Dear Meta-Reviewer and Reviewers,

We sincerely thank you for your thoughtful feedback and constructive suggestions during the May ARR cycle. This document outlines the revisions made in our resubmission of the paper, *"CHIMERA: A Knowledge Base of Idea Recombination in Scientific Literature."*

We use the following color scheme for clarity:

- **Yellow**: Original reviewer comments
- **Green**: Descriptions of the corresponding revisions

Additionally, this document includes a color-coded version of our resubmission (page 10), where the discussed revisions are highlighted in **Blue**.

Thank you again for your time and consideration.

Best regards,
The Authors

Overview of Changes

- **Contribution summary.** We added a bullet-pointed list in the introduction to clearly highlight our key contributions.
- **Motivation and impact.** The introduction and related work sections were strengthened to better position CHIMERA relative to prior work. We expanded the discussion of its practical value, with concrete applications (e.g., meta-scientific analysis, supervised ideation) and examples of insights enabled by the dataset.
- **Extraction figure.** Figure 2 now includes step-by-step illustrative examples and a new "Application" panel to clarify the methodology and downstream utility.
- **Annotation setup.** We now emphasize annotator expertise (PhDs in scientific fields), detail the selection process, and report both F1 and Cohen's Kappa agreement scores. We also cited works with a similar setup, and added a brief analysis of common disagreement types.
- **Extraction quality.** We added a new large-scale LLM-as-a-Judge assessment of the graph data to complement the extraction model evaluation.
- **Clarified taxonomy.** We emphasized the cognitive grounding of the blend/inspiration schema, and added a new analysis demonstrating its coverage of more nuanced recombination types. We now explicitly discuss future directions for refining the taxonomy in the Limitations section.
- **New prediction illustration and analysis.** We added a new figure (Figure 5) to visually clarify the prediction setup. Additionally, we now include a qualitative error analysis of reranking outputs to shed light on the reranker behavior and common failure cases.
- **Increased user-study scale**. We now annotate 100 examples from 5 researchers.
- **Expanded limitations.** The discussion of limitations was extended to more clearly articulate the scope of our extraction and prediction evaluation.

Point-by-Point Responses

Area Chair aLC2

Comment: "Concerns were raised about the quality of CHIMERA. The annotators are came from a crowdsourcing platform, raising questions about their expertise. Since most of the data was extracted automatically by the model, ensuring extraction accuracy remains a significant challenge. The current blend/inspiration schema is also somewhat coarse-grained; reviewers noted it may fail to capture more nuanced relations."

Response:

- Regarding annotators, both annotators hold PhDs in scientific fields with extensive past experience in scientific text annotations. They were selected through a rigorous screening process (including a test annotation sample) from a pool of candidates we had previously successfully collaborated with. We now emphasize their qualifications and our selection process more clearly in the revised manuscript (Section 3.1, lines 229-235). It is very common to hire expert annotators from platforms such as UpWork.

- Regarding the quality of the extracted data, we evaluate the extraction model on a high-quality, expert-verified test set. To reinforce the quality of extraction, we now include in the revised manuscript a large-scale assessment of the graph data (Section 4.2, lines 353-368, Appendix B.7), showing that over 80% of the extracted examples are accurate, with the majority of errors stemming from extractions of correct recombinations, with insufficiently informative entities (e.g., extracting a combination of "neural" and "feature-based approaches" from "we combine neural and feature based approaches"). Importantly, a core contribution of our work is the introduction of a **new information extraction task** for identifying idea recombinations in scientific texts. We expect that future work building on our expert-annotated training data will further improve extraction performance. However, the presented downstream applications already demonstrate that it's feasible to identify meaningful recombination patterns and gain inspiring insights using the current data.

- Regarding the schema, we added an analysis showing how the schema is able to capture many nuanced types of recombinations, such as reduction and abstraction (Appendix D.2). We also further clarify that we intentionally adopted the blend/inspiration distinction to strike a balance between expressiveness and annotator consistency. These two categories reflect fundamental modes of idea recombination, and our work, to our knowledge, is the first to extract this information from scientific papers. While we agree that more fine-grained distinctions (e.g., abstraction, reduction) could offer additional nuance, our current taxonomy is grounded in cognitive theory and serves as a tractable and scalable starting point

Comment: "The motivation and impact of CHIMERA are not made sufficiently compelling. The authors should better highlight, in the related work section, how this dataset differs from existing similar datasets. In the introduction, it would help to include stronger justification—such as practitioner surveys or other evidence—that demonstrates the clear need for such a dataset. Furthermore, the authors could more explicitly demonstrate the dataset's value by showing what meaningful insights it enables—especially findings that could guide future research directions."

Response:
In the revised manuscript, we significantly expanded both the introduction and related work sections to more clearly articulate the motivation for CHIMERA. We emphasize that we demonstrate CHIMERA's value through several analyses and findings grounded in the knowledge base.

- The introduction now more clearly motivates the need for CHIMERA by showing increased interest in scientific analysis and ideation (lines 065 - 071), and by highlighting limitations in prior work in scientific analysis (lines 084-087) and ideation (lines 099-100). We also clarify how CHIMERA directly addresses these gaps, offering a precise and semantically rich resource for studying innovation and enabling meta-scientific analysis (lines 079-091), and training hypothesis generation models (lines 100-108). A revised Figure 2 also highlights CHIMERA's practical value and potential impact in a range of downstream applications, e.g., meta-scientific analysis and scientific ideation.
- The related work section has been revised to emphasize how our dataset and approach differ from existing IE work that is not able to capture the recombination phenomenon (lines172-184). We also more clearly reference an expanded discussion of prior work limitations in Appendix G.
- In Section 4.3, we present empirical insights enabled by CHIMERA on domain-level recombination trends, including interdisciplinary inspiration patterns and the evolving influence of NLP on other fields (e.g., Figure 4). In the revised version, we go a step further by including concrete examples of predicted research directions that human evaluators found particularly inspiring (Appendix F.1), highlighting CHIMERA's potential to spark novel scientific ideas.

We hope these revisions help clarify the motivation, contributions, and unique value of CHIMERA.

Reviewer QWBs

Comment: "The initial seed data for training the model to extract the idea recombination is annotated by only two people. In this regard, the annotated seed data and the resulting trained model may reflect only the narrow view of those two people (i.e., may be biased). It would be great if the authors could increase the number of annotators, and further measure and report the annotator agreements more rigorously."

Response: We use two expert annotators with PhDs, and, in addition, a third annotator further reviews the annotations. Importantly, we note that other recent work involving the annotation of information extraction tasks also uses two annotators (e.g., arXiv:2410.03594, EMNLP (main) '24, arXiv:2311.09736, ACL findings '24; our inter-annotator agreement is comparable to theirs).

The task involves comprehending a scientific text, identifying entities and relationships. Annotators are required to follow our carefully constructed annotation guidelines. While increasing the number of annotators could be worthwhile, it is also resource-intensive in our setting.

We have included examples of works using a similar annotation setup in Section 3.1 (lines 224–228) and added detailed information on our agreement computation (Appendix A). Additionally, we revised the results discussion in Section 4.1 (lines 305–309) to more clearly reference the implementation details of the agreement analysis.

Comment: "It is unclear, from the perspective of exploring idea recombinations, what is the difference and the ultimate main novelty of this work against existing idea recombination works (Kang et al., 2022;

Radensky et al., 2024a). I understand that the authors train the model to automatically extract the idea recombination; however, the novelty and impact of this work on simply training the model seems marginal."

Response: Our work is very different in nature and focus from Kang et al. (2022) and Radensky et al. (2024a). Those papers focus on interactive, human-in-the-loop ideation tools that help researchers discover analogous ideas. In contrast, our work focuses on extracting a knowledge base of real recombinations that appear in scientific papers, which supports numerous potential applications. Using our KB, for example, it is possible to explore all analogical inspirations proposed in a given scientific topic, to study how subfields draw inspiration or combine methods from others, to analyze how recombinations evolve over time, to train models to predict new recombinations, etc. We indeed expect future ideation tools to build on our work.

Comment: "It is unclear how to fine-tune the retriever for the recombination prediction task. I assume the pair of the head entity and relation is the query, and all the tail entities are the corpus, from which the authors embed them into the vector space to do dense vector-based retrieval based on the similarities between the query and all the elements in the corpus."

Response: We clarify that the task is formulated as retrieving the tail entity, given a query constructed from a context string, a head entity, and a relation type. For example, given the triple:

```
("Design ideas and categories dedicated to motivating users to engage with
privacy-related information", inspiration, "the Protection Motivation Theory")
```

The goal is to predict the tail entity "the Protection Motivation Theory", using the following input query:

```
Raising users' privacy awareness is crucial in modern technology
environments, yet existing efforts have largely failed to systematically
address the challenge of motivating users to engage with privacy-related
information. This gap highlights the need for effective design strategies
that can enhance user motivation and attention to privacy issues. What would
be a good source of inspiration for "Design ideas and categories dedicated to
motivating users to engage with privacy-related information"?
```

We generate these queries using relation-specific templates (e.g., for inspiration: `<CONTEXT>\nWhat would be a good source of inspiration for "<HEAD>"?`). For blends, which are symmetric, we train in both directions (head → tail and tail → head).

Each query is embedded using a sentence encoder, and candidate tail entities are retrieved from the vector space of entity embeddings. Fine-tuning is performed using contrastive loss, where positive samples are true recombination edges and negatives are sampled from non-existent edges.

We revised the task description in Section 5 (lines 419-431) and added a new figure (Figure 5) illustrating example inputs and outputs to clarify the prediction task setup.

Comment: "The performance degradation on the recombination prediction task when using the reranker is not intuitive, as typically, when using the reranker, the performance improvement is significant when k (of Hit at k) is small."

Response: We agree that the performance degradation is counterintuitive. The reranker may down-rank the gold answer in cases where multiple plausible candidates exist. For example:

Query: ...What would be a good source of inspiration for "A novel approach to sketch colourisation"?
Gold answer: the universal childhood activity of colouring and its professional applications in design and story-boarding

Top-20 (before reranking):

1. the universal childhood activity of colouring and its professional applications in design and story-boarding
2. the artists painting process
3. the exercises artists use by looking at amorphous ink splotches for creative inspiration
4. the field of computer graphics
5. recent diffusion-based image editing techniques

...

10. the colorization of gray images

...

20. the process of erasing pencil script

Top-20 (after reranking):

1. recent diffusion based image editing techniques
2. the recently proposed method for 2D image editing -Delta Denoising Score
3. the colorization of gray images

...

8. the universal childhood activity of colouring and its professional applications in design and story-boarding

...

20. the process of erasing pencil script

As shown, the reranker ranks down the gold answer “the universal childhood activity of colouring…” (1→8), and promotes other plausible inspiration sources instead (such as “the colorization of gray images” or “recent diffusion based image editing techniques”). We address this point more clearly in the revised version (Section 5.2, lines 464-471).

Comment: “I am wondering if there are other types of recombination beyond blend and inspiration, which can be potentially explored in future work.”

Response: We present *blend* and *inspiration* as foundational categories that capture two core modes of idea recombination. That said, we agree that more fine-grained types, such as analogy, simplification, or reduction, are worth exploring in future work. To support this direction, we now include a qualitative analysis showing how our taxonomy covers a wide range of recombination patterns (Appendix D.2), along

with illustrative examples (Table 17).

Follow-up comment: “Thank you for your response. I have carefully reviewed it, and I believe that some concerns remain (while the concerns not noted here are all addressed). Specifically, I am still concerned about the reliability of annotating the initial seed data with only two people, which may reflect only their narrow perspective. Also, the agreement can be measured with its dedicated metric, such as Cohen's kappa, while I still believe that increasing the number of annotators or justifying the concern on bias is needed. Additionally, it is still unclear to me why the performance becomes inferior with reranking when k is small. The specific example provided by the authors makes sense, but it may not support the overall trend (i.e., having many plausible candidates for each sample seems to be justified more).”

Response: Regarding the number of annotators, as we mentioned, other published papers (EMNLP, ACL) also use the same number. In fact, one of the most widely used datasets for scientific IE, SciERC (EMNLP arxiv:1808.09602), was annotated by only one annotator. There are more examples (arXiv:2106.14463, NeurIPS 2021; arXiv:2006.03039, ACL (main) 2020). We further stress that this annotation task with two PhD annotators is costly and labor-intensive. In the revised version, we cite previous work using a similar annotation setup, and briefly address the tradeoff between annotation quality and quantity (Section 3.1, lines 224-228).

Regarding F1 for measuring agreement, we reemphasize this is standard for computing annotators' agreement in information extraction settings (e.g., arXiv:2410.03594, EMNLP (main) '24, arXiv:2311.09736, ACL findings '24, arXiv:2006.03039, ACL (main) 2020, all use F1 to present inter-annotator' agreement). The revised version cites related information extraction work that uses F1 as an agreement metric (Section 4.1, lines 305–309) and reports additional Cohen’s Kappa scores for completeness (Section 4.2 lines 332-340).

Regarding the reranker degradation, we agree that having many plausible candidate answers may not entirely explain the lower performance. We have added a qualitative reranker error analysis aiming to better understand this phenomenon (Appendix E.3), and provided examples where the reranker prefers an alternative answer to the gold one (Table 19).

Reviewer Epy3

Comment: “The limitations of using single abstracts reflect potential incompleteness, as recombination often comes from the broader context of sections like the introduction, experiments, and related work—only partially reflected in abstracts.”

Response: We importantly note that the great majority of scientific IE tasks focus on paper abstracts, despite losing coverage of potentially valuable information in the full text (e.g., arXiv:2305.05471 EMNLP findings '23, arXiv:2311.09736, ACL findings '24, arXiv:2406.06357, NAACL findings '25). Annotating full texts with experts introduces much greater annotation complexities and costs. That said, we fully acknowledge this limitation; indeed, we are actively exploring extending CHIMERA to full-text documents, starting with the introduction and background sections.

We added a paragraph to the limitations section addressing the extraction scope explicitly (Limitations, lines 522-537)

Comment: “Though Figure 3a and 3c show that cross-disciplinary inspirations are frequent, they focus on high-level categories (e.g., cs.CL, cs.CV), without deeper exploration of how combinations evolve in specific subdomains or their scientific impact.”

Response: In this work, which lays the foundations of the task of recombination extraction and analysis, we chose to rely on arXiv's curated taxonomy because it provides a widely adopted, human-defined categorization scheme that is relatively stable, interpretable, and comparable across papers. This allowed us to conduct principled, reproducible analyses of cross-domain recombination without introducing additional assumptions about what constitutes a "subdomain."

Reviewer 7gN3

Comment: "The task lacks a single "correct" answer, leading to potential false negatives in automated evaluation. While human studies show annotators prefer CHIMERA's suggestions, the sample size is small (70 examples), and the evaluation relies on volunteers with specific expertise, risking bias. This makes it hard to quantify the model's true utility compared to baselines like GPT-4o."

Response: Assessing hypothesis/idea generation is inherently open-ended, with no single "correct" answer - a challenge shared by many generation tasks in science and creativity.

Our paper includes multiple evaluations and analyses of the prediction use case:

1. A large-scale automatic evaluation, showing that fine-tuning prediction models over CHIMERA significantly improves their performance.
2. A complementary human study, designed as a proof of concept, in which expert raters consistently preferred outputs from the CHIMERA-fine-tuned model over those from a strong zero-shot LLM baseline (GPT-4o).

We revised the limitations section (lines 538–549) to acknowledge the open-ended nature of the prediction task and the challenges in quantifying model utility.

Regarding the human evaluation of the pilot prediction task, the revised version presents an extended user study, showing results for 100 examples, annotated by five researchers (Section 5.1, lines 472-488). Importantly, we emphasize that **our main contribution/task is the CHIMERA dataset itself**: the first knowledge base of scientific idea recombinations, which enables a variety of downstream tasks. Recombination prediction serves as one illustrative use case, showcasing how CHIMERA can be leveraged in practice.

Comment: **"**The pipeline heavily relies on GPT-4o for domain classification, context extraction, and span similarity judgment, but only experiments with Mistral-7B, Llama-3.1, and GoLLIE for extraction. Other LLMs (e.g., open-source models without API access) or non-generative approaches (e.g., BERT variants) are understudied, leaving uncertainty about performance in resource-constrained settings or with different architectural choices."

Response: We use and evaluate open-source models extensively. The core extraction model, used to construct CHIMERA at scale, is Mistral-7B, an open-source model fine-tuned with parameter-efficient methods (LoRA) to ensure accessibility and reproducibility.

We also evaluated a broad range of extraction baselines, including:

- **Open-source LLMs** (Mistral-7B, LLaMA 3.1, GoLLIE),
- **Proprietary LLMs** (GPT-4o),
- **Non-generative models**, such as PURE, a SciBERT-based token classifier for recombination entity span extraction.

Finally, the recombination prediction model uses lightweight sentence encoders (e.g., all-mpnet-base-v2, ~108M parameters), further underscoring the viability of the collected data in resource-constrained settings.

Comment: “The annotation guidelines (e.g., distinguishing blend/inspiration) may have hidden ambiguities. While inter-annotator agreement is reported, the paper does not provide detail specific disagreements or how they were resolved, leaving issues about schema consistency.”

Response: Our annotation process followed a structured protocol, with detailed guidelines refined through multiple pilot rounds to ensure consistency. While inter-annotator agreement scores were already reported, we agree that illustrating common disagreement patterns can further clarify how we addressed potential ambiguities in the schema. In the revised version, we now include representative examples of such disagreements, along with explanations of how they were resolved (Appendix A.1).

Comment: “The framework defines only two coarse-grained types (blend/inspiration), missing nuanced relations like metaphor, metonymy, or hierarchical abstraction. For example, the paper cites "nature-inspired optimization algorithms" as inspiration, but complex analogies (e.g., quantum computing + machine learning) might better fit a blend-inspiration hybrid category.”

Response: Our framework intentionally focuses on two high-level categories (blend/inspiration) as foundational types that capture a wide range of recombination phenomena. These categories are grounded in established cognitive science theories of creativity and were designed to support scalable, high-quality annotation while maintaining conceptual clarity. We emphasize that our work is the first to collect this type of information. While more fine-grained relations, such as metaphor or abstraction, exist, our goal in this work is to establish a robust and generalizable taxonomy as a first step toward structured modeling of scientific recombination.

In the revised manuscript, we emphasize the theoretical grounding of our taxonomy with references to relevant cognitive science literature (Section 2, lines 133–139). We also provide a qualitative analysis illustrating its coverage across diverse recombination patterns (Appendix D.2) and include illustrative examples of more nuanced recombination types extracted using our schema (Table 17).

Comment: “The recombination prediction test set includes post-2024 data, but the model’s ability to generalize to emerging fields or sudden scientific shifts (e.g., new subdomains in AI) is untested. The human study involves only three volunteers, limiting confidence in its generalizable impact across research communities.”

Response: Our work encompasses multiple evaluations and analyses, including both automated and human evaluations of predictions. While evaluating predictions on emerging fields or sudden scientific shifts should be an interesting direction, this applies in general to any work on scientific hypothesis/idea generation, with many non-trivial methodological challenges.

We re-emphasize that our primary contribution is building the first knowledge base of scientific idea recombinations, and demonstrating its multiple uses.

Comment: “Explicitly link recombination types (blend/inspiration) to cognitive science theories (e.g., Gentner’s structure mapping theory) to ground the work in deeper theoretical foundations. This would elevate the paper’s contribution beyond empirical analysis to theoretical innovation.”

Response: We revised the related work section to emphasize this connection (Section 2, lines 133-139).

Comment: “Some figures (e.g., Figure 3 on domain relations) lack clear labels or legend explanations. Adding more detailed captions and color-coding to distinguish between blend/inspiration edges would help readers interpret cross-domain patterns more easily.”

Response: Figure 3 separates blend and inspiration edges into distinct subfigures: 3a shows inspiration, and 3b shows blends. This separation makes additional color-coding for edge types unnecessary. Furthermore, we already apply color-coding to node domains to aid interpretation. To improve clarity, we have revised the figure captions to provide a clearer explanation of the structure and contents of each subfigure.

Comment: “The paper cites works up to 2024, but given its 2025 context, including recent breakthroughs in scientific IE (e.g., new LLM-based extraction models from 2025) would ensure the literature review is current.”

Response: While the specific 2025 works referenced by the reviewer are unclear, we have surveyed recent publications from 2025 and identified a few that are relevant to idea recombination. We have updated the related work section to include these (Section 2, lines 151–153).

CHIMERA: A Knowledge Base of Scientific Idea Recombinations for Research Analysis and Ideation

Anonymous ACL submission

Abstract

A hallmark of human innovation is *recombination*—the creation of novel ideas by integrating elements from existing concepts and mechanisms. In this work, we introduce CHIMERA, a large-scale Knowledge Base (KB) of over 28K recombination examples automatically mined from the scientific literature. CHIMERA enables large-scale empirical analysis of how scientists recombine concepts and draw inspiration from different areas, and enables training models that propose novel, cross-disciplinary research directions. To construct this KB, we define a new information extraction task: identifying recombination instances in scientific abstracts. We curate a high-quality, expert-annotated dataset and use it to fine-tune a large language model, which we apply to a broad corpus of AI papers. We showcase the utility of CHIMERA through two applications. First, we analyze patterns of recombination across AI subfields. Second, we train a scientific hypothesis generation model using the KB, showing that it can propose novel research directions that researchers rate as inspiring. We release our data and code at https://anonymous.4open.science/r/CHIMERA-0510.

1 Introduction

Recombination—the creation of novel conceptual or physical solutions by combining existing mechanisms, methods, perspectives—is a widely recognized mechanism of ideation and innovation (Uzzi et al., 2013; Youn et al., 2015; Shi and Evans, 2023). It involves reinterpreting prior ideas by breaking them into components and blending them into new solutions (Knoblich et al., 1999; McCaffrey, 2012). This often requires forming abstract structural mappings across domains (Gentner et al., 1997; Gentner and Markman, 1997; Gentner and Kurtz, 2005; Chan et al., 2011; Frich et al., 2019)—e.g., as in bio-inspired algorithms that apply biological principles to computational problems.

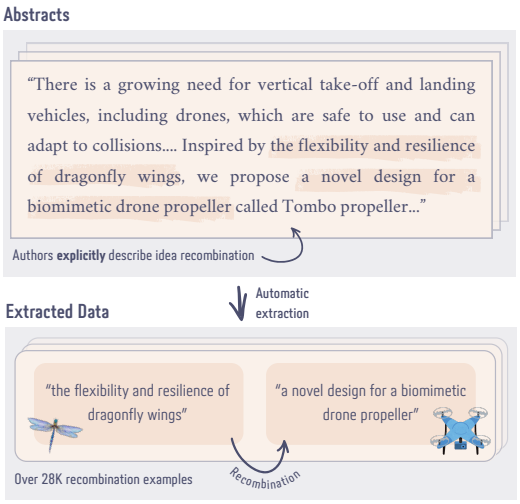


Figure 1: We propose a new task of extracting *recombinations*: examples of how scientists connect ideas in novel ways. The extracted information enables applications in research analysis and automated ideation.

In this work, we introduce a new task: extracting recombinations from scientific papers. We present CHIMERA, a large-scale knowledge base (KB) of recombination examples automatically mined from papers. Figure 1 shows one such case, where a robotic design is inspired by animal mechanics. CHIMERA enables exploring, analyzing, and training models on such examples, capturing a fundamental pattern of human ingenuity.

Unlike simpler concept co-occurrence methods (Krenn et al., 2022) or general scientific extraction schemas (Luan et al., 2018), CHIMERA targets cases where authors *explicitly* describe recombination as central to their contribution. We focus on two broad recombination types: ***blends***, which combine concepts into novel approaches (e.g., augmenting classical ML with quantum computing), and ***inspirations***, where ideas from one domain spark solutions in another (e.g., using bird flock behavior to coordinate drones). CHIMERA captures both intra- and cross-domain cases, including analogies,

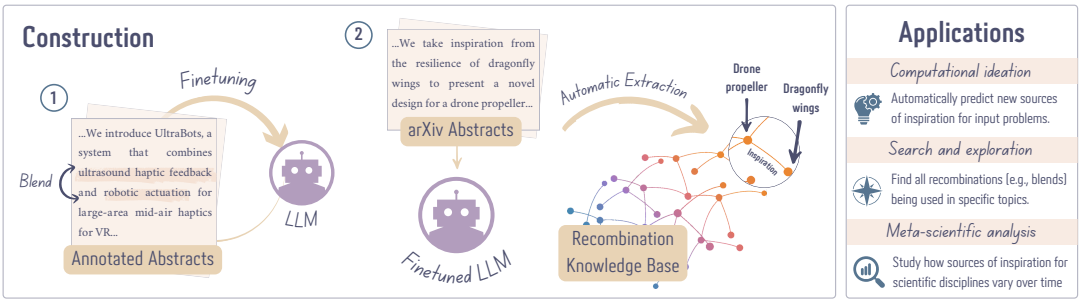


Figure 2: CHIMERA KB construction and applications. **Construction**: (1) We use human-annotated recombination examples to fine-tune an LLM for information extraction; (2) the model extracts recombinations from arXiv abstracts to build a large-scale KB. **Applications**: CHIMERA supports diverse use cases, including computational ideation, exploration of recombination patterns across scientific domains, and meta-scientific analysis.

abstractions, and reductions.

The resulting KB and methods enable diverse uses (Figure 2). In this paper, we focus on two applications that have seen growing interest in recent years: **Science Analysis** (Fortunato et al., 2018; Wahle et al., 2023; Pramanick et al., 2025) and **Scientific Ideation** (Wang et al., 2024; Si et al., 2024; Radensky et al., 2024; Garikaparthi et al., 2025).

Science Analysis. We demonstrate how CHIMERA supports meta-scientific analysis (also known as *science of science* or *scientometrics*) (Fortunato et al., 2018): empirical studies of how innovation unfolds. Researchers conducting meta-science analyses aim to understand how a field (e.g., AI) evolves over time and identify trends (e.g., emerging connections across areas). CHIMERA allows analysis of how ideas are combined within and across domains (Shi and Evans, 2019), and of how disciplines, topics and concepts inspire one another. This provides a *direct and precise* alternative to traditional citation-based (Wang et al., 2015; Myers et al., 2013; Wahle et al., 2023) or co-occurrence-based approaches (Frohnert et al., 2024), which are often coarse and noisy. Unlike these methods, CHIMERA allows to identify how a scientific idea is formed by blending concepts or by taking inspiration from another concept, unlocking new and also more granular analyses.

Scientific Ideation. We show how CHIMERA supports training and evaluating scientific hypothesis generation models (Wang et al., 2024), by learning from patterns of past recombinations to propose novel concept blends or inspirations (e.g., new analogical inspirations). Prior work has explored suggesting analogical recombinations via unsupervised discovery (Radensky et al., 2024; Hope et al., 2017); in contrast, CHIMERA provides the first large-scale resource with *real*, author-described examples of how research problems were addressed via recombination. This enables supervised recombination models to observe many examples of how recombinations have been applied to specific problems (e.g., the cross-domain inspiration in Figure 1), and *learn* to suggest relevant blends or inspiration directions for new problems.

Finally, CHIMERA also enables faceted search and exploration (Katz et al., 2024). Researchers can search the KB to find cases of cross-domain inspirations within a topic of interest (e.g., search for all robotics ideas inspired by zoology), sparking new creative directions.

To conclude, our contributions are as follows:

- We present CHIMERA, the first knowledge base of idea recombination examples described by authors in scientific papers. CHIMERA distinguishes between two core types: *blends* and *inspirations*, enabling nuanced analysis in downstream tasks.
- We define a novel extraction task to identify recombinations in scientific abstracts, and release a high-quality, expert-verified dataset of 500+ manually annotated examples, accompanied by fine-tuned extraction baselines.
- We show CHIMERA's utility through two applications: a) *Meta-scientific analysis* of recombination patterns, and b) *Computational ideation*, where models trained on CHIMERA propose novel recombination directions.

2 Related Work

Recombinant creativity Blending concepts and analogical inspiration are core mechanisms of ideation and innovation in cognitive science and creativity research (McKeown, 2014; 201, 2019; Holyoak and Thagard, 1994). These processes in-

Recombination extraction examples
Abstract: "...Current archaeology depends on trained experts to carry out bronze dating... we propose to integrate advanced deep learning techniques and archaeological knowledge..."
Blend: "*advanced deep learning techniques*" ⟷ "*archaeological knowledge*"
Abstract: "...Traditional approaches to enhance dialogue planning in LLMs, ... either face efficiency issues or deliver suboptimal performance. Inspired by the dual-process theory in psychology... we propose the Dual-Process Dialogue Planning (DPDP) framework..."
Inspiration: "*the dual-process theory in psychology*" ⟶ "*enhance dialogue planning in LLMs*".

Table 1: Example **blend** and **inspiration**. Note that blend is a symmetric relation, while inspiration is not.

volve combining or re-representing existing ideas to produce novel concepts and solutions.

Recent work explores how idea recombination can enhance LLM-powered ideation tools. For example, `CreativeConnect` (Choi et al., 2023) lets users recombine keywords to generate graphic sketches, while `Luminate` (Suh et al., 2023) supports recombination of dimensional values to produce diverse LLM responses. `Scideator` (Radensky et al., 2024) is another recent work that helps researchers explore ideas through interactive concept recombination. Other studies focus on recombining ideas from input and analogous artifacts (Srinivasan and Chan, 2024; Chilton et al., 2019) or searching for useful recombinations via iterative idea generation (Yang et al., 2025a,b).

In this work, we build `CHIMERA`, the first KB of scientific idea recombinations, and show how it enables a new approach for recombinant ideation: training models that *learn* from past examples of how ideas have been recombined in scientific texts, to suggest new recombination directions.

Scientific information extraction Information extraction (IE) from scientific texts has been widely studied in NLP. A foundational resource is `SciERC` (Luan et al., 2018), which labels scientific entities (e.g., methods, tasks, metrics) and generic relations (e.g., conjunction) across 500 abstracts. Later datasets, such as `SciREX` (Jain et al., 2020) and `SciDMTAL` (Pan et al., 2024), expand IE to full documents, but similarly focus on standard schema involving scientific concepts and their relations. However, existing extraction approaches are not designed to capture recombination relationships, often resulting in noisy, irrelevant, or misleading outputs, as we illustrate in Appendix G, Figure 20.

In this work, we introduce a focused IE schema tailored specifically to idea recombination, along with a taxonomy that distinguishes between key recombination types: *blend* and *inspiration*. This enables a more precise and semantically rich analysis of cross-domain ideation. For instance, our knowledge base includes numerous analogical inspirations identified in AI research (Figure 1) - patterns that existing scientific IE schemas fail to capture.

3 Extracting Recombinations

Problem definition We focus on scientific abstracts where authors *explicitly link* their contribution to a novel combination or clear source of inspiration. As outlined in the introduction, we capture this with two coarse-grained relation types: **blend** and **inspiration**. **Blend** refers to the fusion of multiple concepts–such as methods, models, or theories–into a new solution or framework. We use the terms "concept blend" and "concept combination" interchangeably. **Inspiration**, by contrast, refers to transferring knowledge or insight from one entity (the *source*) to another (the *target*). This transfer may be realized through analogies, abstraction, or more general links to influential prior work.

Each relation is defined over free-form text spans that represent scientific concepts (see Figure 1; additional examples in Table 1). In blend relations, we refer to the participating entities as *combination-elements*; in inspiration relations, we refer to them as the *inspiration-source* and *inspiration-target*. This schema captures diverse recombination phenomena, such as metaphor, reduction, or abstraction (as illustrated in Appendix D.2) while remaining conceptually clear and efficient to annotate. It offers practical annotation advantages and strong alignment with ideation theory (McKeown, 2014; 201, 2019; Holyoak and Thagard, 1994).

3.1 Recombination Mining

We begin by curating a dataset of annotated recombination examples, which we use to train an information extraction model. The trained model

is then applied to extract recombinations at scale. This process is illustrated in Figure 2.

Example type	# Train	# Test	# Total
blend	124	76	100
inspiration	45	24	69
not-present	195	116	311
All	364	216	580

Table 2: Human-annotated corpus. We include also negative examples without recombinations ("*not-present*").

Data sourcing We annotate AI-related papers from the unarXive corpus (Saier and Färber, 2020)[1]. The data undergo an initial keyword-based filtering to identify works that are more likely to specify idea recombination. Table 8 in Appendix B.1 lists the keywords used in this process. We then assign the remaining abstracts to annotators.

Annotation process Our annotation setup follows standard IE practices, using two trained annotators and expert review to balance quality and feasibility (Naik et al., 2024; Sharif et al., 2024; Pramanick et al., 2025). Following a screening phase, we recruited two annotators with scientific PhDs via Upwork[2], selected from a pool of highly experienced workers we had previously collaborated with. Screening involved annotating examples using a detailed guidelines document[3], followed by a one-hour training session covering additional examples and edge cases. Annotation was conducted using `LightTag` (Perry, 2021), a web-based annotation platform. This process yielded 580 annotated abstracts, summarized in Table 2. To monitor annotation quality, we assign 10% of the examples to both annotators and review this shared subset after each batch. Disagreements are resolved through discussion and revision. All annotations are then reviewed by an NLP expert, who verifies correctness, refines spans, and consolidates annotations.

Automatic recombination mining We use the collected data to fine-tune an LLM-based extraction model. We instruct the model to extract the most salient recombination from the text, if one exists. The model must determine whether the text discusses recombination, infer its type, and identify entities in a single query. We devise the test set from examples where at least two annotators (out of three) agree on the recombination type (or absence), ensuring high-quality, low-ambiguity data. Table 2 summarizes the train and test sets.

[1]We focus on the following arXiv categories: cs.AI, cs.CL, cs.CV, cs.CY, cs.HC, cs.IR, cs.LG, cs.RO, cs.SI

[2]`https://www.upwork.com`

[3]`https://tinyurl.com/4mfdrx2f`

Category	# Interdisciplinary	# Total
Inspiration Edges	5,182 (54.1%)	9,578
Blend Edges	1,792 (9.6%)	18,586
Edges (total)	6,974 (24.8%)	28,164
Nodes (total)	n/a	43,393

Table 3: `CHIMERA` contains over 28K recombinations, a quarter of them interdisciplinary.

3.2 The `CHIMERA` Knowledge Base

We construct the `CHIMERA` knowledge base by mining recombination examples from scientific abstracts, categorizing them, and representing them in a graph where nodes are scientific concepts and edges denote recombination relations.

Large-scale mining We use abstracts from the arXiv dataset[4], which updates monthly and includes more recent papers than unarXive (Saier and Färber, 2020). We apply our fine-tuned extraction model over publications from 2019-2024 within the same CS categories used for the annotation task. We then filter out predictions that don't conform to the data schema or cannot be parsed.

Categorization We apply `GPT-4o` to identify the scientific domain of each extracted entity given the abstract. This enables analyses we perform in Section 4.3. Further, each node is assigned a higher-level discipline—either the arXiv group name (e.g., "*computer-science*" for cs.AI) or a relevant non-arXiv domain. Additional technical details regarding this step appear in Appendix C.

KB building We normalize entities by clustering semantically similar ones. Next, we enrich each edge in the graph with the publication date and arXiv categories of the paper citing it. For simplicity, we focus on binary relations. Table 3 summarizes the resulting KB, including counts of interdisciplinary blends and inspirations.

4 Results

4.1 Experimental Settings

Evaluation criteria We evaluate (1) **Abstract classification**–does the text discuss recombina-

[4]`https://tinyurl.com/mrzksbky`

Task	Baseline	P	R	F1
***Abstract classification**:* *Does it discuss a recombination?*	Human-agreement	0.786	**0.795**	**0.789**
	E2E$_{Mistral-7B-Instruct-v0.3}$	**0.815**	0.762	0.763
	E2E$_{Llama-3.1-8B-Instruct}$	0.630	0.628	0.620
	E2E$_{GoLLIE-13B}$	0.677	0.667	0.667
	E2E$_{GPT-4o}$	0.720	0.580	0.572
	Abstract-classifier$_{Mistral-7B-Instruct-v0.3}$	0.622	0.607	0.602
	Abstract-classifier-CoT$_{Mistral-7B-Instruct-v0.3}$	0.774	0.748	0.749
***Entity extraction**:* *What are the relevant entities?*	Human-agreement	**0.863**	**0.585**	**0.665**
	E2E$_{Mistral-7B-Instruct-v0.3}$	0.587	0.352	0.440
	E2E$_{Llama-3.1-8B-Instruct}$	0.249	0.259	0.252
	E2E$_{GoLLIE-13B}$	0.259	0.187	0.217
	E2E$_{GPT-4o}$	0.138	0.293	0.217
	Entity-extractor$_{GPT-4o}$	0.268	0.263	0.247
	Entity-extractor$_{SciBERT}$	0.324	0.248	0.276
	Entity-extractor$_{PURE_{SciBERT}}$	0.187	0.536	0.271
***Relation extraction**:* *What is the recombination?*	Human-agreement	**0.793**	**0.574**	**0.641**
	E2E$_{Mistral-7B-Instruct-v0.3}$	0.598	0.366	0.454
	E2E$_{Llama-3.1-8B-Instruct}$	0.264	0.294	0.276
	E2E$_{GoLLIE-13B}$	0.301	0.219	0.253
	E2E$_{ICL-GPT-4o}$	0.223	0.385	0.244

Table 4: Recombination extraction results. **Bold** text signifies the best result, while underlined text signifies the second-best. We observe that surprisingly large and capable models struggle with the extraction tasks.

tion?, (2) **Entity extraction**–what entities are described? and (3) **Relation extraction**–what is the relation discussed? For abstract classification, we report precision, recall, and F1. For entity and relation extraction, we adopt a soft matching approach: two entities of the same type match if they refer to semantically similar concepts. We use `GPT-4o-mini`[5] to judge similarity (see prompt and details in Appendix B.4).

A predicted entity may match at most one gold entity, and vice versa; extra matches are ignored. We compute precision, recall, and F1 under this soft matching. For relations, we use partial matching: a predicted relation contributes to the true positive count proportionally to the number of correctly matched entities in a gold relation of the same type.

We measure inter-annotator agreement using the same precision, recall, and F1 metrics, following standard practice in information extraction, where one annotator is treated as the gold reference (Naik et al., 2023; Sharif et al., 2024).

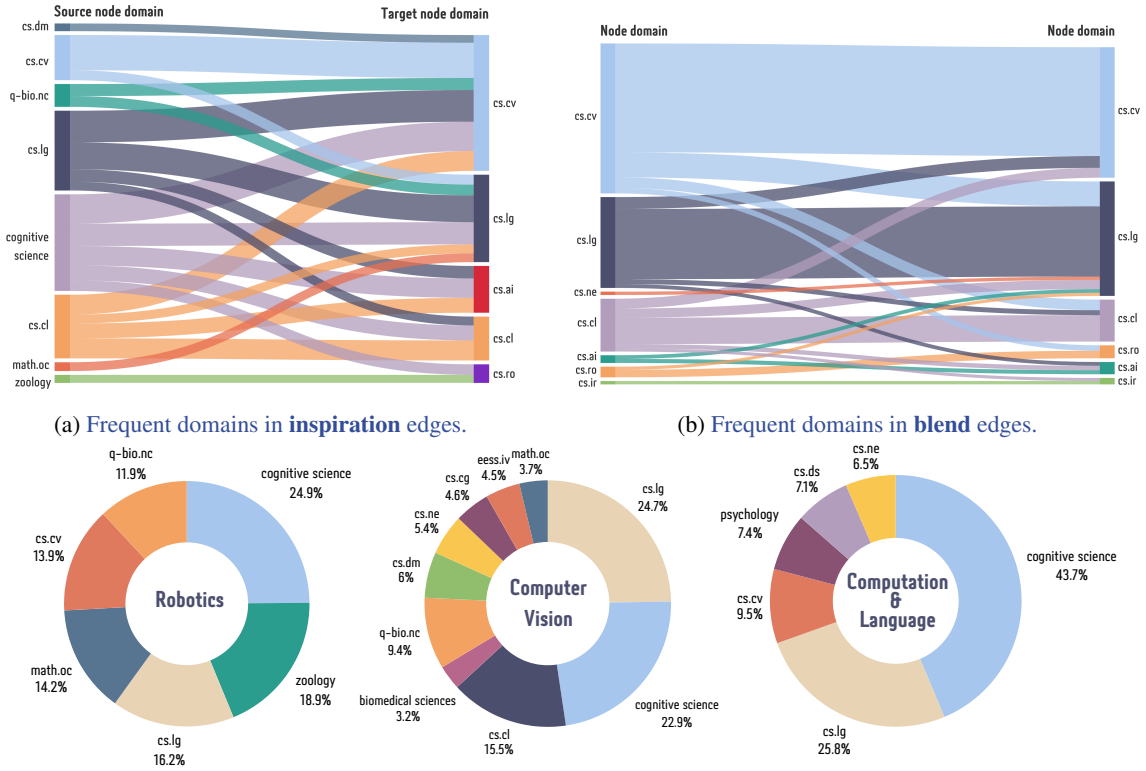
Extraction baselines We evaluate several extraction baselines, including end-to-end (E2E) models that jointly predict whether an abstract discusses recombination, identify its type, and extract the involved entities. In these models, the prediction of any relation is treated as a positive signal for abstract-level classification. We also assess specialized models for individual sub-tasks: *Abstract classifiers*, which predict whether the text discusses recombination, and *Entity extractors*, identify relevant entities. Implementation details for the baselines are in Appendix B.2. To contextualize model performance, we compare results against inter-annotator agreement, used as a proxy for human-level performance. Appendix A presents additional details concerning agreement computation.

[5]Performed on par with `GPT-4o`.

4.2 Extraction Results

Table 4 reports results for abstract classification, entity extraction, and relation extraction. Human agreement scores are 0.760, 0.675, and 0.651 respectively, aligning with soft annotator agreement reported in similar complex extraction tasks (Naik et al., 2023; Sharif et al., 2024). Cohen's κ also indicates moderate to substantial agreement: $\kappa = 0.578$ for abstract classification, 0.631 for entity extraction, and 0.542 for relation extraction. Analysis of annotator disagreement is provided in Appendix A.1—most disagreements concern the presence of a recombination or the identification of its constituent entities, while disagreements over the recombination type are relatively rare.

Fine-tuning `Mistral-7B` on our data yields the best performance across all subtasks. We observe that entity and relation extraction are more challenging than classification for both humans and SOTA LLMs. However, humans still significantly outperform automatic extraction approaches. Ap-



(a) Frequent domains in **inspiration** edges.

(b) Frequent domains in **blend** edges.

(c) Common **sources of inspiration** in leading domains.

Figure 3: Recombinations between areas. cs.*, q-bio.nc and math.oc are arXiv categories. Inspirational connections are often cross-domain (Figure 3a), whereas blends tend to occur within the same domain (Figure 3b). Figure 3c zooms in on a few domains, for example, revealing that robotics often draws inspiration from zoology.

pendix B.5 presents an analysis of extraction errors. Interestingly, focusing on a smaller portion of the recombination extraction task is not necessarily easier than performing it end-to-end, as seen in the lower performance of abstract classifiers. We discuss this point further in Appendix B.3.

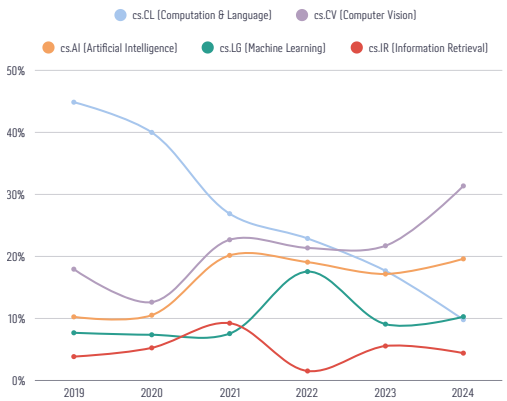


Figure 4: Prevalent domains inspired by cs.CL concepts (NLP). Note the *decrease* in within-domain inspiration.

Large-scale evaluation To assess extraction quality at scale, we evaluate 2,000 CHIMERA examples using a strong LLM-based judge (GPT-4.1). An example is labeled correct if (1) the extracted entities reflect meaningful scientific concepts, and (2) their relation captures a central recombination explicitly described in the abstract. We first validate the judge's reliability by showing high agreement with human annotations on a representative subset. Applied to the full sample, the judge estimates an extraction accuracy of 80.55%, supporting the robustness of our approach. Notably, most extraction errors are minor, typically involving correct recombinations where the extracted entities are less informative than those in the original abstract (see examples and additional details in Appendix B.7).

4.3 KB Meta-Science Analysis

Blends vs. inspirations Figures 3a and 3a present the predominant domain pairs for inspiration and blend relations in CHIMERA (above the 0.9 quantile). The analysis reveals an interesting pattern of a distinct difference in behavior between inspirations and blends: inspirations span a broader range of domains, while blends tend to link within the same or similar domains. This suggests that when human researchers take inspiration they tend

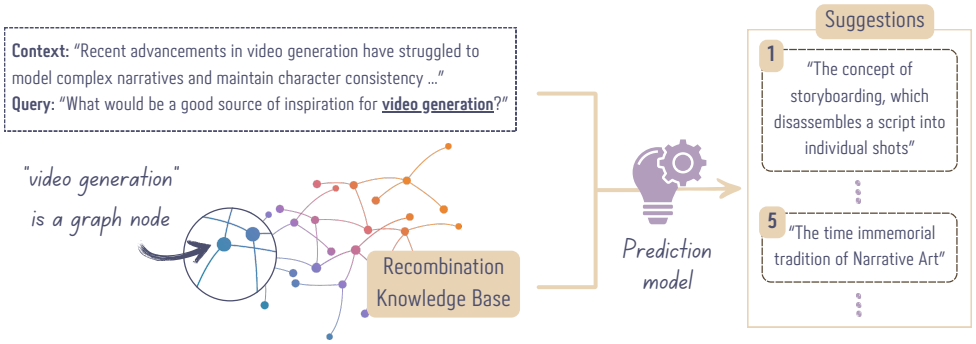


Figure 5: Recombination prediction. Given a context string and a query about recombining a graph node, a model trained on CHIMERA suggests plausible recombination directions, leveraging patterns *learned* from prior examples.

to look across more areas other than their own, but tend to look within their own domain when they build approaches by integrating together mechanisms. Inspirations also link more often to areas not covered by the arXiv taxonomy, e.g., cognitive science and zoology. More research building on our initial analysis and KB can shed additional light on the different ways in which scientists combine concepts to form ideas. Table 16 in Appendix D.1 provides a tabular view of this analysis for clarity.

Inspiration analysis We next analyze how different fields draw inspiration from each other. Figure 3c shows the top 10% cross-domain inspiration sources for three prevalent domains in the graph: **cs.RO** (Robotics), **cs.CV** (Computer Vision) and **cs.CL** (NLP). We observe that while some sources of inspiration (like *cognitive-science*) are commonly shared across related fields, domains may draw inspiration from unique sources (e.g., from *zoology* to *cs.RO* as seen in Figure 1). Interestingly, cs.CV takes more inspiration from cs.CL than vice versa. cs.CL also takes considerably more inspiration from cognitive science than cs.CV, and also takes inspiration from psychology (see example in Table 1), while cs.CV takes more inspiration from biomedical sources. cs.CV also takes inspiration from mathematical topics (discrete math, optimization and control). Appendix B.6 presents examples of such interdisciplinary inspirations.

Split	# Inspiration	# Blend	# Total
Train	5,408	19,909	25,317
Validation	119	411	530
Test	2,026	8,591	10,617

Table 5: We divide prediction data by the publication years associated with each query (training and validation sets < 2024, test set ≥ 2024) to avoid contamination.

Figure 4 shows the percentage of target nodes in domains drawing inspiration from cs.CL (NLP) over five years. We observe two trends: a decrease in intra-domain inspiration (where cs.CL concepts inspire other cs.CL concepts), and an increase in cs.CV (Computer Vision) concepts drawing inspiration from cs.CL.

5 Recombination Prediction

We demonstrate how CHIMERA could be used to train supervised models that recombine concepts and generate novel scientific ideas.

Figure 5 illustrates the recombination prediction task. Given a context string (e.g., "*Recent advancements in video generation have struggled to model complex narratives...*") and a query about recombining a graph node (e.g., "*What would be a good source of **inspiration** for **video generation**?*") the goal is to predict a suitable entity to complete the recombination (e.g., "*The concept of storyboarding...*"). Formally, given a query with a context string (e.g., a problem, experimental settings, goals), an entity e and a recombination type τ, the task is to predict a different entity e' such that (e, τ, e') is a valid edge in CHIMERA.

Data preparation We start by converting edges to pairs of queries and answers. The queries describe the task inputs: a single graph node, the edge recombination type, and a context string, which we extract from the corresponding abstract using `GPT-4o-mini`. Note that this process might leak information regarding the answer (the other graph node) into the query. Therefore, we follow it by applying `GPT-4o-mini` to identify leakages (see examples and implementation details for this step in Appendix E.1). We discard approximately 22% of pairs due to leaks and split the remainder by pub-

Baseline	H@3	H@5	H@10	H@50	H@100	MRR	MedR
all-mpnet-base-v2	0.033	0.042	0.061	0.126	0.170	0.033	1305
bge-large-en-v1.5	0.041	0.053	0.076	0.151	0.199	0.041	1135
e5-large-v2	0.024	0.033	0.050	0.113	0.155	0.026	1590
all-mpnet-base-v2$_{finetuned}$	**0.110**	**0.135**	0.178	**0.320**	**0.402**	**0.106**	**194**
bge-large-en-v1.5$_{finetuned}$	0.104	0.130	0.168	0.306	0.392	0.102	222
e5-large-v2$_{finetuned}$	0.107	0.133	0.173	0.317	0.397	0.103	212
all-mpnet-base-v2$_{finetuned}$ + RankGPT	0.100	0.130	**0.192**	**0.320**	**0.402**	0.097	**194**

Table 6: Recombination prediction results. **MedR** = Median Rank. Fine-tuning on CHIMERA improves MedR **10×**. Interestingly, reranking the top-20 answers using RankGPT boosts the H@10 but slightly reduces H@3,5 and MRR.

lication year, with all papers published after 2024 in the test set. Table 5 summarizes the data splits.

Prediction We experiment with zero-shot and finetuned retrievers based on encoders trained before the test set cutoff year (2024). We next explore applying a GPT-4o-based reranker (Sun et al., 2023) to the top 20 retrieved results to improve our predictions further. The GPT-4o data cutoff is October 2023, meaning the reranker is also unfamiliar with our test set. Appendix E.2 provides additional implementation details for the prediction baselines.

5.1 Prediction Results

Table 6 presents our results. Fine-tuning greatly improves retrievers, decreasing the median rank of the gold answer by an order of magnitude. The last row reports results using RankGPT (Sun et al., 2023) with GPT-4o as a reranker, applied to the top-20 candidates from the best-performing retriever (all-mpnet-base-v2$_{\text{finetuned}}$). While reranking improves Hits@10, it lowers performance on Hits@3, Hits@5, and MRR. These seemingly counterintuitive results are further examined in Appendix E.3. We find that the reranker can inadvertently lower the rank of the gold answer in cases where (i) multiple plausible answers are present, or (ii) the gold answer appears alongside semantically similar variants, making it difficult to distinguish between highly relevant alternatives and the annotated gold.

User study We recruit five volunteers with verified research experience (at least one published paper) and assign them examples based on their expertise. Each example includes an inspiration query and suggestions from six sources: (1) **Ours**: our method, including reranking (2) **Gold**: the gold answer, (3) **Random**: a random test-set node, (4) **GPT-4o**: a GPT-4o generated suggestion, (5) **ZS-CHIMERA**: zero-shot prediction using our test nodes as candidates, and (6) **ZS-SciERC**: zero-shot prediction using SciERC-extracted candidates (Luan et al., 2018). For baselines returning a ranked list of suggestions, we only use the top result.

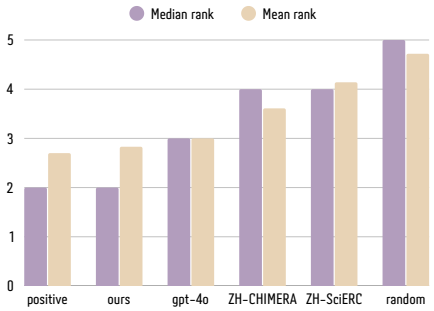


Figure 6: Researchers find our recombination suggestions almost as helpful as the gold answer in inspiring ideas, validating our automated evaluation.

Annotators ranked baseline suggestions by their *helpfulness* in inspiring interesting ideas. Figure 6 reports the median and average rank across 100 examples, where lower values indicate better performance. Our approach receives a similar rank as the gold answer, and annotators prefer it to all other baselines. This gives a complementary signal to the automatic evaluation, showing that our recombination prediction approach learns to create helpful recombinations. Appendix F includes further study details and examples of model predictions that participants found especially inspiring.

Conclusions

We present CHIMERA, a novel knowledge base of 28K+ scientist-authored recombinations, capturing how scientists blend concepts and draw inspiration from different areas. CHIMERA supports a wide range of applications—we show its utility for meta-scientific analysis and for fine-tuning models that predict novel, inspiring recombination directions.

Limitations

Extraction quality As with any automatically-extracted knowledge base, CHIMERA naturally contains some extraction errors (see Appendix B.5). Importantly, our work is the first to explore the new task of extracting recombinations from papers, revealing a gap between the performance of extraction models and humans on the task. As is the case with newly-introduced NLP tasks, future methods trained on our annotated corpus are expected to further improve extraction results, and hence the quality of CHIMERA. However, our analysis shows we already reach good extraction quality overall with minor errors (see Section 4.2), and our downstream applications further demonstrate that the data in CHIMERA can be used to derive utility in scientific meta-analysis and ideation.

Abstract-level scope CHIMERA focuses on extracting recombination instances from scientific abstracts rather than full papers. This design choice, common in scientific IE tasks (Gonzalez et al., 2023; Zhang et al., 2024; Naik et al., 2024), enables more scalable annotation and leverages the fact that abstracts typically summarize key contributions—including conceptual recombinations. In our setting, we focus on capturing cases where a recombination is at the core of a paper's contribution, hence likely to appear in the abstract. Confirming this intuition, we further conduct an analysis that finds that abstracts cover the vast majority of these cases. Extending extraction methods to full papers could reveal additional recombination patterns in future work.

Recombination prediction evaluation As in other open-ended creative tasks (Jentzsch and Kersting, 2023; Meng et al., 2023; Huot et al., 2024), the recombination prediction task admits no single correct answer. Given a problem description, there are many valid ways to blend ideas or draw inspiration, which can lead to false negatives and an overly conservative estimate of model performance. To mitigate this, we conduct a complementary human evaluation. However, due to the expertise required from evaluators, the scale and depth of this assessment are necessarily limited.

Experimenting with additional models Our work leverages a diverse set of models for extraction and prediction, including open-source LLMs (e.g., Mistral-7B, all-mpnet-base), proprietary models (e.g., GPT-4o), and non-generative baselines (e.g., PURE). GPT-4o is used for auxiliary tasks, such as evaluation (judging entity span similarity), analysis (identifying entity's scientific domain), and to enrich our data (generating a context string for the extracted recombinations). As our primary focus is on building and analyzing the recombination knowledge base, we limit our experiments to these models. Exploring a broader range of models for these auxiliary tasks is an important direction for future work.

Ethical Considerations

To collect human-annotated recombination examples, we recruited crowdworkers through the Upwork platform. All annotators were informed in advance about the nature, purpose, and scope of the annotation task. They were compensated fairly for their time, at rates ranging from $26 to $30 per hour. Annotation quality was monitored through overlapping assignments and expert review to ensure reliability and accuracy.

For our human evaluation study, three volunteers with prior research experience participated in ranking model outputs. Participation was entirely voluntary, and no personal or identifying information about the annotators or participants is collected or disclosed.

To support transparency and reproducibility, we release our code, model checkpoints, and the annotated data under an open license. We used AI-based coding assistants (e.g., GitHub Copilot) and language tools for minor code and grammar refinements during development.

References

2019. The cambridge handbook of creativity.

Joel Chan, Katherine Fu, Christian Schunn, Jonathan Cagan, Kristin Wood, and Kenneth Kotovsky. 2011. On the benefits and pitfalls of analogies for innovative design: Ideation performance based on analogical distance, commonness, and modality of examples. *Journal of mechanical design*, 133(8):081004.

Lydia B. Chilton, S. Petridis, and Maneesh Agrawala. 2019. Visiblends: A flexible workflow for visual blends. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

DaEun Choi, Sumin Hong, Jeongeon Park, John Joon Young Chung, and Juho Kim. 2023. Creative-connect: Supporting reference recombination for graphic design ideation with generative ai. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Rosario Delgado and Xavier-Andoni Tibau. 2019. Why cohen's kappa should be avoided as performance measure in classification. *PloS one*, 14(9):e0222916.

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Stasa Milojevic, Alexander Michael Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert Ĺaszló Barabási. 2018. Science of science. *Nature*, 214:1–2.

Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. 2019. Mapping the landscape of creativity support tools in hci. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Felix Frohnert, Xuemei Gu, Mario Krenn, and Evert P L van Nieuwenburg. 2024. Discovering emergent connections in quantum physics research via dynamic word embeddings. *Machine Learning: Science and Technology*, 6.

Aniketh Garikaparthi, Manasi Patwardhan, Lovekesh Vig, and Arman Cohan. 2025. Iris: Interactive research ideation system for accelerating scientific discovery. *Preprint*, arXiv:2504.16728.

Dedre Gentner, Sarah Brem, Ron Ferguson, Philip Wolff, Arthur B Markman, and KD Forbus. 1997. Analogy and creativity in the works of johannes kepler. *Creative thought: An investigation of conceptual structures and processes*, pages 403–459.

Dedre Gentner and Kenneth J. Kurtz. 2005. Relational Categories. In *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin*, APA decade of behavior series. American Psychological Association, Washington, DC, US.

Dedre Gentner and Arthur B Markman. 1997. Structure mapping in analogy and similarity. *American psychologist*, 52(1):45.

Fernando Gonzalez, Zhijing Jin, Bernhard Schölkopf, Tom Hope, Mrinmaya Sachan, and Rada Mihalcea. 2023. Beyond good intentions: Reporting the research landscape of nlp for social good. *arXiv preprint arXiv:2305.05471*.

Moritz Hennen, Florian Babl, and Michaela Geierhos. 2024. Iter: Iterative transformer-based entity recognition and relation extraction. In *Conference on Empirical Methods in Natural Language Processing*.

Keith J. Holyoak and Paul Thagard. 1994. Mental leaps: Analogy in creative thought.

Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 235–243, New York, NY, USA. ACM.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2024. Agents' room: Narrative generation through multi-step collaboration. *ArXiv*, abs/2410.02603.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. Scirex: A challenge dataset for document-level information extraction. In *Annual Meeting of the Association for Computational Linguistics*.

Sophie F. Jentzsch and K. Kersting. 2023. Chatgpt is fun, but it is not funny! humor is still challenging large language models. *ArXiv*, abs/2306.04563.

Uri Katz, Mosh Levy, and Yoav Goldberg. 2024. Knowledge navigator: Llm-guided browsing framework for exploratory search in scientific literature. *Preprint*, arXiv:2408.15836.

G. Knoblich, S. Ohlsson, H. Haider, and D. Rhenius. 1999. Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(6):1534–1555. 00691.

Mario Krenn, Lorenzo Buffoni, Bruno Coutinho, Sagi Eppel, Jacob Gates Foster, Andrew Gritsevskiy, Harlin Lee, Yichao Lu, João P. Moutinho, Nima Sanjabi, Rishi Sonthalia, Ngoc M. Tran, Francisco Valente, Yangxinyu Xie, Rose Yu, and Michael Kopp. 2022. Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. *Nature Machine Intelligence*, 5:1326–1335.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *ArXiv*, abs/1808.09602.

T. McCaffrey. 2012. Innovation Relies on the Obscure: A Key to Overcoming the Classic Problem of Functional Fixedness. *Psychological Science*, 23(3):215–218. 00117.

Céline McKeown. 2014. The cognitive science of science: explanation, discovery, and conceptual change. *Ergonomics*, 57:632 – 633.

Yan Meng, Liangming Pan, Yixin Cao, and Min-Yen Kan. 2023. Followupqg: Towards information-seeking follow-up question generation. In *International Joint Conference on Natural Language Processing*.

William K. Myers, Simon J. George, Yaser Nejaty-Jahromy, James R. Swartz, and R. Britt. 2013. Atypical combinations and scientific impact. In *unknown*.

Aakanksha Naik, Bailey Kuehl, Erin Bransom, Doug Downey, and Tom Hope. 2023. Care: Extracting experimental findings from clinical literature. *ArXiv*, abs/2311.09736.

Aakanksha Naik, Bailey Kuehl, Erin Bransom, Doug Downey, and Tom Hope. 2024. Care: Extracting experimental findings from clinical literature. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4580–4596.

Huitong Pan, Qi Zhang, Cornelia Caragea, Eduard Constantin Dragut, and Longin Jan Latecki. 2024. Scidmt: A large-scale corpus for detecting scientific mentions. In *International Conference on Language Resources and Evaluation*.

Tal Perry. 2021. Lighttag: Text annotation platform. In *Conference on Empirical Methods in Natural Language Processing*.

Aniket Pramanick, Yufang Hou, Saif M. Mohammad, and Iryna Gurevych. 2025. The nature of NLP: Analyzing contributions in NLP papers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25169–25191, Vienna, Austria. Association for Computational Linguistics.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Daniel S Weld, and Tom Hope. 2024. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *arXiv preprint arXiv:2409.14634*.

Tarek Saier and Michael Färber. 2020. unarxive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics*, 125:3085 – 3108.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier López de Lacalle, German Rigau, and Eneko Agirre. 2023. Gollie: Annotation guidelines improve zero-shot information-extraction. *ArXiv*, abs/2310.03668.

Omar Sharif, Joseph Gatto, Madhusudan Basak, and Sarah Masud Preum. 2024. Explicit, implicit, and scattered: Revisiting event extraction to capture complex arguments. In *Conference on Empirical Methods in Natural Language Processing*.

Feng Shi and James Evans. 2023. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications*, 14(1):1641.

Feng Shi and James Allen Evans. 2019. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications*, 14.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Arvind Srinivasan and Joel Chan. 2024. Improving selection of analogical inspirations through chunking and recombination. *Proceedings of the 16th Conference on Creativity & Cognition*.

Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2023. Luminate: Structured generation and exploration of design space with large language models for human-ai co-creation. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *ArXiv*, abs/2304.09542.

You Can Teach, Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. 2020. You can teach an old dog new tricks! on training knowledge graph embeddings. In *International Conference on Learning Representations*.

Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. Atypical combinations and scientific impact. *Science*, 342(6157):468–472.

Jan Philip Wahle, Terry Ruas, Mohamed Abdalla, Bela Gipp, and Saif Mohammad. 2023. We are who we cite: Bridges of influence between natural language processing and other academic fields. *ArXiv*, abs/2310.14870.

Jian Wang, Reinhilde Veugelers, and Paula E. Stephan. 2015. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Ewing Marion Kauffman Foundation Research Paper Series*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. Scimon: Scientific inspiration machines optimized for novelty. *ACL*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Zonglin Yang, Wanhao Liu, Ben Gao, Yujie Liu, Wei Li, Tong Xie, Lidong Bing, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. 2025a. Moose-chem2: Exploring llm limits in fine-grained scientific hypothesis discovery via hierarchical search. *arXiv preprint arXiv:2505.19209*.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2025b. Moosechem: Large language models for rediscovering unseen chemistry scientific hypotheses. *Preprint*, arXiv:2410.07076.

Hyejin Youn, Deborah Strumsky, Luis MA Bettencourt, and José Lobo. 2015. Invention as a combinatorial process: evidence from us patents. *Journal of the Royal Society interface*, 12(106):20150272.

Xingjian Zhang, Yutong Xie, Jin Huang, Jinge Ma, Zhaoying Pan, Qijia Liu, Ziyang Xiong, Tolga Ergen, Dongsub Shim, Honglak Lee, and 1 others. 2024. Massw: A new dataset and benchmark tasks for ai-assisted scientific workflows. *arXiv preprint arXiv:2406.06357*.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *North American Chapter of the Association for Computational Linguistics*.

A Annotator Agreement

Following standard practice in information extraction (Naik et al., 2023; Sharif et al., 2024), we assess inter-annotator agreement using precision, recall, and F1 scores. Agreement is computed by treating one annotator's labels as the reference and the other's as predictions. In addition to measuring entity-level and relation-level agreement, we also evaluate agreement on recombination presence—that is, whether a text expresses a recombination instance, regardless of its type.

We apply the soft entity and relation matching procedure described in Section 4.1 to compute entity and relation agreement. All agreement scores are based on the 49 documents annotated by both annotators (approximately 10% of the full dataset). We treat these agreement measures as a proxy for human-level performance on this task.

A.1 Disagreement Analysis

To better understand the sources of annotation disagreement, we conducted a qualitative analysis. The most common cause stems from differences in identifying whether a recombination is present at all (see examples in Table 7). In such cases, disagreements were resolved via discussion and expert adjudication. The primary criterion for resolution was whether the authors explicitly describe a recombination as contributing to their approach.

Interestingly, once annotators agreed that a recombination was present, they rarely disagreed on its type, and only a single example exhibited this form of conflict. However, disagreements over which entities the recombination includes were more frequent. These typically fell into two categories:

1. **Boundary disagreements**, where annotators selected different spans with overlapping meaning. Here, the expert favored the span that preserved more context (e.g., "*reinforcement learning which uses traditional time series stock price data*" was preferred over "*traditional time series stock price data*").

2. **Conceptual disagreements**, where annotators identified fundamentally different entities. These were resolved through further discussion and clarification.

B Additional Extraction Details

B.1 Recombination keywords

We use keyword-based filtering to identify works that are more likely to discuss recombination before assigning papers to human annotators. Table 8 presents the list of keywords used for this step.

B.2 Extraction baselines implementation

E2E recombination extraction We use `Mistral-7B` as the backbone for our recombination extraction baseline. We fine-tune the model using `mistral-finetune`[6] on a single NVIDIA RTX A6000 48GB GPU over 500 steps. The training was conducted using the default learning rate of $6.e-5$ and weight decay of 0.1. We use a batch size of 1 and a maximum sequence length of 4096 tokens. `mistral-finetune` implements Low-Rank Adaptation of LLM (LoRA), a parameter efficient fine-tuning method (Hu et al., 2021), which we use with the default rank of 64. The evaluation uses the corresponding repository, mistral-inference[7]. We rerun the same experiment using `Llama-3.1-8B` as a backbone, using an additional 500 warm-up steps, a learning rate of $2e-5$ and a weight decay of 0.01. Figure 7 presents the prompt for these experiments.

[6]https://github.com/mistralai/mistral-finetune
[7]https://github.com/mistralai/mistral-inference

Annotators' Disagreement Examples
Abstract: "... This research proposed a framework based on Long Short-Term Memory (LSTM) deep learning network to generate day-ahead hourly temperature forecast... A case study is shown which uses historical in-situ observations and Internet of Things (IoT) observations for New York City, USA. By leveraging the historical air temperature data from in-situ observations, the LSTM model can be exposed to more historical patterns that might not be present in the IoT observations. Meanwhile, by using IoT observations, the spatial resolution of air temperature predictions is significantly improved..." **Annotator 1**: [**Blend**: "*Internet of Things (IoT) observations for New York City, USA*" $\longleftrightarrow$ "*historical air temperature data from in-situ observations*"] **Annotator 2**: [] **Resolution**: Upon expert review, Annotator 1's interpretation was selected, as the authors explicitly describe how the two sources of data serve complementary roles in their method.
Abstract: "...we propose an integrated system that can perform large-scale autonomous flights and real-time semantic mapping in challenging under-canopy environments. We detect and model tree trunks and ground planes from LiDAR data, which are associated across scans and used to constrain robot poses as well as tree trunk models. The autonomous navigation module utilizes a multi-level planning and mapping framework and computes dynamically feasible trajectories that lead the UAV to build a semantic map of the user-defined region of interest in a computationally and storage efficient manner. A drift-compensation mechanism is designed to minimize the odometry drift using semantic SLAM outputs in real time, while maintaining planner optimality and controller stability..." **Annotator 1**: [**Blend**: "*LiDAR data*" $\longleftrightarrow$ "*a multi-level planning and mapping framework*"] **Annotator 2**: [] **Resolution**: Annotator 2's judgment was selected after expert review, as the relation between the two components is not clearly described as a recombination.

Table 7: Examples of annotation disagreements and resolutions by expert review.

Recombination keywords					
combines	analogies	aggregate	intermingle	unify	blending
combined	equivalence	aggregation	intermingling	unification	blends
combine	equivalent	align	join	weave	blend
combination	reduction	alignment	joining	weaving	blends
combinations	reframing	amalgamate	juxtapose	hybrid	merge
combining	reframe	amalgamation	juxtaposition	merge	merges
mixing	reformulating	assemble	link	merges	unites
mixture	casting	assembling	linkage	merging	analogy
mix	cast	associate	meld	merged	analogize
mixed	casts	association	melding	conflation	analogies
integrates	viewing	bond	mesh	couple	equivalence
integrating	viewed	bonding	meshing	unite	equivalent
integrate	view	bridge	perceive	unites	correlate
integrated	inspire	bridging	perception	interplay	correlation
connection	inspired	coalesce	relate	interconnect	envision
synergy	inspiration	coalescence	relation	harmonize	envisioning
fusion	inspires	compose	splice	harmony	harmonize
fuses	inspiring	composition	splicing	incorporate	harmony
unify	interconnect	incorporation	synthesis	reduction	synthesis
aggregate	align	inspiring	inspire	couple	conjunction
aggregation	reframing	inspiration	fuse	unite	conjoin
alignment	reframe	inspires	synthesis		

Table 8: Recombination keywords. We use a predefined list of keywords to identify works that are more likely to discuss idea recombination.

You are an AI assistant tasked with analyzing scientific abstracts for idea recombination. Your goal is to identify the most salient recombination in the given abstract and format it as a JSON string. Follow these instructions carefully:

1. First, familiarize yourself with the possible entity types for recombinations:

<entity_types>
combination-element: An idea, method, model, technique, or approach combined in the text with other elements.
inspiration-source: A concept, idea, problem, approach, or domain the authors drew inspiration from.
inspiration-target: A concept, idea, problem, approach, or domain in which the authors utilize the inspiration they drew from the inspiration source.
</entity_types>

2. Now, carefully read the following scientific abstract: <abstract>{TEXT}</abstract>

3. Your task is to extract the most salient recombination from this abstract. A recombination can be either:
a) Combination: The authors combine two or more ideas, methods, models, techniques, or approaches to obtain a certain goal.
b) Inspiration: The authors draw inspiration or similarities from one concept, idea, problem, approach, or domain and implement it in another.

4. After identifying the recombination, you will format it as a JSON string in the following structure:

<recombination>{recombination_type: {entity_type_1: [ent_1, ent_2], entity_type_2: [ent_3],...}}</recombination>

If you don't think the text discusses a recombination, or that the recombination is not a central part of the work, return an empty JSON object: {}.

5. Before providing your final answer, use the following scratchpad to think through the process:

<scratchpad>
1. Identify the main ideas, methods, or approaches discussed in the abstract.
2. Determine if there is a clear combination of ideas or if one idea inspired the application in another domain.
3. Identify the specific entities involved in the recombination.
4. Classify the entities according to the provided entity types.
5. Determine the recombination type (combination or inspiration).
</scratchpad>

6. Now, provide your final output in the specified JSON format. Ensure that the output is a valid JSON string. If the output is empty, return {}. Place your answer within <answer> tags.

Remember to carefully analyze the abstract and only identify a recombination if it is clearly present and central to the work described.

Figure 7: E2E extraction prompt. {TEXT} is the placeholder for the input abstract text.

```
@dataclass
class Inspiration(Template):
 """An inspiration describes drawing inspiration or similarities from one concept, idea, problem,
  approach, or domain and implementing it in another. For example, taking inspiration from the human brain to
 design a learning algorithm, performing a reduction from one problem to another, or using a technique from one
 domain in another."""

 inspiration_src: str # The source of the inspiration (e.g., the human brain)
 inspiration_target: str # The target of the inspiration (e.g., a learning algorithm)

@dataclass
class Combination(Template):
 """A combination describes joining two ideas, methods, models, techniques to obtain a certain goal. For example,
 combining two models to improve performance, combining two methods to solve a problem, or combining two ideas to
 create a new concept."""

 comb_element_1: str # The first element of the combination (e.g., model A)
 comb_element_2: str # The second element of the combination (e.g., model B)
```

Figure 8: GoLLIE guidelines.

In addition to fine-tuning LLMs on our data, we experiment with GoLLIE (Sainz et al., 2023), a general IE model fine-tuned to follow any annotation guidelines in a zero-shot fashion. We apply GollIE-13B on our data, using a single NVIDIA RTX A6000 48GB GPU, 1-beam search, and limit the new token number to 128. GoLLIE is finetuned from CODE-LLaMA2, and receives guidelines in the form of data classes describing what objects and properties the model should extract. Figure 8 depicts the guidelines we used to test GoLLIE as an E2E recombination extraction model. In the rare cases where the model returns more than a single recombination type (< 10), we select the first.

We also experiment with GPT-4o in few-shot settings. We select 45 examples for each example type (*blend*, *inspiration*, *not-present*) from the training data (a total of 135). As Table 2 describes, the training set only has 45 *inspiration* examples (as opposed to > 100 *blend* and *not-present* examples). 45 is, therefore, the maximal number of examples per class we can sample while keeping the ICL set balanced. We run each experiment 5 times, sampling a new set of few-shot examples in each, and report the average. Figure 9 presents the prompt for this experiment.

Specialized baselines The recombination extraction model has to execute multiple tasks at once (classifying the document, extracting entities, inferring relations), which might be more challenging than performing them separately. To explore this question, we examine our model classification and extraction abilities against designated models for each task. We use Mistral-7B as a specialized classifier and experiment with two versions of the training data. The first includes binary responses (*present*, *not-present*), while the other contains a short CoT-style analysis string as well as the gold class. We construct the analysis string by incorporating the human entity annotations into predetermined templates (e.g., "*This paper discusses a recombination since the authors take inspiration from [inspiration-source] and implement it in [inspiration-target]*").

To evaluate entity extraction, we compare our model against GPT-4o in few-shot settings and include 45 cases per example type, similarly to the E2E experiment. To account for variability due to example selection, we run each experiment 5 times, sampling a new set of few-shot examples in each, and report the average. The total cost of this process sums up to 50$. The prompt template for this experiment is available on Figure 10.

We experiment with non-generative approaches as well, and compare our model to a SciBERT (Zhong and Chen, 2021) based token classifier. The encoder uses a standard Hugging-Face implementation of SciBERT, which we train on a single NVIDIA RTX A6000 48GB GPU over 500 steps. We use a weight decay of 0.1, a learning rate of $6.e-5$ and a batch size of 1. We also experiment with PURE (Zhong and Chen, 2021), a well-known information extraction baseline. We finetune PURE over our train set using the default parameters, except for max_span_length, which we set to 40 to accommodate for the longer entities in our data.

B.3 E2E vs Specialized extraction

This section reflects on the results described in Section 4, drawing on implementation details of the baselines (described in Appendix B.2). In Section 4, we observe that narrowing the focus to a smaller portion of the recombination extraction task does not always improve performance - in fact, it can lead to worse results. This pattern emerges across three Mistral-based classifiers: the end-to-end version (E2E), the specialized version (Abstract-classifier), and the specialized version trained with synthetic CoT strings (Abstract-classifier-CoT). We hypothesize that identifying recombination relations in text may be analogous to Chain-of-Thought prompting (CoT), a technique known to enhance LLM performance across various tasks (Wei et al., 2022). This hypothesis is

You are an AI assistant tasked with analyzing scientific abstracts for idea recombination. Your goal is to identify the most salient recombination in a given abstract and format it as a JSON string. Follow these instructions carefully:

1. First, familiarize yourself with the possible entity types for recombinations:

<entity_types>
comb-element: An idea, method, model, technique, or approach combined in the text with other elements.
inspiration-src: A concept, idea, problem, approach, or domain the authors drew inspiration from.
inspiration-target: A concept, idea, problem, approach, or domain in which the authors utilize the inspiration they drew from the inspiration source.
</entity_types>

2. Review the following examples to understand the expected output format and the process of identifying recombinations:

<examples>{EXAMPLES}</examples>

3. Now, carefully read the following scientific abstract: <abstract>{TEXT}</abstract>

4. Your task is to extract the most salient recombination from this abstract. A recombination can be either:
a) Combination: The authors combine two or more ideas, methods, models, techniques, or approaches to obtain a certain goal.
b) Inspiration: The authors draw inspiration or similarities from one concept, idea, problem, approach, or domain and implement it in another.

5. After identifying the recombination, you will format it as a JSON string in the following structure:

<recombination>{recombination_type: {entity_type_1: [ent_1, ent_2], entity_type_2: [ent_3],...}}</recombination>

If you don't think the text discusses a recombination, or that the recombination is not a central part of the work, return an empty JSON object: {}.

6. Before providing your final answer, use the following scratchpad to think through the process:

<scratchpad>
1. Identify the main ideas, methods, or approaches discussed in the abstract.
2. Determine if there is a clear combination of ideas or if one idea inspired the application in another domain.
3. Identify the specific entities involved in the recombination.
4. Classify the entities according to the provided entity types.
5. Determine the recombination type (combination or inspiration).
</scratchpad>

7. Now, provide your final output in the specified JSON format. Ensure that the output is a valid JSON string. If the output is empty, return {}. Place your answer within <recombination> tags.

Remember to carefully analyze the abstract and only identify a recombination if it is clearly present and central to the work described.

Figure 9: E2E ICL prompt. {TEXT} is a placeholder for the abstract text, and {EXAMPLES} for the ICL examples.

```
You are tasked with identifying specific types of entities in a given scientific abstract. The entity types you need to identify are:

1. comb-element: An idea, method, model, technique, or approach combined in the text with other elements.
2. inspiration-src: A concept, idea, problem, approach, or domain the authors drew inspiration from.
3. inspiration-target: A concept, idea, problem, approach, or domain in which the authors utilize the inspiration they drew from the inspiration source.

Here is the text you need to analyze:

<text>{TEXT}</text>

Please read the text carefully and identify all entities that belong to the types listed above. Pay close attention to the context and relationships between concepts to accurately categorize each entity.

After identifying the entities, you should output them in a valid JSON format. Use the entity types as keys and lists of entities as values. For example:

{"comb-element": ["entity1", "entity2"],
 "inspiration-src": ["entity3"],
 "inspiration-target": ["entity4", "entity5"]}

Ensure that your JSON output is valid:
- Use double quotes around strings
- Do not include a trailing comma after the last item in a list or object
- Escape any double quotes that appear within entity names

Enclose your final JSON output in <output_json> tags.

Remember to review your output for accuracy and completeness before submitting your final answer.
```

Figure 10: Entity extraction prompt. {TEXT} is a placeholder for the input abstract.

```
You are tasked with comparing two spans extracted from a scientific text to determine if they discuss the same {ENTITY_TYPE}. Follow these instructions carefully:

1. First, read the full text for context:
<full_text>{TEXT}</full_text>

2. Now, consider these two spans extracted from the text above:
<span1>{SPAN1}</span1>
<span2>{SPAN2}</span2>

3. Your task is to carefully analyze these two spans and determine if they discuss the same {ENTITY_TYPE}. The idea the spans discuss should be exactly the same, up to minor lexical or semantic variations.

4. In your analysis, consider the following:
 a. The main topic or idea presented in each span
 b. The context in which these spans appear in the full text
 c. Any potential contradictions between the spans

5. After your analysis, provide a justification for your determination. Explain your reasoning clearly, referencing specific elements from the spans and the full text if necessary.

6. Based on your analysis and justification, provide a "Yes" or "No" answer to whether the spans discuss the same {ENTITY_TYPE}.

7. Present your response in the following format:
 <justification>[Your detailed justification here]</justification>
 <answer>[Your "Yes" or "No" answer here]</answer>
```

Figure 11: Span similarity prompt. {ENTITY_TYPE} is either "combination-element", "inspiration-source" or "inspiration-target". {TEXT} is a placeholder for the paper's abstract. {SPAN1}, {SPAN2} are placeholders for the compared spans.

supported by the superior performance of Abstract-classifier-CoT compared to its non-CoT counterpart.

B.4 Span similarity

We provide our span similarity prompt in Figure B.4. We use it in the extraction evaluation process as discussed in Section 4.1. To mitigate position bias, we query the model twice per pair with reversed orderings, accepting a match only if both judgments are positive. We prefer `GPT-4o-mini` over `GPT-4o` based on a comparison which found only 3 disagreements across the test set.

B.5 Extraction error analysis

We perform analysis over the test set, revealing different sources of error which may inspire future improvements. Our focus is on understanding how different types of input texts can influence the result, specifically, in cases where the extraction model struggles. We use our best-performing fine-tuned E2E model for this analysis.

Context dependent or subtle phrasing We observe that, unsurprisingly, cases in which the recombination is implied or subtle are more challenging for the model. For instance (see also Table 9, row 1), *"Kahneman & Tversky's prospect theory"* inspires the design of a loss function that *"directly maximizes the utility of generations"*, but this is not stated directly. Moreover, abstracts that express idea recombination while referencing previously mentioned entities are also harder to detect.

Multiple recombinations Some papers present a salient recombination along with other insignificant ones. We notice that in those cases, the model might extract a non-salient recombination or mix multiple ones (see Table 9, row 2 for such a case).

Borderline cases The role of a recombination as a core element in the work is sometimes debatable. Table 9, row 3 presents an example of such a case where the authors explicitly mention integrating *"embedding space comparison"* with *"computational notebook environment"*, which may be interpreted as a recombination (the usage of notebook in these environments is completely new and novel), or simply as a way to present the tool's environment. We notice that the extraction model

Bad extraction examples (human annotated test set)
Abstract: "...Kahneman & Tversky's prospect theory tells us that humans perceive random variables in a biased but well-defined manner (1992) ... Using a Kahneman-Tversky model of human utility, we propose a HALO [Human Aware Loss Function] that directly maximizes the utility of generations instead of maximizing the log-likelihood of preferences, as current methods do..." **Gold** = [**Inspiration**: "*Kahneman & Tversky's prospect theory*" $\longrightarrow$ "*a HALO*"] **Pred** = []
Abstract: "...We address the problem by proposing a Wasserstein GAN combined with a new reverse mask operator, namely Reverse Masking Network (R-MNet), a perceptual adversarial network for image inpainting ... Additionally, we propose a new loss function computed in feature space to target only valid pixels combined with adversarial training..." **Gold** = [**Blend**: "*a Wasserstein GAN*" $\longleftrightarrow$ "*...R-MNet*"] **Pred** = [**Blend**: "*a Wasserstein GAN*" $\longleftrightarrow$ "*...R-MNet*" $\longleftrightarrow$ "*a new loss function*"]
Abstract: "... In order to characterize model flaws and choose a desirable representation, model builders often need to compare across multiple embedding spaces, a challenging analytical task supported by few existing tools. We first interviewed nine embedding experts in a variety of fields to characterize the diverse challenges they face and techniques they use when analyzing embedding spaces. Informed by these perspectives, we developed a novel system called Emblaze that integrates embedding space comparison within a computational notebook environment..." **Gold** = [**Blend**: "*embedding space comparison*" $\longleftrightarrow$ "*...notebook environment*"] **Pred** = []

Table 9: In the first row, the extraction model misses an inspiration relation because of subtle phrasing. In the second row, when analyzing an abstract with multiple recombinations, the model fails to identify the most important one and confuses entities across different relations. In the third row, the model fails to detect a weak recombination example.

tends to miss those cases.

B.6 Extraction examples

Table 10 presents examples of interdisciplinary, automatically extracted inspiration recombinations.

B.7 Large-scale extraction assessment

You are tasked with reviewing outputs from an information extraction system that processes scientific abstracts. Your job is to assess whether a particular extracted relation and its associated entities are both meaningful and accurate, according to strict scientific criteria. Below is the information you will use for your review:

<abstract> {{ABSTRACT}} </abstract>
The information extraction model has produced the following output:

Relation type: {{EXTRACTED_RELATION}}
Entity 1: {{ENTITY1}}
Entity 2: {{ENTITY2}}

Possible relation types are:

combination: A combination indicates that the authors describe joining together two or more ideas, methods, models, techniques, or approaches to achieve a specified goal.
inspiration: Inspiration means the authors describe taking inspiration, analogy, abstraction, reduction, reformulation, or similarities from one concept, idea, problem, approach, or domain and applying it to another.
Please evaluate the model's extraction according to the two criteria below:

Criterion 1: Do the extracted entities (Entity 1 and Entity 2) each represent clear, meaningful scientific concepts, methods, models, problems, or approaches?

Criterion 2: Is the stated relation (combination or inspiration) one that described in the abstract as occurring between these two entities?

Provide your answer by filling in the YES/NO values for each criterion below:

<answer> { "1": "[YES/NO]", "2": "[YES/NO]" } </answer>
Replace [YES/NO] with your judgment for each criterion. Make sure your response is a valid JSON object and fits the format exactly. Do not provide additional explanation or commentary outside of the JSON object.

Figure 12: Large-scale evaluation prompt. {ABSTRACT} is a placeholder for the original abstract text. {EXTRACTED_RELATION}, {ENTITY1}, and {ENTITY2} are placeholders for the relation type and entities extracted by our model.

To complement our human annotation efforts and enable large-scale evaluation, we conducted a qualitative assessment of the automatically extracted recombination examples in CHIMERA using GPT-4.1 as an LLM-based judge.

Validating the LLM Judge. We first assessed GPT-4.1's reliability by comparing its judgments against those of a domain expert. A PhD student with NLP expertise manually reviewed 100 randomly sampled recombination examples and labeled each as correct if: (1) the extracted entities corresponded to meaningful scientific concepts, and (2) the relation between them captured a central recombination explicitly described in the abstract. Upon analyzing the identified extraction errors, we observe a significant portion stems from extracting correct recombinations with uninformative entities (criteria 2) and not from a conceptual misunderstanding of the text. We provide examples of such cases in Table 11.

GPT-4.1 was prompted with the same examples using an evaluation template aligned with the assessment criteria (see Figure 12). Given the imbalance nature of the data, we report the F1 score instead of Cohen's κ, following the recommendations of previous work (Delgado and Tibau, 2019). The resulting F1 score of 0.912 indicates substantial agreement, supporting the use of GPT-4.1 as a reliable proxy for large-scale quality assessment.

Large-Scale Evaluation. Following validation, we applied GPT-4.1 to a larger sample of 2,000 automatically extracted examples from CHIMERA. The model labeled 799 of these examples as correct, resulting in an estimated extraction accuracy of 80.55%. These results provide further evidence for the overall quality and robustness of our extraction pipeline.

C Graph nodes domains

We identify the scientific domain of each entity using `GPT-4o` in a zero-shot setting. Given the abstract and the extracted recombination entities, the model assigns to each entity an *arXiv category* and a broader *scientific branch*. If the model successfully assigns an arXiv category, we treat it as the entity's domain. Otherwise, the model selects a branch from a predefined list of outer-arXiv domains (see Table 12) and sets it as the domain. If neither a standard arXiv category nor a branch can be assigned, the entity is labeled as belonging to the `Other` domain.

Entities in the `Other` domain are excluded from the analysis in Section 4.3, as they are often too noisy, overly broad, or miscellaneous to interpret reliably. Figures 13 and 14 present the prompts used for analyzing blend and inspiration relations, respectively. Running this classification process over the full corpus cost approximately 250$.

The `Other` domain We assign the `Other` domain to nodes the model fails to classify. In total, 2,127 graph nodes fall into this category. We manually examined a sample of 150 such nodes and found that many were either too ambiguous or too general to categorize meaningfully. Interestingly, some of

Detected Inspiration	**Abstract**
Inspiration-Source: "the shepherding behavior of herding dogs" [zoology] *Inspiration-Target*: "Frontier exploration" [cs.ro]	"Efficient exploration of large-scale environments remains a critical challenge in robotics... The presented bio-inspired framework heuristically models **frontier exploration** similar to the **shepherding behavior of herding dogs**. This is achieved by modeling frontiers as a sheep swarm reacting to robots modeled as shepherding dogs..."
Inspiration-Source: "the multi-granular diagnostic approach of pathologists" [biomedical sciences] *Inspiration-Target*: "Histopathological image classification" [cs.cv]	"**Histopathological image classification** constitutes a pivotal task in computer-aided diagnostics... In the diagnostic process of pathologists, a multi-tiered approach is typically employed to assess abnormalities in cell regions at different magnifications...Inspired by **the multi-granular diagnostic approach of pathologists**, we perform feature extraction on cell structures at coarse, medium, and fine granularity, enabling the model to fully harness the information in histopathological images..."
Inspiration-Source: "the dual-process theory in psychology" [psychology] *Inspiration-Target*: "enhance dialogue planning in LLMs" [cs.cl]	"...Traditional approaches to **enhance dialogue planning in LLMs**, ... either face efficiency issues or deliver suboptimal performance. Inspired by **the dualprocess theory in psychology**, which identifies two distinct modes of thinking - intuitive (fast) and analytical (slow), we propose the Dual-Process Dialogue Planning (DPDP) framework. DPDP embodies this theory through two complementary planning systems: an instinctive policy model for familiar contexts and a deliberative Monte Carlo Tree Search (MCTS) mechanism for complex, novel scenarios..."
Inspiration-Source: "the Global Workspace Theory in conscious processing" [cognitive science] *Inspiration-Target*: "learning effective feature embeddings for CTR prediction" [cs.lg]	"**Click-Through Rate (CTR) prediction is a pivotal task in product and content recommendation, where learning effective feature embeddings is of great significance**....inspired by **the Global Workspace Theory in conscious processing**, which posits that only a specific subset of the product features are pertinent while the rest can be noisy and even detrimental to human-click behaviors, we propose a CTR model that enables Dynamic Embedding Learning with Truncated Conscious Attention for CTR prediction, termed DELTA..."

Table 10: Inter-domain inspiration examples from the CHIMERA knowledge graph.

Bad Extraction Examples (arXiv)
Abstract: "The detection of allusive text reuse is particularly challenging due to the sparse evidence on which allusive references rely—commonly based on none or very few shared words. Arguably, lexical semantics can be resorted to since uncovering semantic relations between words has the potential to increase the support underlying the allusion and alleviate the lexical sparsity. A further obstacle is the lack of evaluation benchmark corpora, largely due to the highly interpretative character of the annotation process. In the present paper, we aim to elucidate the feasibility of automated allusion detection. We approach the matter from an Information Retrieval perspective in which referencing texts act as queries and referenced texts as relevant documents to be retrieved, and estimate the difficulty of benchmark corpus compilation by a novel inter-annotator agreement study on query segmentation..." ✗ **Automatic extraction** (incorrect entities): [**Inspiration**: "In an Information Retrieval perspective, referencing texts act as queries and referenced texts as relevant documents to be retrieved" ⟶ **"a task-oriented dialog system"**]
Abstract: "Supervised deep learning with pixel-wise training labels has great successes on multi-person part segmentation. However, data labeling at pixel-level is very expensive. To solve the problem, people have been exploring to use synthetic data...the results are much worse compared to those using real data and manual labeling. The degradation of the performance is mainly due to the domain gap, i.e., the discrepancy of the pixel value statistics between real and synthetic data. In this paper, we observe that real and synthetic humans both have a skeleton (pose) representation. We found that the skeletons can effectively bridge the synthetic and real domains during the training. Our proposed approach takes advantage of the rich and realistic variations of the real data and the easily obtainable labels of the synthetic data to learn multi-person part segmentation on real images without any human-annotated labels... " ✗ **Automatic extraction** (uninformative entities): [**Combination**: **"real"** ⟷ **"synthetic humans"**]
Abstract: "...In this paper we propose an LLM feature-based framework for dialogue constructiveness assessment that combines the strengths of feature-based and neural approaches, while mitigating their downsides. The framework first defines a set of dataset-independent and interpretable linguistic features, which can be extracted by both prompting an LLM and simple heuristics. Such features are then used to train LLM feature-based models...We also find that the LLM feature-based model learns more robust prediction rules instead of relying on superficial shortcuts, which often trouble neural models." ✗ **Automatic extraction** (uninformative entities): [**Combination**: **"neural"** ⟷ **"feature-based"**]

Table 11: Representative examples of bad automatic extraction. Many errors stem from uninformative entity spans, as presented by the two bottom examples.

Non-arXiv scientific domains		
Agricultural Science	Anatomy	Animal Science
Anthropology	Archaeology	Behavioral Science
Biochemistry	Bioinformatics	Bioclimatology
Biomedical Engineering	Biophysics	Biotechnology
Botany	Cardiology	Chemical Engineering
Civil Engineering	Clinical Psychology	Cognitive Science
Criminology	Cryosphere Science	Cytology
Demography	Dentistry	Dermatology
Developmental Biology	Ecology	Ecotoxicology
Economics	Educational Psychology	Electrical Engineering
Emergency Medicine	Endocrinology	Energy Science
Engineering Science	Entomology	Environmental Engineering
Environmental Science	Epidemiology	Ethology
Food Science	Forestry	Gastroenterology
Genetics	Genomics	Geography
Geology	Geophysics	Glaciology
Health Informatics	Histopathology	Hydrodynamics
Hydrogeology	Hydrology	Immunogenetics
Immunology	Industrial/Organizational Psychology	Landscape Architecture
Linguistics	Marine Biology	Materials Science
Mechanical Engineering	Medical Microbiology	Meteorology
Microbiology	Mineralogy	Molecular Biology
Mycology	Nanotechnology	Neurology
Neuroscience	Nuclear Engineering	Nutritional Science
Obstetrics	Oceanography	Oncology
Ophthalmology	Ornithology	Orthopedics
Otology	Paleoclimatology	Paleontology
Pathobiology	Pathology	Pediatric Medicine
Pedagogy	Petrology	Pharmacogenomics
Pharmacology	Philosophy	Physiology
Political Science	Proteomics	Psychiatry
Psychology	Psychopathology	Public Health
Pulmonology	Radiology	Rheumatology
Seismology	Social Psychology	Sociology
Surgery	Systems Biology	Thermodynamics
Toxicology	Urban Planning	Urology
Veterinary Science	Virology	Volcanology
Wildlife Biology	Zoology	

Table 12: Non-arXiv scientific domains. We complement arXiv category taxonomy using a broader list of scientific fields.

You are an AI assistant tasked with analyzing a scientific abstract to determine the arXiv categories and scientific branches of combined elements. Your goal is to identify the most appropriate arxiv taxonomy category and most suitable scientific domain for each element provided.

Here is the abstract you will be analyzing:
<abstract>{ABSTRACT}</abstract>

And here is the list of combined elements identified from the abstract:
<elements>{ELEMENTS}</elements>

Here is a list of the standard arXiv categories:
<arxiv>{ARXIV}</arxiv>

And here is a list of scientific branches:
<branches>{BRANCHES}</branches>

For each element in the list, you need to:
1. Identify the best matching arXiv taxonomy category from the provided list. If it doesn't match any category, use "other". If there's insufficient information, use "insufficient-info".
2. Identify the scientific branch from the provided branches list. If there's insufficient information, use "insufficient-info". If no branch name in the list describes the source properly, use "other".

Return your output in the following format:
<output>
[{"text": "element1",
 "arxiv_category": "category1",
 "scientific_branch": "branch1"},
 {"text": "element2",
 "arxiv_category": "category2",
 "scientific_branch": "branch2"}, ...]
</output>

Format your response as a valid JSON string.

Now, analyze the provided elements from the abstract and generate your response in the specified JSON format. Make sure to include all elements from the provided list, and ensure that your output is properly formatted as a valid JSON string.

Figure 13: blend domain analysis prompt. {ELEMENTS} is a placeholder for the recombination entities extracted from {ABSTRACT}. {ARXIV} is a placeholder for full arXiv category names and their descriptions. {BRANCHES} is a placeholder for the list of non-arXiv domains given in Appendix C, Table 12.

You will be analyzing the scientific branches and arXiv taxonomy categories of an inspiration source and target based on an abstract from a scientific paper. Here's the information you'll be working with:

<abstract>{ABSTRACT}</abstract>

<inspiration_source>{INSPIRATION_SOURCE}</inspiration_source>

<inspiration_target>{INSPIRATION_TARGET}</inspiration_target>

<arxiv>{ARXIV}</arxiv>

<branches>{BRANCHES}</branches>

Your task is to identify the arXiv taxonomy category and most suitable scientific branch for both the inspiration source and the inspiration target.

For the inspiration source:
1. Identify the best matching arXiv taxonomy category from the provided list. If it doesn't match any category, use "other". If there's insufficient information, use "insufficient-info".
2. Identify the scientific branch from the provided branches list. If there's insufficient information, use "insufficient-info". If no branch name in the list describes the source properly, use "other".

Repeat the same process for the inspiration target.

Provide your analysis in the following format:

<source-branch>[Insert the scientific branch of the inspiration source here]</source-branch>
<source-arXiv>[Insert the arXiv taxonomy category of the inspiration source here]</source-arXiv>
<target-branch>[Insert the scientific branch of the inspiration target here]</target-branch>
<target-arXiv>[Insert the arXiv taxonomy category of the inspiration target here]</target-arXiv>

Ensure that you only include the requested information within each tag, without any additional explanation or reasoning.

Figure 14: inspiration domain analysis prompt. {INSPIRATION_SOURCE} and {INSPIRATION_TARGET} are placeholders for the inspiration entities extracted from {ABSTRACT}. {ARXIV} is a placeholder for full arXiv category names and their descriptions. {BRANCHES} is a placeholder for the list of non-arXiv domains given in Appendix C, Table 12.

Type	**Examples**
Non-Academic	"the snap-through action of a steel hairclip", "yoga", "origami, the traditional Japanese paper-folding technique, is a powerful metaphor for design and fabrication of reconfigurable structures", "Tangram, a game that requires replicating an abstract pattern from seven dissected shapes"
Noisy	"a deep", "word-", "at the context level", "a neural part", "post", "text–audio", "end-to-end multi-modal model only X-VLM only X-VLM only X-VLM only X-VLM only X-VLM only X-VLM only X-VLM only X-VLM only X-VLMs", "a user's long-term"
Overly-general	"human experiences", "a styling method", "local search method", "a pipeline inspired by experts' work", "a new modality", "feature based approaches"
Misclassified	"Reinforcement learning, or RL", "Facial Expressions Recognition(FER)", "a Kullback-Liebler regularization function", "K-nearest neighbors algorithm", "Shapley values from game theory", "Gaussian Stochastic Weight Averaging"

Table 13: Examples of graph nodes in the "other" domain. We analyze a sample of 150 nodes in this domain and identify groups with common traits, as shown in the table.

these nodes refer to non-academic or highly niche concepts (see examples in Table 13).

domain grouping To avoid sparsity, we group similar domains as displayed in Table 14. Table 15 presents the node distribution of common domains after applying this grouping process.

D Additional Knowledge Base Analysis

D.1 Predominant recombination relations

We provide a tabular version of Figure 3 in Section 4.3 on Table 16 for better readability.

D.2 Nuanced recombination types

In Section 3, we defined the two broad recombination types used in CHIMERA: *blends* and *inspirations*. In this section, we demonstrate that this taxonomy is both robust and expressive, offering broad coverage of more nuanced recombination phenomena.

To support this, we perform a qualitative analysis of 30 *inspiration* examples from the CHIMERA dataset. We identify distinct subtypes of inspiration, such as *analogy*, *metaphor*, *reduction*, *abstraction*, and *application of existing knowledge*. These subtypes emerge naturally within our current schema, illustrating its extensibility and broad coverage. Table 17 provides examples for each. This analysis lays the groundwork for future refinement and expansion of the taxonomy.

E Additional Prediction Details

E.1 Prediction data preprocessing

Context extraction and leakage filtering We use GPT-4o-mini to extract a few sentences from each abstract describing the background or motivation of the authors using recombination (See prompt on Figure 15). Adding these contexts to the queries helps them be more specific and limits the search space. However, this might introduce leaks into the queries - cases where the extracted context reveals the answer. Table 18 presents leak examples. We utilize GPT-4o-mini again to filter out such cases from the data, using the prompt shown in Figure 16. In a qualitative analysis of 50 randomly sampled query-answer pairs, we find that a human annotator agrees with 87% of the model's predictions (whether there is a leak). Finally, we divide the remaining query-answer pairs into splits as described in Table 5 is Section 5.

E.2 Prediction baselines

We use a bi-encoder architecture for recombination prediction and experiment with three popular encoders as backbones: `all-mpnet-base-v2` (109M parameters), `bge-large-en-v1.5` (Xiao et al., 2023) (335M parameters) and `e5-large-v2` (Wang et al., 2022) (335M parameters). These models' checkpoints predate 2024, meaning they are unfamiliar with our test set. The model receives a query string composed of a context description, a graph entity, and a relation type and returns a ranked list of answers (other graph nodes). We perform HPO (random grid search of 10 trails) to select the number of training epochs, warmup ratio and learning rate for each model. We use contrastive loss and generate 30 negatives per positive example. Following the literature standard (Teach et al., 2020), we report metrics in the filtered settings to avoid false negatives. Given the difficulty of the task we focus on ranking only the 12751 test set entities. A full summary of our data splits is available on 5. The examples we use to train and evaluate our prediction models contain all collected nodes, including those classified as belonging to the "other" domain.

We utilize RankGPT (Sun et al., 2023) as a strong reranker and apply it to rerank the top-20 predicted results. We employ RankGPT with GPT-4o, a window size of 10 and a step size of 5. Note the information cutoff of GPT-4o is October 2023 [8], meaning it is unfamiliar with our test set as well. We use the implementation available in [9]. However, we find that adjusting the default prompt works better for our task. Figure 17 shows the modified reranking prompt. The cost of applying the reranker to our data was 60$.

E.3 Reranker error analysis

In Section 5.1, we show that reranking the top-20 answers retrieved by our best-performing prediction model (all-mpnet-base-v2$_{\text{finetuned}}$) can sometimes **lower** the rank of the gold candidate. To better understand the underlying causes of such reranking failures, we conduct an error analysis of 30 representative cases. Our goal in this section is to describe common patterns in these errors and highlight particularly challenging scenarios that may inform future progress.

[8] As stated in https://platform.openai.com/docs/models/gpt-4o

[9] https://github.com/sunnweiwei/RankGPT/tree/main

You are tasked with extracting the rationale behind the selection of a specific methodology in a scientific study. You will be provided with an abstract and a statement about the methodology used. Your goal is to extract the reasons for choosing this methodology from the abstract.

First, carefully read the following abstract: <abstract>{{ABSTRACT}}</abstract>

Next, inspect the following examples of background descriptions:

1. Large language models (LLMs) commonly employ autoregressive generation during inference, leading to high memory bandwidth demand and consequently extended latency.
2. Reconstructing deformable tissues from endoscopic videos is essential in many downstream surgical applications. However, existing methods suffer from slow rendering speed, greatly limiting their practical use.
3. Many industrial tasks-such as sanding, installing fasteners, and wire harnessing-are difficult to automate due to task complexity and variability.
4. Multi-legged robots offer enhanced stability in complex terrains, yet autonomously learning natural and robust motions in such environments remains challenging.

Now, consider this methodology statement: <methodology_statement>{{METHODOLOGY_STATEMENT}}</methodology_statement>

To complete this task, follow these steps:

1. Analyze the abstract thoroughly, focusing on:
 - The context or reasons that justify the methodology choice
 - Any challenges, limitations, or research needs the methodology addresses
 - Mentions of previous research or knowledge gaps that the methodology aims to target

2. When formulating your response:
 - Phrase your response as a general 1-2 sentence description of a challenge, limitation research needs, etc.
 - Use exclusively the information from the abstract. Do not incorporate external knowledge or assumptions.
 - Minimize including information from the methodology statement in your answer.
 - Do not include information about the used methodology in your answer.
 - If the background details are unclear, return an empty response.

3. Format your response as follows:
 <background>
 [1-2 background sentences]
 </background>

Remember to base your response strictly on the provided abstract and statement. Do not include additional information or assumptions.

Figure 15: Context extraction prompt. {{ABSTRACT}} is a placeholder for the input abstract. {{METHODOLOGY_STATEMENT}} is a sentence describing the recombination. We build it by filling one of the following templates with the extracted recombination entities: "*Combine <source-entity> and <target-entity>*" for blends and "*Take inspiration from <source-entity> and apply it to <target-entity>*" for inspirtions.

Group	Scientific domains
Geosciences	Geology, Geophysics, Petrology, Mineralogy, Hydrology, Hydrogeology, Seismology, Volcanology, Cryosphere Science, Glaciology, Geography
Environmental Sciences	Environmental Science, Environmental Engineering, Ecology, Ecotoxicology
Biomedical Sciences	Biochemistry, Immunology, Immunogenetics, Neuroscience, Oncology, Pathology, Pathobiology, Pharmacology, Toxicology
Health and Medicine	Cardiology, Neurology, Urology, Gastroenterology, Obstetrics, Pediatric Medicine, Rheumatology, Dermatology, Ophthalmology, Otology, Pulmonology, Emergency Medicine, Surgery, Radiology, Orthopedics, Psychiatry, Dentistry, Public Health, Epidemiology, Health Informatics, Clinical Psychology, Psychopathology
Zoology	Zoology, Entomology, Ornithology, Wildlife Biology, Animal Science, Veterinary Science, Ethology
Agriculture	Agricultural Science, Forestry
Food Sciences	Nutritional Science, Food Science
Psychology	Educational Psychology, Social Psychology, Psychology, Industrial/Organizational Psychology
Microbiology	Microbiology, Medical Microbiology
Humanities	Linguistics, Philosophy, Pedagogy
Social Sciences	Sociology, Anthropology, Political Science, Demography

Table 14: Scientific domains grouped by category. We group similar non-arXiv scientific domains (see Table 12) to thicken infrequent ones.

```
You are an AI assistant tasked with identifying potential leakages in a given query. A leakage occurs when a query reveals or implies the answer. Follow these steps carefully:

1. Read the following query: <query>{{QUERY}}</query>
2. Now, read the corresponding answer: <answer>{{ANSWER}}</answer>
3. Analyze the query for any information that might disclose the answer. Look for words, phrases, or implications in the query that directly relate or reveal information from the answer.

4. Write your analysis in the following format:
<analysis>
[If you identified a leakage, briefly explain what information from the answer is included in the query. If you did not identify a leakage, write "no leakage".]
</analysis>

5. Based on your analysis, determine if there is a leakage.

6. Provide your response in the following format:
<leakage>
[Write "yes" if there is a leakage, or "no" if there is no leakage. Do not include any additional explanation or reasoning.]
</leakage>

Remember, your task is to identify leakages, not to answer the query or explain your reasoning. Stick strictly to the output format provided.
```

Figure 16: Leak detection prompt.

```
{'role': 'user','content': f"I have a scientific query describing settings and requesting a suggestion. I will provide you with {num} suggestions, each indicated by number identifier [].\nRank the suggestions based on their potential usefulness in addressing the query: {query}."},
{'role': 'assistant', 'content': 'Okay, please provide the passages.'},
...
{'role': 'user', 'content': f"[{rank}] {content}"},
{'role': 'assistant', 'content': f'Received passage [{rank}].'},
...
{'role': 'user', 'content': "Scientific Query: {query}. \nRank the {num} suggestions above based on their potential usefulness in addressing the query. The suggestions should be listed in descending order using identifiers. The most relevant suggestions should be listed first. The output format should be [] > [], e.g., [1] > [2]."}
```

Figure 17: Adjusted RankGPT prompt.

(i) Multiple plausible answers. In some cases, the reranker correctly identifies a strong and highly relevant candidate, and ranks it above the gold even though both answers are valid. These errors stem not from a lack of understanding, but from the presence of several equally reasonable responses. For instance, in Table 19 (top), the reranker promotes a conceptually grounded strategy from game theory over a more generic gold response about rational decision principles.

Domain	Count	Domain	Count	Domain	Count
cs.cv	12504	cs.lg	8440	cs.cl	4697
cs.ro	2241	cs.ai	2091	cognitive science	936
cs.ir	884	cs.ne	864	cs.si	655
cs.hc	645	q-bio.nc	441	cs.ds	409
cs.cg	382	cs.cy	378	cs.gr	367
math.oc	356	eess.iv	278	cs.dm	269
cs.db	254	eess.sp	242	cs.lo	204
cs.ma	203	cs.ce	185	cs.sy	177
cs.cr	164	stat.me	138	cs.gt	132
psychology	116	eess.sy	108	cs.se	104
zoology	101	cs.it	100	math.pr	96
cs.dc	89	behavioral science	88	cs.mm	82
eess.as	79	nlin.ao	79	cs.ar	74
cs.na	66	cs.pl	65	biomedical sciences	63
physics.med-ph	60	stat.ml	56	health and medicine	56
physics.bio-ph	52	cs.ni	48	physics.ao-ph	44
stat.th	43	anatomy	41	math.na	40
math.ds	39	cs.fl	38	humanities	38
q-bio.pe	32	cs.dl	32	cs.sc	30
math-ph	27	cond-mat.stat-mech	25	math.ap	24
math.dg	22	physics.class-ph	22	cs.sd	22
econ.th	21	math.ca	21	math.mg	20
physics.comp-ph	20	physics.optics	20	cs.et	20

Table 15: Node domains distribution. The table presents the number of graph nodes from each domain with above-median frequency.

Inspirations			*Blends*		
Source	**Target**	**Count**	**Source**	**Target**	**Count**
cs.cv	cs.cv	334	cs.cv	cs.cv	4329
cs.lg	cs.cv	300	cs.lg	cs.lg	2793
cognitive science	cs.cv	278	cs.cl	cs.cl	1049
cs.lg	cs.lg	254	cs.lg	cs.cv	992
cognitive science	cs.lg	211	cs.cv	cs.lg	470
cs.cl	cs.cl	190	cs.cl	cs.cv	422
cs.cl	cs.cv	188	cs.cv	cs.cl	391
cognitive science	cs.ai	184	cs.lg	cs.cl	363
cognitive science	cs.cl	142	cs.ro	cs.ro	299
cs.cl	cs.ai	141	cs.ro	cs.cv	218
cs.lg	cs.ai	118	cs.cl	cs.lg	197
q-bio.nc	cs.cv	114	cs.ai	cs.cl	174
q-bio.nc	cs.lg	102	cs.ai	cs.ai	161
cognitive science	cs.ro	100	cs.ai	cs.lg	151
cs.cv	cs.lg	94	cs.lg	cs.ai	146
cs.lg	cs.cl	84	cs.lg	cs.ne	133
cs.cl	cs.lg	84	cs.ir	cs.ir	132
math.oc	cs.lg	83	cs.lg	cs.ro	124
zoology	cs.ro	76			

Table 16: Predominant inspiration and blend relations. The above is a tabular version of Figures 3b, 3a in Section 4.3. It presents edges with (source-domain, target-domain) pairs frequency above the 0.98 quantile.

(ii) Semantically similar variants. Another common error involves the reranker prioritizing paraphrased or reformulated versions of the gold answer. While these candidates are semantically close to the gold, the gold itself may fall in rank due to redundancy. As shown in Table 19 (bottom), several variants of "Direct Preference Optimization" receive high rankings, but the original mention of the method is pushed downward, possibly due to lexical overlap penalties or insufficient canonicalization.

These examples highlight nuanced challenges in reranking systems, such as handling redundancy and conceptual equivalence.

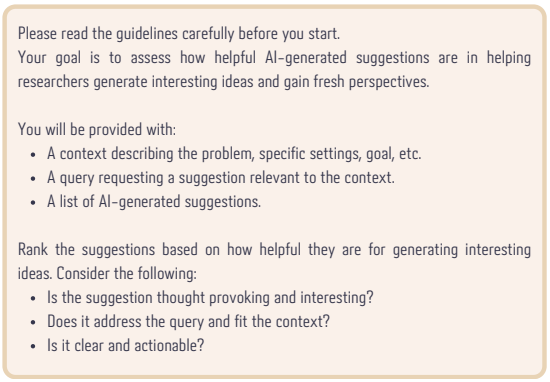


Figure 18: User study guidelines.

F User study additional details

We request each to fill out a form asking in what scientific domains they feel comfortable reading papers and a short description of their research area. We then used `granite-embedding-125m-english` to retrieve semantically similar contexts to this description from the relevant arXiv categories. We manually verify that the retrieved contexts match the description and discard examples with poorly extracted information (e.g., the context begins with "*This study reviews the problem of...*" instead of directly describing the source study problem). In addition, we let the volunteers mark an example as "ill-defined", in which case we ignore their inputs. We conduct a 10-minute training session with each volunteer, requesting them to read the instructions and explain the task. Figure 18 presents the instructions given to the participants in the study. Figure 19 presents the web interface of the annotation platform.

Nuanced recombination types examples
Reduction
Abstract: "Register allocation is one of the most important problems for modern compilers...This work demonstrates the use of casting the register allocation problem as a graph coloring problem..."
Inspiration: "*a graph coloring problem*" $\longrightarrow$ "*Register allocation*"
Metaphor
Abstract: "Affective sharing within groups strengthens coordination and empathy...we propose HeartBees, a bio-feedback system for visualizing collective emotional states, which maps a multi-dimensional emotion model into a metaphorical visualization of flocks of birds..."
Inspiration: "*flocks of birds*" $\longrightarrow$ "*a multi-dimensional emotion model*"
Analogy
Abstract: "Physics-informed Graph Neural Networks have achieved remarkable performance...by mitigating common GNN challenges...Despite these advancements, the development of a simple yet effective paradigm that appropriately integrates previous methods for handling all these challenges is still underway. In this paper, we draw an analogy between the propagation of GNNs and particle systems in physics, proposing a model-agnostic enhancement framework..."
Inspiration: "*particle systems in physics*" $\longrightarrow$ "*the propagation of GNNs*"
Application of existing knowledge
Abstract: "Object detection in high-resolution satellite imagery is emerging as a scalable alternative to on-the-ground survey data collection...However, performing object detection over large geographies can still be prohibitively expensive due to the high cost of purchasing imagery and compute. Inspired by traditional survey data collection strategies, we propose an approach to estimate object count statistics over large geographies through sampling..."
Inspiration: "*traditional survey data collection strategies*" $\longrightarrow$ "*Object detection in high-resolution satellite imagery*"
Abstraction
Abstract: "While visual question-answering (VQA) benchmarks have catalyzed the development of reasoning techniques, they have focused on vertical thinking. Effective problem-solving also necessitates lateral thinking...To bridge this gap, we formulate visual lateral thinking as a multiple-choice question-answering task..."
Inspiration: "*a multiple-choice question-answering task*" $\longrightarrow$ "*visual lateral thinking*"

Table 17: Examples of nuanced inspiration types found within CHIMERA. While all examples are labeled as *inspiration*, they illustrate finer-grained mechanisms such as metaphor, reduction, and analogy. This suggests that our taxonomy is expressive enough to capture a rich diversity of recombination strategies.

Query	Answer
Understanding the human brain's processing capabilities can inspire advancements in machine learning algorithms and architectures. Previous methods in brain research were limited to identifying regions of interest for one subject at a time, restricting their applicability and scalability across multiple subjects. What would be a good source of inspiration for "**a highly efficient processing unit**"?	The human brain
Existing models for link prediction in knowledge graphs primarily focus on **representing triplets in either distance or semantic space**, which limits their ability to fully capture the information of head and tail entities and utilize hierarchical level information effectively. **This indicates a need for improved methods that can leverage both types of information** for better representation learning in knowledge graphs. What could we blend with "**distance measurement space**" to address the described settings?	Semantic measurement space

Table 18: Leakages examples. Examples of leaks - queries that reveal or strongly imply the answer.

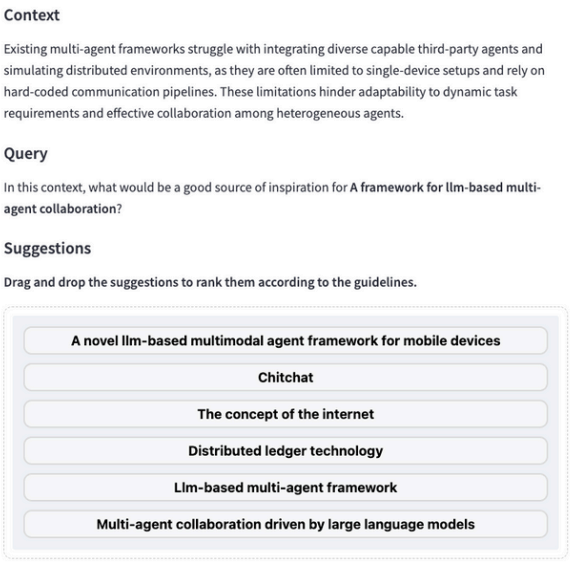


Figure 19: User study interface.

F.1 Predictions examples

Table 20 shows a selection of model predictions that participants rated as most helpful for inspiring research directions. These examples highlight how CHIMERA-trained models can move beyond surface-level associations to propose insightful cross-domain inspirations, for instance, linking harmful meme detection to visual commonsense reasoning, or drawing on neuroscience to improve LLM knowledge retention. Such predictions demonstrate CHIMERA's potential to power ideation tools that help researchers identify novel, actionable directions for future work.

G Comparison to other information extraction methods

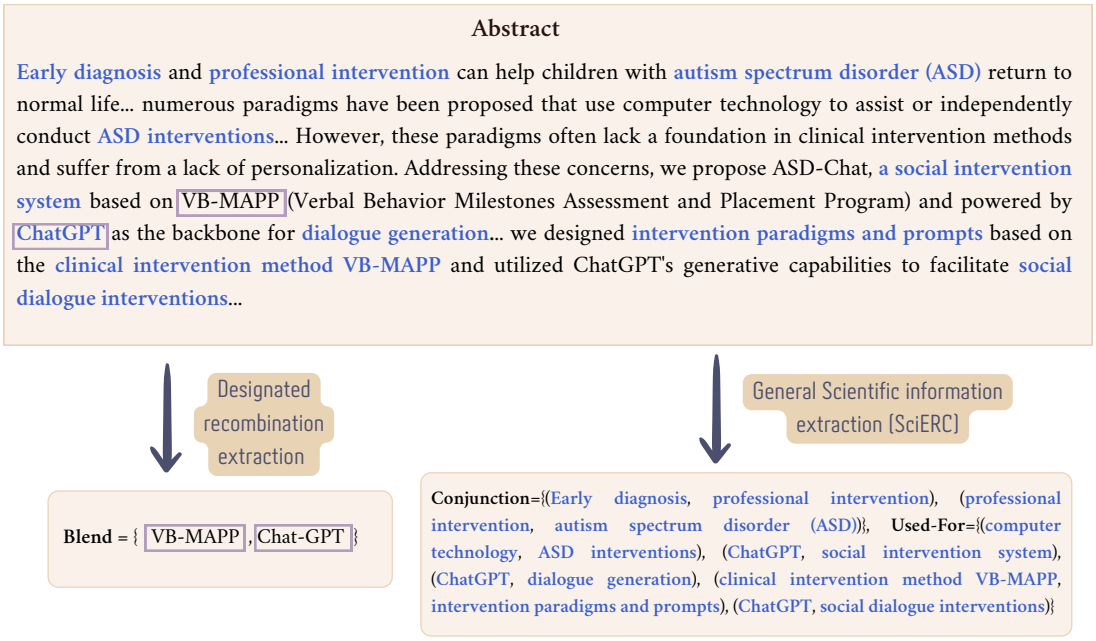
Both general scientific extraction and concept co-occurrence struggle to capture concise and accurate recombination relations, as can be seen in Figure 20. Figure 20a presents how general scientific IE schemas lack relation types to model recombinations. The figure presents the results of our specialized extraction method besides a transformer-based extraction model (Hennen et al., 2024) fine-tuned on SciERC (Luan et al., 2018), a general IE schema. While our new data schema easily models the recombinant connection between two techniques: "*BV-MAPP (Verbal Behavior Milestones Assessment and Placement Program)*", "*ChatGPT*"

Reranking error examples	
(i) Multiple plausible answers	
Query: "Traditional reasoning methods in language models often rely on historical information and employ a uni-directional reasoning strategy... This leads to suboptimal decision-making... What would be a good source of inspiration for *enhancing the decision rationality of language models*?"	
Pre-reranking (top-20)	*Post-reranking (top-20)*
1. the inherent human attribute of engaging in logical reasoning to facilitate decision-making 2. principles of rational decision-making 3. the Level-K framework from game theory and behavioral economics, which extends reasoning from simple reactions to structured strategic depth ...	1. the Level-K framework from game theory and behavioral economics, which extends reasoning from simple reactions to structured strategic depth 2. Bayesian inference: conditioning a prior on evidence ... 6. principles of rational decision-making
Query: "...while Large Language Models (LLMs) excel in various NLP tasks, their ability to generate comprehensive data stories remains underexplored... What would be a good source of inspiration for *Data-driven storytelling*?"	
Pre-reranking (top-20)	*Post-reranking (top-20)*
1. the human storytelling process 2. story writing 3. Interactive digital stories ... 9. narrative structure designs ...	1. story analysis and generation systems 2. generative artificial intelligence (Gen-AI)-driven narrative personalization 3. narrative structure designs 4. the human storytelling process ...
(ii) Semantically similar variants	
Query: "Prior methods for aligning large language models face challenges in tuning to maximize non-differentiable and non-binary objectives...This highlights a need for a more flexible approach that can generalize to various user preferences... while maintaining alignment... What could we blend with *reinforcement learning via human feedback* to address the described settings?"	
Pre-reranking (top-20)	*Post-reranking (top-20)*
1. aligning Large Language Models with human preferences 2. Direct Preference Optimization for preference alignment 3. direct preference optimization ... 5. State-of-the-art language model fine-tuning techniques, such as Direct Preference Optimization ...	1. Direct Preference Optimization for preference alignment 2. State-of-the-art language model fine-tuning techniques, such as Direct Preference Optimization 3. contrastive learning-based methods like Direct Preference Optimization 4. a Semi-Policy Preference Optimization method 5. direct preference optimization ...

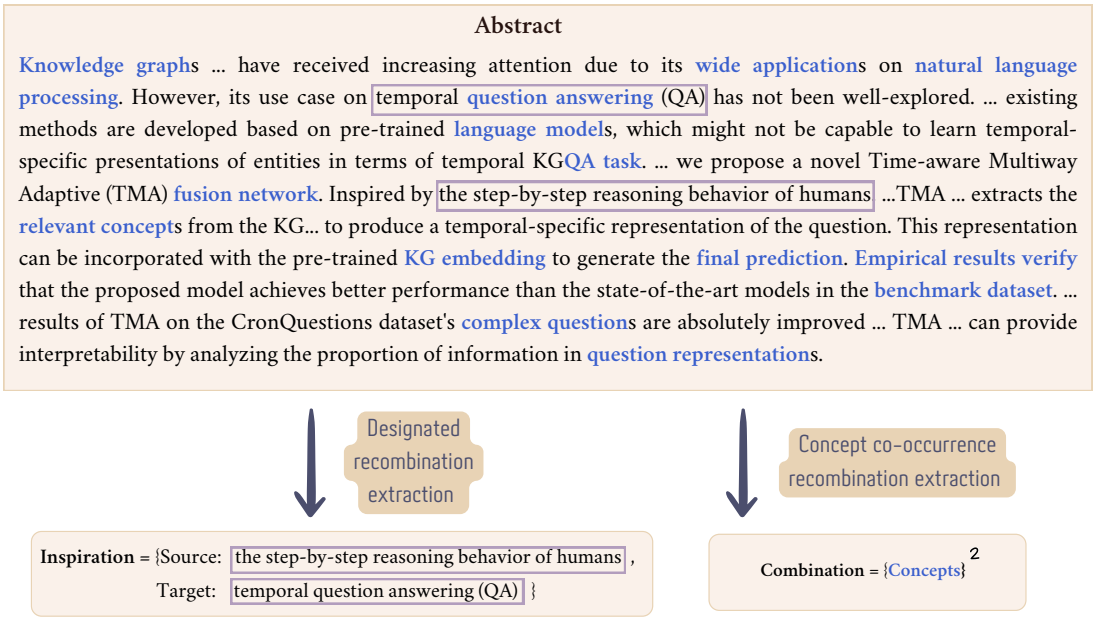
Table 19: Illustrative examples where the reranker preferred a different answer over the gold one.

Prediction Examples
Query: Existing large language models (LLMs) underperform in legal judgment prediction due to challenges in understanding case complexities and distinguishing between similar charges. This highlights a need for improved methodologies that can effectively address these issues to enhance judicial efficiency. In this context, what would be a good source of inspiration for *legal judgment prediction*? **Prediction**: The Issue, Rule, Application, Conclusion (Issue, Rule, Application, Conclusion) framework, a sequential reasoning approach used by lawyers.
Query: The exponential increase in scientific literature makes it challenging for researchers to stay current with recent advances and identify meaningful research directions. Existing methods for idea generation either trivially prompt large language models or directly expose them to extensive literature without indicating useful information, highlighting a need for more effective approaches to research ideation. In this context, what would be a good source of inspiration for *automating the generation of novel research ideas*? **Prediction**: Emulate the human process of inducing collective creativity through engaging discussions with participants from diverse backgrounds and perspectives.
Query:The challenge of detecting harmful memes lies in the implicit meanings embedded within them, which are not explicitly conveyed through surface text and images. Existing harmful meme detection methods fail to provide readable explanations that reveal these implicit meanings, highlighting a significant gap in the ability to support detection decisions. In this context, what would be a good source of inspiration for *detecting harmful memes*? **Prediction**: Visual commonsense discovery in computer vision.
Query: Large language models (LLMs) often struggle to provide up-to-date information due to their one-time training and the constantly evolving nature of the world. Existing approaches to keep LLMs current face difficulties in extracting stored knowledge, highlighting a need for improved methods of knowledge acquisition from raw documents. In this context, what would be a good source of inspiration for *improving an llm's ability to effectively acquire new knowledge from raw documents*? **Prediction**: Neuroscience, where the human brain often sheds outdated information to improve the retention of crucial knowledge and facilitate the acquisition of new information.

Table 20: Examples of recombination directions predicted by our model and rated as most inspiring by user study participants. Each prediction links a scientific challenge with a cross-domain concept, illustrating CHIMERA 's potential to support creative research ideation.



(a) Comparison to recombination extraction using a general scientific IE schema (SciERC)



(b) Comparison to recombination extraction using concept co-occurrence.

Figure 20: Comparison of our designate recombination extraction method to alternative approaches. Figure 20a: General recombination extraction schemas lack fitting relation types to capture recombinations, which results in capturing plenty of irrelevant relations ("*Early diagnosis*" $\longleftrightarrow$ "*professional intervention*"). Figure 20b: Recombination extraction using concept co-occurrence might be nonsensical ("*wide application*" $\longleftrightarrow$ "*final prediction*") or even misleading ("*question answering*" $\longleftrightarrow$ "*language models*")).

as a concept *blend*, the SciERC extraction schema isn't equipped with proper relation types for this. As a result, it captures mostly irrelevant information for our task (e.g background details as "*Early diagnosis*" or "*professional intervention*"). Figure 20b shows how recombination extraction using concept co-occurrence might be misleading. In this method, each pair of canonical scientific concepts (e.g, *neural networks*) that co-occur within the same abstract are considered a recombination. The figure presents an example of using AI-related concepts curated by Krenn et al. (2022) for recombination extraction, alongside recombination extracted using our designated approach. Note that when using concept co-occurrence, the extracted recombinations are essentially $\{concepts\}^2$, which might be imprecise, and capture meaningless recombinations (e.g., "*wide application*" recombined with "*final prediction*") or misleading recombinations (e.g., "*question answering*" with "*language models*", which explicitly presented by the authors as a lacking approach for the task). In comparison, our new extraction schema neatly models the main recombiant relation presented in the text as taking inspiration from "*the step-by-step reasoning behavior of humans*" for "*temporal question answering*."