our current evaluation is based on quantitative metrics such as diversity and community relevance. While these metrics are useful for assessing the breadth and community-alignment of the generated idea space, they could be insufficient for judging the scientific merit of individual idea under some circumstances. For instance, a generated idea that is lexically unique in terms of diversity, but could be conceptually trivial or scientifically unsound. An idea might achieve high relevance by closely mirroring existing research trends, making it plausible but potentially incremental and not truly novel. Conversely, a truly groundbreaking idea might score low on relevance because it deviates significantly from established paradigms. A more rigorous assessment requires moving beyond surface-level statistics to semantic evaluation, for which human expert judgment remains the gold standard. Experts assess ideas along critical aspects like feasibility, potential impact, and non-obviousness, providing qualitative depth that text-based metrics are not designed to capture. However, large-scale human evaluation is difficult to scale, expensive, and subject to inter-annotator variability. Securing a diverse pool of experts capable of judging ideas across the wide range of generated topics is a major logistical challenge. LLM-as-a-judge frameworks, such as the one proposed by (Li et al., 2023) trained with carefully designed rubrics, might be biased by its training data, potentially favoring well-phrased but shallow ideas over more bad-worded but conceptually deep ones. A future direction is a hybrid evaluation pipeline that leverages our quantitative metrics for initial filtering, employs LLM-as-a-judge for scalable scoring, and incorporates targeted human experts for final validation.

Organizing the elements. While our current implementation, which treats conceptual elements as independent items, has successfully generated a wide breadth of ideas, it can be further enriched by the structured relationships that exist between scientific concepts. One extension is to evolve from simple, flat lists to categorical and hierarchical structures by linking to the keywords in OpenReview or the task and method hierarchies on Papers with Code[9]. This would enable more granular control over ideation — for example, allowing sampling at different level of abstraction. Another possible direction is to explicitly model the exclusiveness of selection preferences to learn which combinations of themes, domains, and methods are most likely to be coherent. By integrating such structured knowledge, our system would transform into a more semantic-aware ideation that is capable of generating ideas that are both novel and conceptually sound.

A.2 Extended Discussion

Related Work: Data Mining from Literature Extracting structured information from academic papers is a critical research area (Zhang et al., 2024). Prior work has explored concept-level understanding through methods such as topic discovery (Lee et al., 2022) and concept matching (Kang et al., 2024), often operating over large sets of concepts using clustering or taxonomy construction. Ontology-based approaches (Müller et al., 2004) similarly aim to organize and retrieve scientific knowledge from literature at the conceptual level. In contrast, our work focuses on identifying a small set of high-quality concepts, specifically, the theme, domain, and method of each article, that are used as rotating wheels in Llull's thinking machines. This design supports combinatorial exploration and enables ideation within and across research communities.

How to view A+B+C? The same A+B+C can lead to different results and execution. For the utility and feasibility of the idea execution, it is vital to analyze why the components are complementary: *e.g.*, why a core problem in a domain requires an architectural change or can be viewed as a specific theme.

[9] paperswithcode.com

Community	A (Theme)	B (Domain)	C (Method)
ICLR 24	representation learning, offline learning, sparsity, interpretability, explainable, unsupervised learning, uncertainty, multi-modal, multi-hop, reference free, fairness, contrastive learning, sampling, heterogeneity, out-of-distribution, active learning, hierarchical structure, meta-learning	planning, safety, reinforcement learning, question answering, calibration, automated research, federated learning, classification, image generation, optimization, memorization, representation learning, segmentation	Ordinary Differential Equations, visualization tool, dynamic programming, matching function, plug-in modules, semi-supervised learning, self-training, Text-to-Image Generators, Linear Discriminant Analysis
COLM 24	in-context learning, in-the-wild, compositionality, self-evolve, long-tail, multi-hop, reference free, multi-modal, generalization, alignment, adaptation, robustness, granularity, multilingual, interpretability	inference, RAG, automated translation, decision-making, drug discovery, text generation, fine-tuning, argument mining, code editing, preference learning	Transformers, Self-attention, Mamba, RWKV, SSMs, state space models, RLHF, PPO, Mixture-of-Experts, LoRA, RNNs, VQAs, Pruning, deep generative models
ACL 24	self-evolve, generalization, domain generalization, temporal generalization, robustness, resilient, parameter-efficient, multilingual, cross-lingual, multi-task learning, cross-task, bias, debiasing, modularity, less is more, adaptive, adaptability, unsupervised adaptation	human-bot interaction, entity grounding, finance, equity research, macroeconomics, tool learning, metaphor interpretation, game playing, open-world games, style transfer, medical diagnostics	RoBERTa, BART, ByT5, diffusion models, Latent Diffusion Model, Brownian Bridge process, PLMs, Spiking Neural Network, generative models, contrastive decoding

Table 6: Lists of Theme (A), Domain (B), and Method (C) written by researchers from different communities. NLP, CV, and RL Theory denote natural language processing, computer vision, and reinforcement learning theory, respectively.

Jaccard.	ACL 22 vs. 23	ACL 23 vs. 24	ACL 22 vs. 24
# Theme (A)	0.08	0.07	0.14
# Domain (B)	0.22	0.17	0.19
# Method (C)	0.05	0.08	0.09

Table 7: Jaccard similarity of different disks for different years of ACL conferences. Compared to theme and method, there is a higher similarity of domains across years.

A.3 Mining the elements and templates (Full)

We present the full table of elements written by humans in Table 6.

A.4 Differences over years

Besides the relevance among conferences, another interesting dimension is the distribution shift over years (Tran et al., 2020). We further compare the elements of different disks through the lens of token-level Jaccard similarity for ACL from 2022 to 2024. Table 7 shows a higher similarity in the elements from domains compared to themes or methods. One potential reason is that ACL typically lists several tracks to guide the paper submission, *e.g.*, *Question Answering* and *NLP applications*. The Jaccard similarity does not change a lot across years, but the similarity is not high in general, which indicates the topical diversity in top-tier conferences. Besides the disappearance of method elements through the years,

we can also observe the occurring interests in certain elements. For the elements that appear uniquely in ACL 2024, disk A (theme) has *perspective awareness*, *Multi-generator*, etc; disk B has *Hateful Meme Detection*, *emotional support*, etc.

A.5 Example ideas from different ideation methods

This section provides examples from three distinct AI-driven ideation methodologies, each producing a different kind of conceptual output. We summarize the idea and omit some details for better presentation.

A.5.1 Example 1: AI Scientist

This method demonstrates the LLM's capacity to generate a comprehensive, structured research plan from a single core concept. The output is an actionable roadmap detailing the necessary steps to investigate an idea for next-stage experimentation.

Title: Adversarial Stereotype Dissolution Prompting: Reducing Social Biases in Large Language Models through Active Counter-Example Generation

1. Problem Statement: Large language models often generate outputs that reinforce existing stereotypes and social biases, even when attempting to be unbiased. This perpetuates harmful societal prejudices and limits the models' ability to provide fair and inclusive responses across diverse user groups.

2. Motivation: Current approaches to reducing bias in language models typically focus on avoiding or counterbalancing stereotypes... By prompting the model to generate adversarial examples that contradict stereotypes, we can encourage it to develop more nuanced and less biased representations ...

3. Proposed Method: We introduce **Adversarial Stereotype Dissolution Prompting (ASDP)**, a technique that challenges the model to actively generate counter-stereotypical examples. The prompt structure includes: ...

4. Step-by-Step Experiment Plan:

Step 1: Dataset Preparation:
Create a dataset of stereotype-sensitive queries across various domains (e.g., gender, race, age, profession), Collect 100-200 such queries for a comprehensive evaluation...

Step 2: Baseline Methods Implementation:
Implement the following baseline methods:

a) Standard prompting (direct query).
b) Disclaimer prompting (adding "Please provide an unbiased response" to queries).
c) Counterbalancing prompting (explicitly asking for examples from different groups).

Step 3: ASDP Implementation

- Implement the Adversarial Stereotype Dissolution Prompting method.
- Create a template that includes the four steps mentioned in the proposed method.
- Ensure the prompt is clear and consistent across different queries.

Step 4: Model Selection Use GPT-4 and GPT-3.5-turbo from OpenAI's API for the experiments. These models are state-of-the-art and widely used, making the results relevant and comparable.

Step 5: Experiment Execution: For each query in the dataset:
a). Generate responses using each baseline method and ASDP.
b). For ASDP, store the intermediate outputs (identified stereotype, counter-examples, analysis, and reformulated query) for later analysis.

Step 6: Evaluation Metrics: Develop a set of evaluation metrics:
a). *Stereotype Adherence Score:* Manually rate responses on a scale of 1-5 for how much they adhere to common stereotypes.
b).*Diversity Score:* Measure the lexical and semantic diversity of the generated responses.
c).*Factual Accuracy:* Verify the factual claims made in the responses.
d). *Plausibility of Counter-Examples:* Rate the realism and plausibility of the generated counter-examples in ASDP.

Step 7: Human Evaluation: Recruit a diverse group of 5-10 human evaluators to rate a subset of the responses (50-100) on bias, fairness, and overall quality...

Step 8: Analysis

- Compare the performance of ASDP against the baselines using the defined metrics.
- Analyze the intermediate outputs of ASDP to understand how the model identifies and challenges stereotypes...

...

A.5.2 Example 2: Research Town (Author Simulation)

This method simulates a domain expert to synthesize a focused and plausible research abstract. By adopting the persona of specific researchers, the system generates a condensed, high-impact summary of a potential scientific contribution.

Seed Idea: Language Models as Memory Augmentation
Simulated Authors: Alexei A. Efros, David A. Forsyth **Generated Idea:**

"title": "Language Models as a Cognitive Prosthesis for Memory Augmentation", "authors": ["Alexei A. Efros", "David A. Forsyth"], "abstract": "Human memory is fallible... We propose a system, 'Cognitive Scribe,' that leverages large language models (LMs) to serve as a personal memory prosthesis...", "method": "Our proposed method involves three main components: (1) A wearable device... (2) A secure, on-device data processing pipeline... (3) A fine-tuned large language model...", ...

A.6 Experimental Details

For Gemini, Claude, and GPT models, we use the official API service. If applicable, we set the max output token to be 8192, temperature to be 0.7, top p to be 0.7, and top k to be 50. For TF-IDF and t-SNE, we use the implementations of Scikit-Learn.

We present the details of prompts for element extraction and element merging as follows:

Element Extraction Prompt
You are a helpful assistant who annotates the paper with its title and the abstract:
Please annotate the paper with the following information:
1. The themes of the paper (As, *e.g.*, few-shot, long-tail, less is more, in-the-wild, self-refine, look-ahead, hindsight, memory, self-, rethink, weak to strong, granularity, in-context learning, reference free, grokking, self-evolve, long-tail, compositionality, multi-hop, modular, etc.) 2. The domains of the paper (Bs, *e.g.*, question answering, argument mining, planning, RAG, calibration, reasoning, safety, debate, memorization, automated research, etc.) 3. The method insights of the paper, especially novel architecture (Cs, *e.g.*, Mamba, RWKV, LLMs, Self-attention, LLMs, etc.) 4. The templates of the paper title/abstract (templates, *e.g.*, Comparing C1 and C2 in B1 with A1, etc.)
Requirements: 1. There can be multiple A, B, C, and one Template. 2. Use generic keywords of A, B, C, and Template to allow reuse, instead of specific ones for each paper. 3. Make sure keywords are exclusive among A, B, C.
Please output the annotation in the following JSON format:
"A": ["few-shot", "long-tail"], "B": ["argument mining"], "C": ["Mamba"], "Template": ["Comparing C1 and C2 in B1 with A1"]
An Example: Title: Thrust: Adaptively Propels Large Language Models with External Knowledge
Abstract: Although large-scale pre-trained language models (PTLMs) are shown to encode rich knowledge in their model parameters, the inherent knowledge in PTLMs can be opaque or static, making external knowledge necessary. However, the existing information retrieval techniques could be costly and may even introduce noisy and sometimes misleading knowledge. To address these challenges, we propose the instance-level adaptive propulsion of external knowledge (IAPEK), where we only conduct the retrieval when necessary. To achieve this goal, we propose to model whether a PTLM contains enough knowledge to solve an instance with a novel metric, Thrust, which leverages the representation distribution of a small amount of seen instances. Extensive experiments demonstrate that Thrust is a good measurement of models' instance-level knowledgeability. Moreover, we can achieve higher cost-efficiency with the Thrust score as the retrieval indicator than the naive usage of external knowledge on 88% of the evaluated tasks, with 26% average performance improvement. Such findings shed light on the real-world practice of knowledge-enhanced LMs with a limited budget for knowledge seeking due to computation latency or costs.
Output: {"A": ["adaptive"], "B": ["RAG"], "C": ["Large Language Models"], "Template": ["A1 application of B1 to C1"]}
You task:
Title: title
Abstract: abstract
Output:

Element Merging Prompt
You are a helpful assistant who merges the keywords or phrases with their semantic similarity.
Here is a list of keywords or phrases for a domain: keywords
Requirements: 1. Please merge the keywords by creating a keyword group in a valid decodable JSON format. 2. No need to merge the keywords that are not to similar. 3. Output the JSON format only. 4. Do not be lazy, please list the full output covering all keywords or phrases without omission.
The potential JSON format is: {{"keyword_group_name": ["keyword1", "keyword2", "keyword3"]}}
The keyword group name should be a short and concise description of the keyword group. An example keyword group: "RAG": [RAG,retrieval augmented generation, retrieval augmentation]
Your output:

Idea Rewriting Prompt
You are a senior professor in AI, and your students propose to do a combination. Can you refine the title into a good one that can be accepted by top conferences such as ACL 2025 and ICLR 2026? Please output one title only, with no other text. Requirements: 1. Do not hallucinate, 2. do not use any existing paper names in your pretraining data. 3. make sure the title is with an outstanding paper quality so that your student can be happy and successfully graduate.

Community	A	B	C
ACL 2024	adaptive, less is more, hierarchical, in-the-wild, self-refine, hindsight, rethink, grokking, long-tail, compositional, multi-hop	agent, planning, retrieval, safety, calibration, reasoning, memorization, persuasion, debate	Mamba, RL, Linear Models, KV Cache, Quantization, Diffusion, Self-attention, Self-supervision
CV	test-time Training, meta learning, active learning, open-set calibration, open-vocab grounding, continual learning, knowledge guided learning, inverse rendering	image classification, detection, segmentation, optical flow estimation, action recognition, style-transfer, denoising	vision transformer, NeRF, ConvNext, style-GAN, point-transformer, Perceiver, Instant-NGP, Yolo, UNet, LoRA
RL Theory	Value A3	Value B3	Value C3

Table 8: Lists of Theme (A), Domain (B), and Method (C) written by researchers from different communities. NLP, CV, and RL Theory denote natural language processing, computer vision, and reinforcement learning theory, respectively.

A.7 (Details) Elements mined from conferences

A.8 Bijective Coverage Evaluation Details

For the bijective coverage analysis in Section 4.3, we implement a two-stage evaluation process using Gemini 2.0 Flash.

Decomposition Prompt. We use the following prompt to test whether research papers can be decomposed into our A+B+C framework:

> You are an expert in AI research taxonomy. I will give you lists of research themes (A), domains (B), and methodologies (C), and a paper title. Your task is to find the MOST SPECIFIC and ESSENTIAL concepts from these lists that capture the core of this paper.
> THEMES (A): {themes}
> DOMAINS (B): {domains}
> METHODOLOGIES (C): {methodologies}
> PAPER TITLE: "{title}"
> Extract the most essential concepts that would allow someone to reconstruct a similar title: - Select relevant themes from list A - Select relevant domains from list B - Select relevant methodologies from list C
> Focus on concepts that are ESSENTIAL to the paper's contribution, not just tangentially related.
> Respond with a JSON object: {{"selected_A": ["theme1", "theme2"], "selected_B": ["domain1"], "selected_C": ["methodology"], "confidence": 0.0-1.0, "explanation": "brief explanation"}}

Reconstruction Prompt. For testing reconstruction capability, we use:

> You are a senior AI researcher. Given these research concepts, generate 5 different realistic paper titles that combine them:
> THEMES: {themes} DOMAINS: {domains} METHODOLOGIES: {methodologies}
> Generate 5 diverse paper titles that would be suitable for a top-tier conference like ACL/EMNLP/NeurIPS.
> Each title should: 1. Combine all the given concepts naturally 2. Sound like a real research paper title 3. Be specific and technical 4. Be different from the others
> Format as a numbered list: 1. [Title 1] 2. [Title 2] 3. [Title 3] 4. [Title 4] 5. [Title 5]

Evaluation Metrics. We consider a paper *decomposable* if it can be successfully mapped to at least one element from each disk (A, B, C). For *reconstruction*, we generate 5 candidate titles and compute Jaccard similarity between each candidate and the original title, taking the maximum similarity. Papers with similarity $\geq 30\%$ are considered successfully reconstructible.

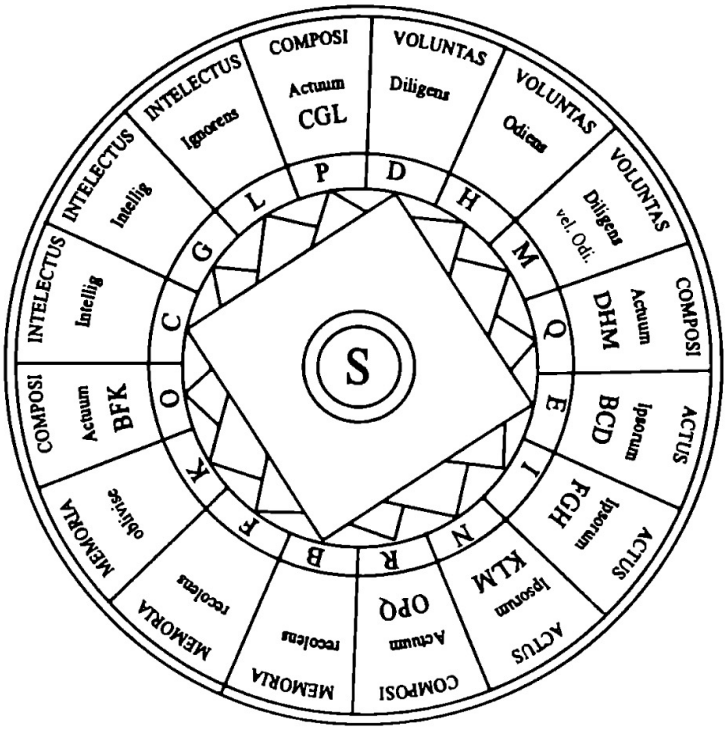

Figure 3: The illustration of the original Ramón Llull's thinking machine.

A.9 Original Ramón Llull's *Ars combinatoria*

We present the original Ramón Llull's thinking machine in Figure 3 from Borges (1937).