Topic 4. Text-based Content and Comparative Table Generation

Yufang Hou

Professor
Interdisciplinary
Transformation University
yufang.hou@it-u.at

Agenda

- ❑ **Short Text Generation with Citation Grounding**
- ❑ **Automatic Research Paper Writing**
- ❑ **Automated Survey & Deep Research Generation**
- ❑ **Meta-analysis Table Generation and Comparative Literature Synthesis**
- ❑ **Summary**

Cite-worthiness Detection / Citation Recommendation

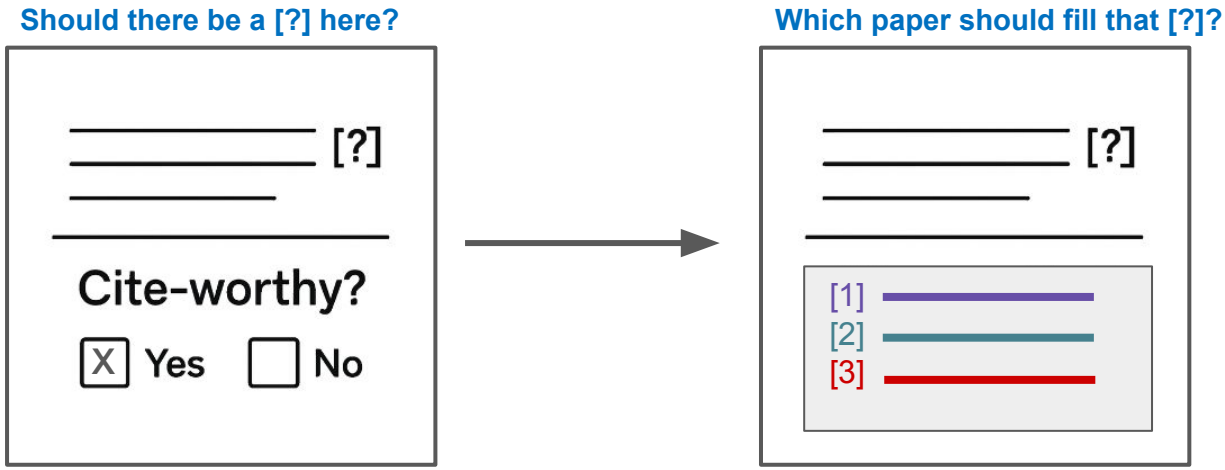
- **Cite-worthiness detection**: given a sentence in a draft manuscript, classify whether it needs a citation [1].

Should there be a [?] here?

[1] Wright & Augenstein, ACL 2021 Findings. CiteWorth: Cite-Worthiness Detection for Improved Scientific Document Understanding.
[2] Çelik & Tekir, EMNLP 2025. CiteBART: Learning to Generate Citations for Local Citation Recommendation.

Cite-worthiness Detection / Citation Recommendation

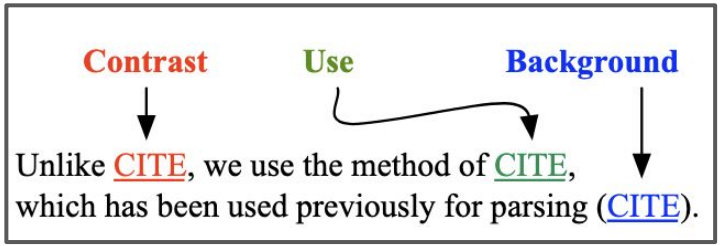
- **Cite-worthiness detection**: given a sentence in a draft manuscript, classify whether it needs a citation [1].
- **Citation recommendation**: given a context that needs a citation (a sentence, paragraph, or full manuscript), retrieve and rank candidate papers to cite, or directly generate citations [2].



[1] Wright & Augenstein, ACL 2021 Findings. CiteWorth: Cite-Worthiness Detection for Improved Scientific Document Understanding.
[2] Çelik & Tekir, EMNLP 2025. CiteBART: Learning to Generate Citations for Local Citation Recommendation.

Citation Analysis: Citation Function

- **Citation function**: classify how scholars use and frame citations [1].

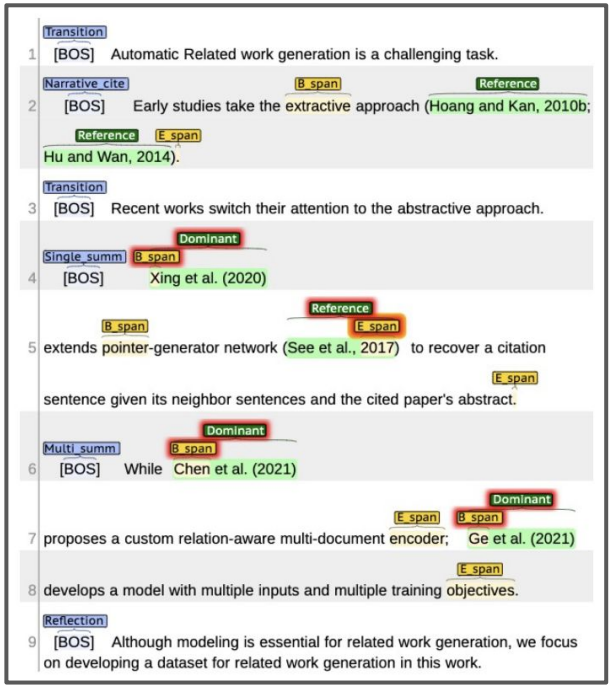


Class	Description	Example
BACKGROUND	P provides relevant information for this domain.	This is often referred to as incorporating deterministic closure (Dörre, 1993).
MOTIVATION	P illustrates need for data, goals, methods, etc.	As shown in Meurers (1994), this is a well-motivated convention [...]
USES	Uses data, methods, etc., from P .	The head words can be automatically extracted [...] in the manner described by Magerman (1994).
EXTENSION	Extends P's data, methods, etc.	[...] we improve a two-dimensional multimodal version of LDA (Andrews et al, 2009) [...]
COMPARISON OR CONTRAST	Expresses similarity/differences to P.	Other approaches use less deep linguistic resources (e.g., POS-tags Stymne (2008)) [...]
FUTURE	P is a potential avenue for future work.	[...] but we plan to do so in the near future using the algorithm of Littlestone and Warmuth (1992).

[1] Jurgen et al., TACL 2018. Measuring the Evolution of a Scientific Field through Citation Frames.

Citation Analysis: CORWA

- **CORWA**: citation oriented related work annotation, which decomposes the related work section with three inter-related annotation tasks [1].



Discourse tagging task tags the role of each related work sentence with one of six labels: {single_summ, multi_summ, narrative_cite, reflection, transition, other}.

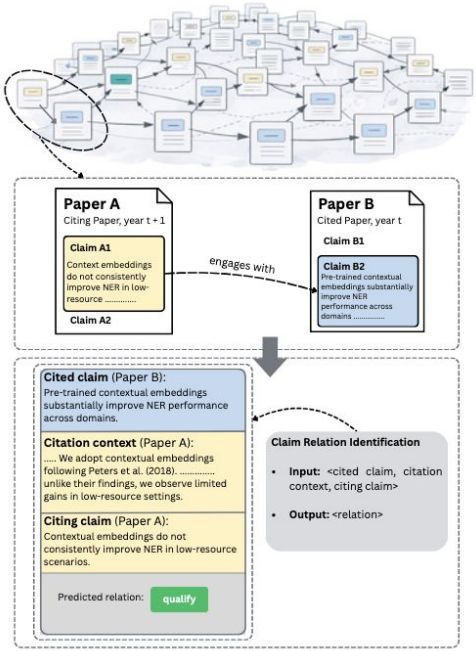
Citation span detection task identifies the span of text whose information is directly derived from a specific cited paper.

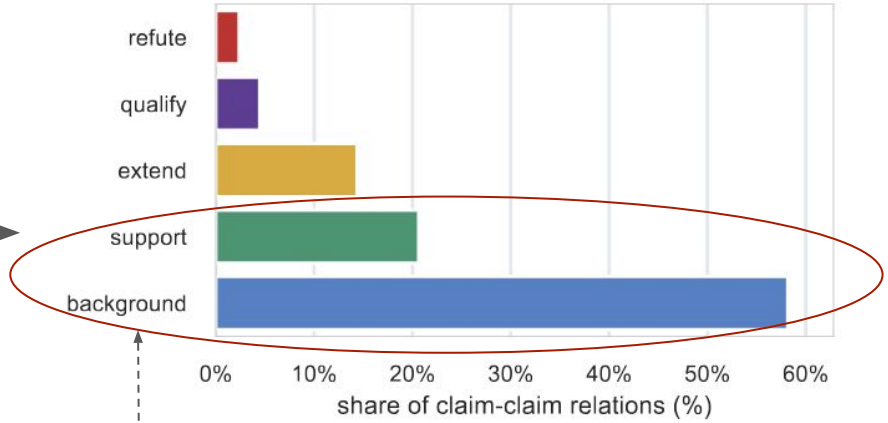
Citation type recognition task indicates whether a cited work is discussed in detail or used to illustrate a high-level concept.

[1] Li et al., NAACL 2022. CORWA: A Citation-Oriented Related Work Annotation Dataset.

Citation Analysis: ClaimFlow

- **Cross-document scientific claim relations classification grounded in citations**: analyze whether a citing paper supports, extends, qualifies, refutes, or references a claim as background [1]



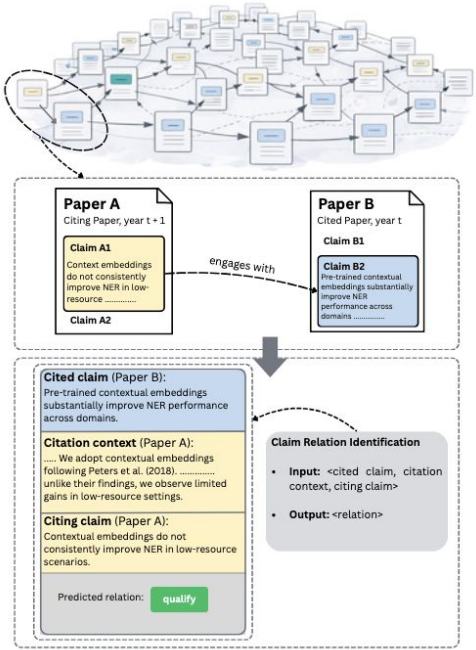


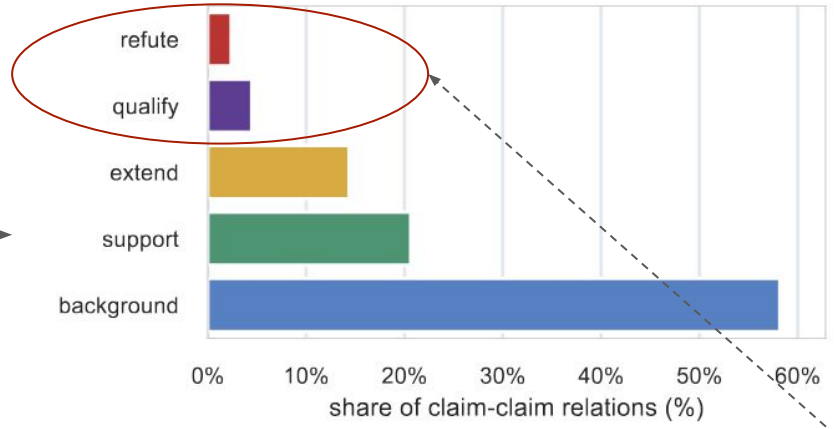
Claims are most often reused as contextual premises or explicitly supported by subsequent work.

[1] Pramanick et al., arXiv 2026. ClaimFlow: Tracing the Evolution of Scientific Claims in NLP.

Citation Analysis: ClaimFlow

- **Cross-document scientific claim relations classification grounded in citations**: analyze whether a citing paper supports, extends, qualifies, refutes, or references a claim as background [1]



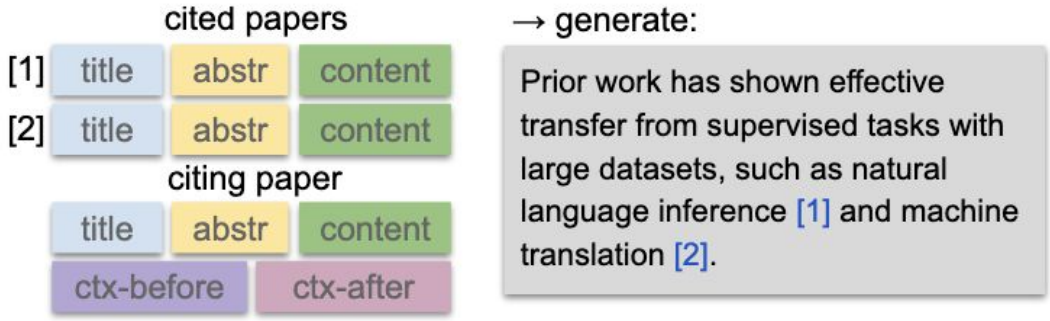


only a small number of claims are ever challenged (11.1%) at least once

[1] Pramanick et al., arXiv 2026. ClaimFlow: Tracing the Evolution of Scientific Claims in NLP.

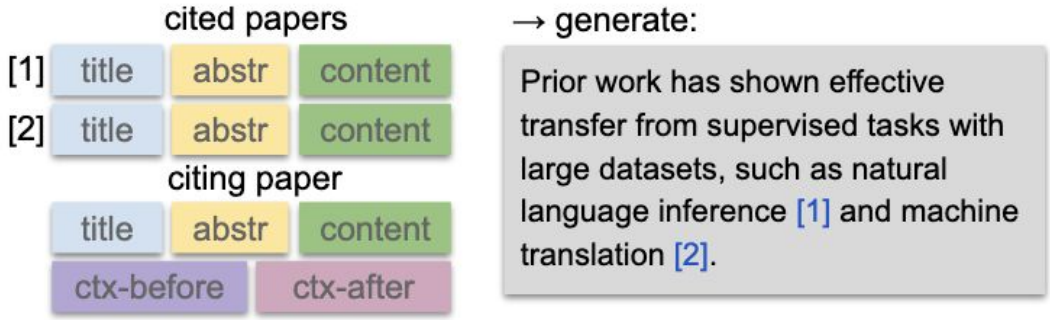
Related Work and Citation Text Generation

- Generating **citation text** for pre-selected cited papers given **the context** of the citing paper



Related Work and Citation Text Generation

- Generating **citation text** for pre-selected cited papers given **the context** of the citing paper



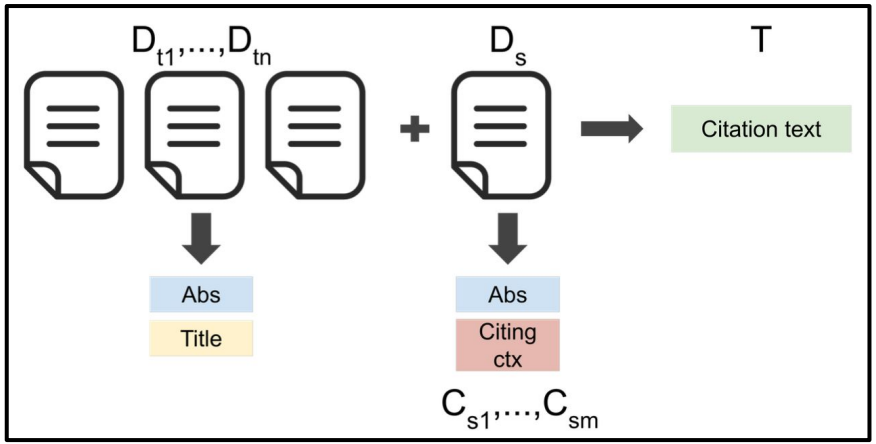
- Task variations

Dataset	Input					Output	
	Cited document (D^t)			Citing context (C^s)		Citation text (T)	
	Single Abs	Multi Abs	Title	Abs	Text	Sent	Para
AbuRa'ed et al. (2020)	✓		✓			✓	
Chen et al. (2021)		✓					✓
Lu et al. (2020)		✓		✓			✓
Xing et al. (2020)	✓			✓	✓	✓	

Related work section generation

CiteBench

- A benchmark for citation text generation[1]
  - **Unified task definition and dataset**
  - **Standardized baselines**: unsupervised, supervised, transfer-based
  - **Evaluation kit:** standard metrics and discourse-based measurements

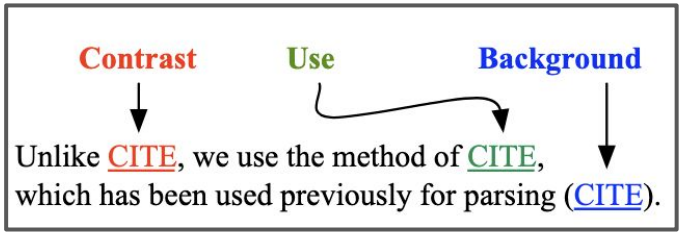


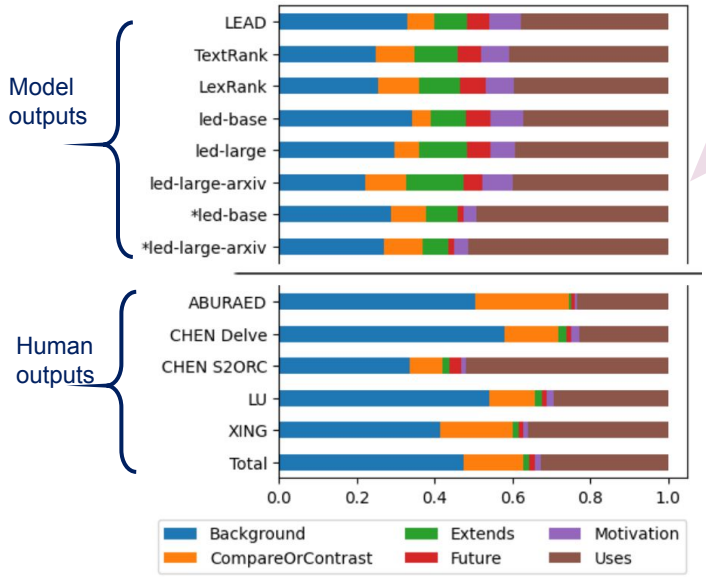
[1] Funkquist et al., EMNLP 2023. CiteBench: A Benchmark for Scientific Citation Text Generation.

CiteBench

- Discourse analysis based on **citation intent**[1] and CORWA (citation oriented related work annotation)

citation intent analysis example





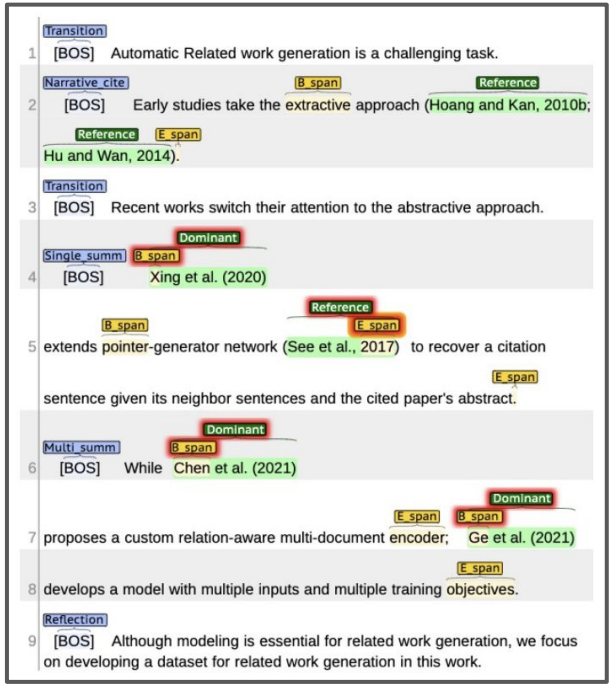
Models tend to under-generate the ***Background*** and ***CompareOrContrast*** sentences, they produce more ***Future***, ***Uses*** and ***Extends*** sentences than the gold reference.

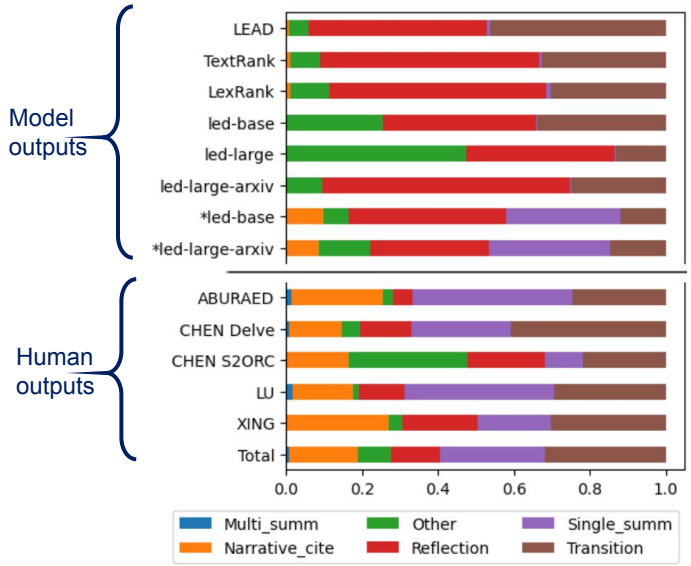
[1] Jurgen et al., TACL 2018. Measuring the Evolution of a Scientific Field through Citation Frames.

CiteBench

- Discourse analysis based on citation intent and **CORWA** (citation oriented related work annotation)[1]

CORWA analysis example

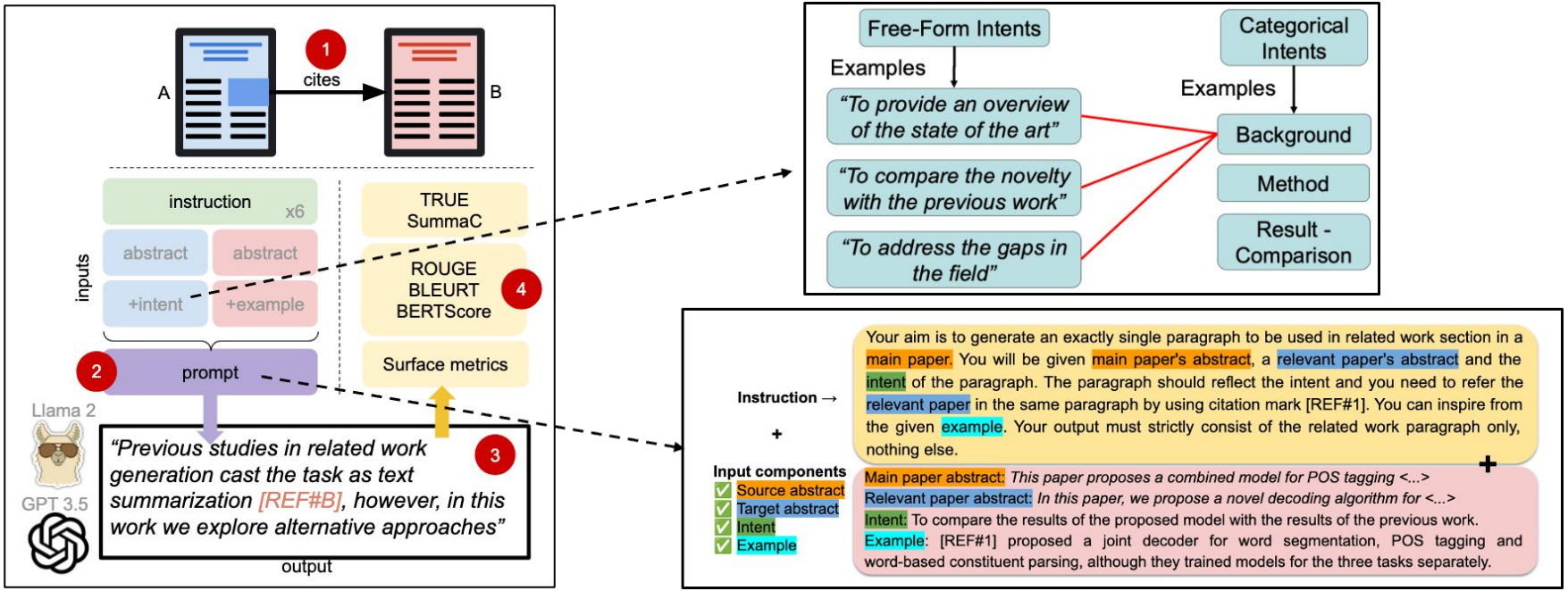




Models tend to under-generate the ***Narrative_cite*** and ***Single_summ*** class, while over-generating the ***Reflection***, compared to the distributions in the gold reference texts.

[1] Li et al., NAACL 2022. CORWA: A Citation-Oriented Related Work Annotation Dataset.

Key Important Input Components for Citation Text Generation

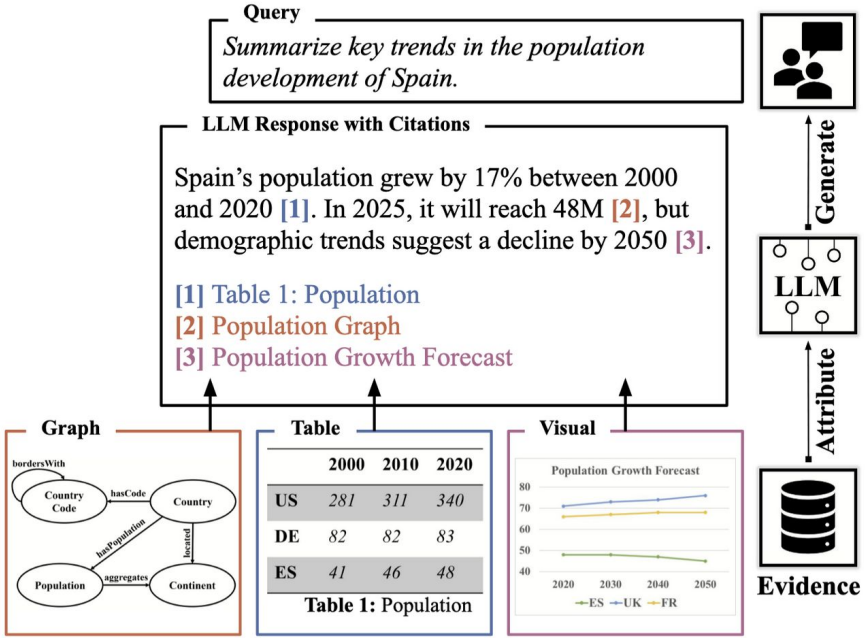


- Joint use of **citation intent** and **example sentences** gives best results for both models and human annotators
- **Free-form** citation intents are more effective than **categorical** intents

[1] Şahinuç et al., ACL 2024. Systematic Task Exploration with LLMs: A Study in Citation Text Generation.

LLM Citation Generation and Attribution

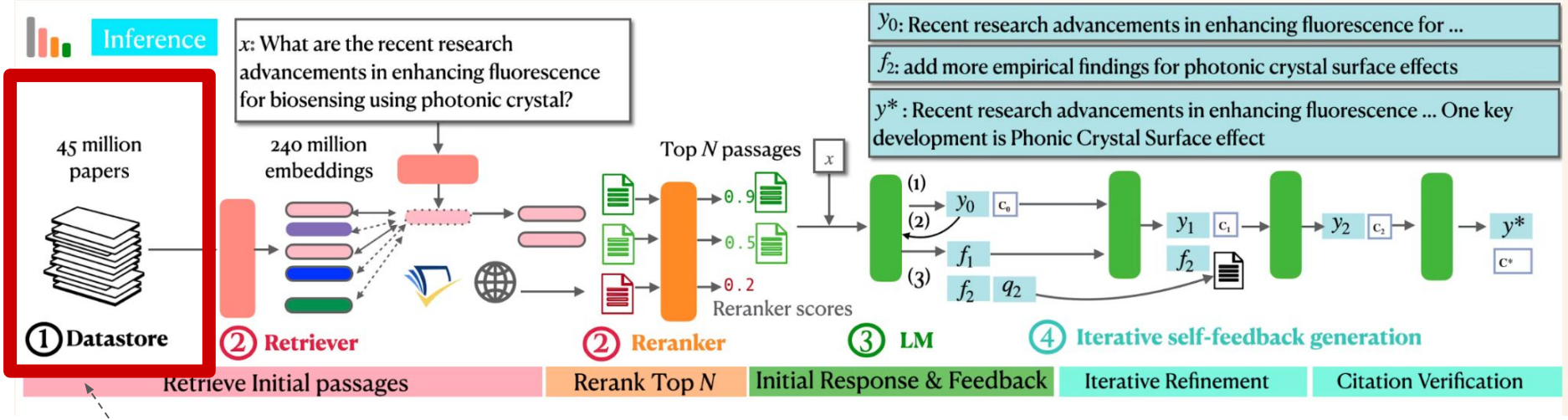
- **Evidence-based text generation**: given a query, LLMs generate a response where each claim is grounded with inline citations to identifiable source passages, making outputs verifiable.



[1] Li et al., arXiv 2023. A Survey of Large Language Models Attribution.
[2] Schreieder et al., arXiv 2025. Attribution, Citation, and Quotation: A Survey of Evidence-based Text Generation with Large Language Models.

OpenScholar

- Inference

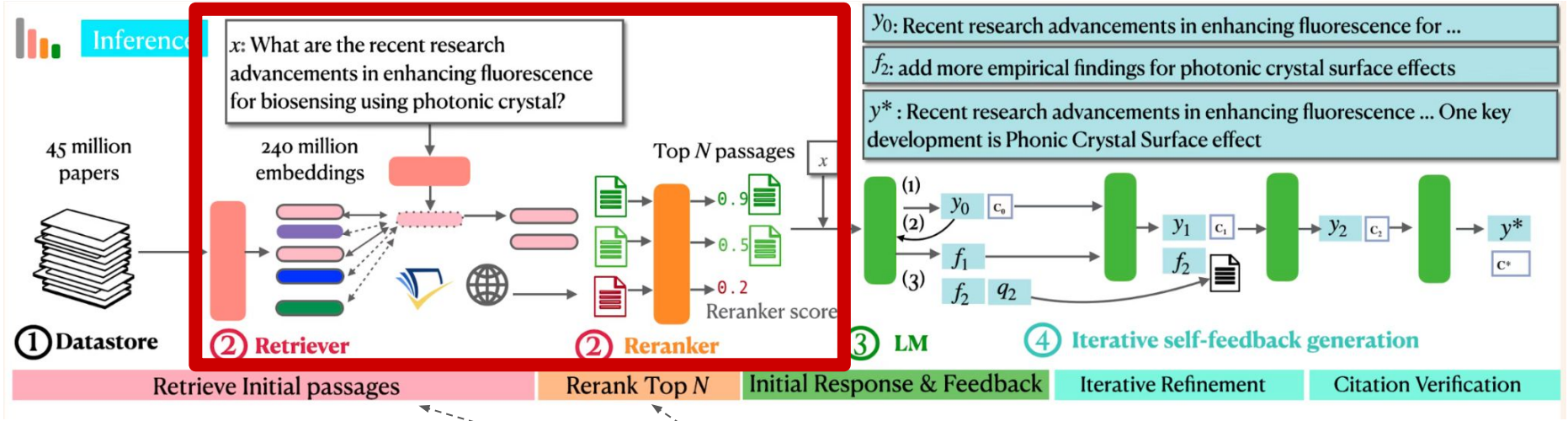


A collection of more than 45M papers from Semantic Scholar and ~250M corresponding passage embeddings.

[1] Asai et al., Nature 2026. Synthesizing scientific literature with retrieval-augmented language models.

OpenScholar

- Inference

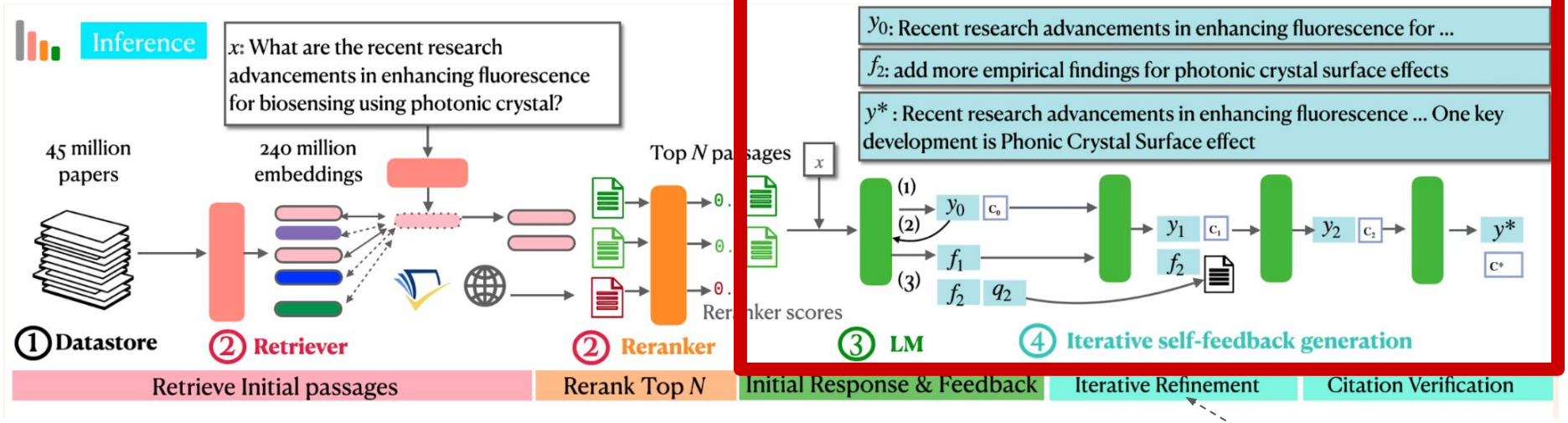


Specialized Retrievers and Rerankers: These tools are trained specifically to identify relevant passages from our scientific literature datastore

[1] Asai et al., Nature 2026. Synthesizing scientific literature with retrieval-augmented language models.

OpenScholar

- Inference

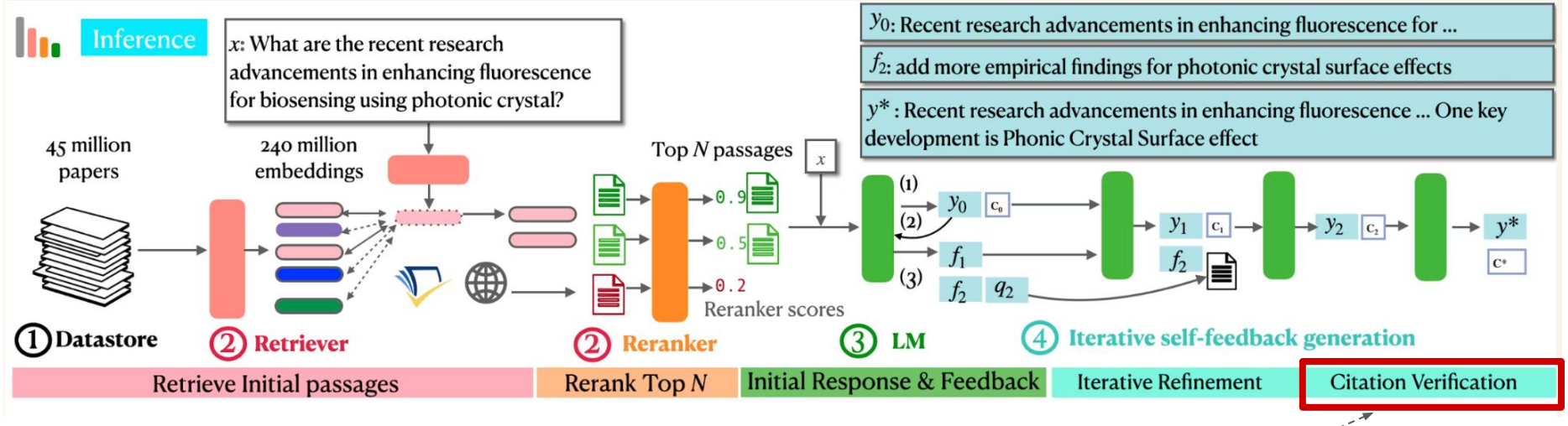


Iterative self-feedback to refine model outputs through natural language feedback. Each iteration involves additionally retrieving more papers.

[1] Asai et al., Nature 2026. Synthesizing scientific literature with retrieval-augmented language models.

OpenScholar

- Inference

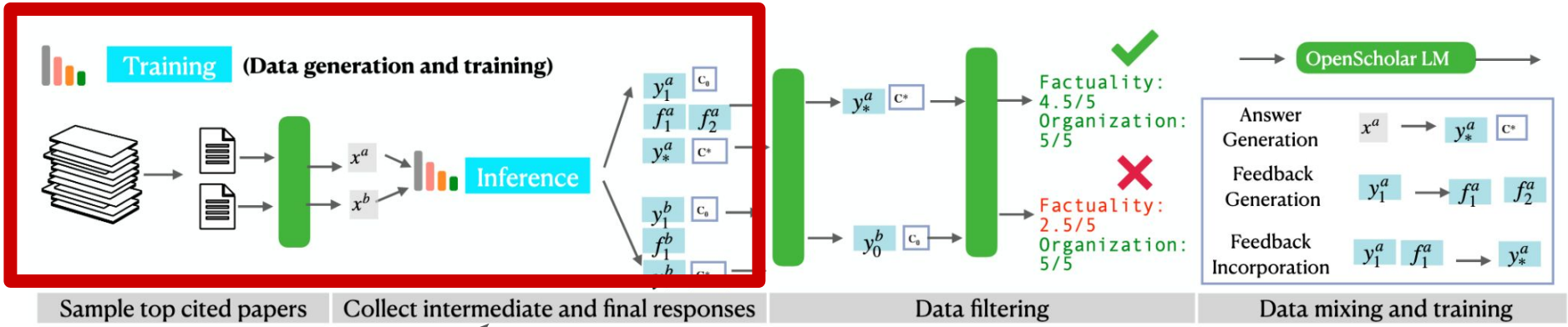


The generator LM ensures that all citation-worthy statements are adequately supported by references from the retrieved passages. If any claims lack proper citations, the LM performs a **post hoc insertion** to ensure that citation-worthy statements are supported by passages.

[1] Asai et al., Nature 2026. Synthesizing scientific literature with retrieval-augmented language models.

OpenScholar

- Training LLama 3.1 8B



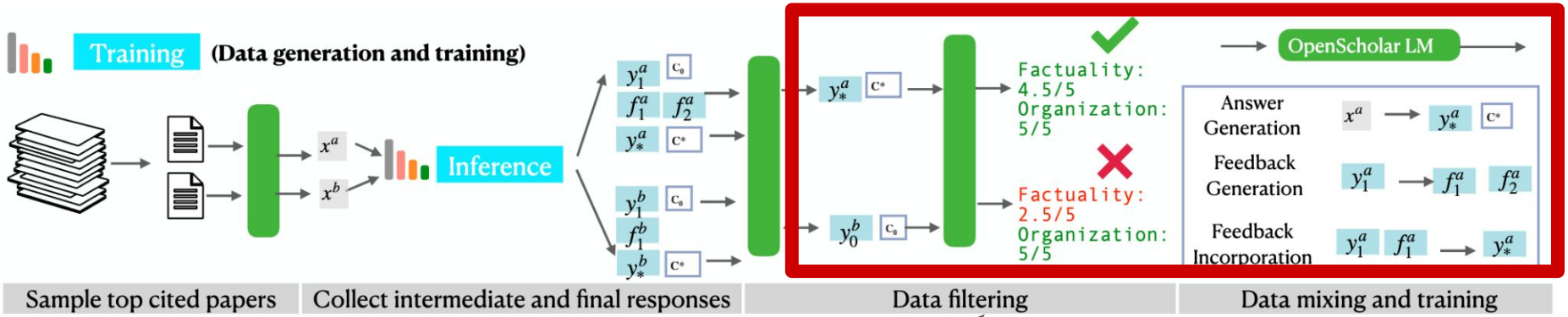
LLama 3.1 70B

generating information-seeking queries based on their abstracts that may require multiple papers to answer

[1] Asai et al., Nature 2026. Synthesizing scientific literature with retrieval-augmented language models.

OpenScholar

- Training LLama 3.1 8B



LLama 3.1 70B

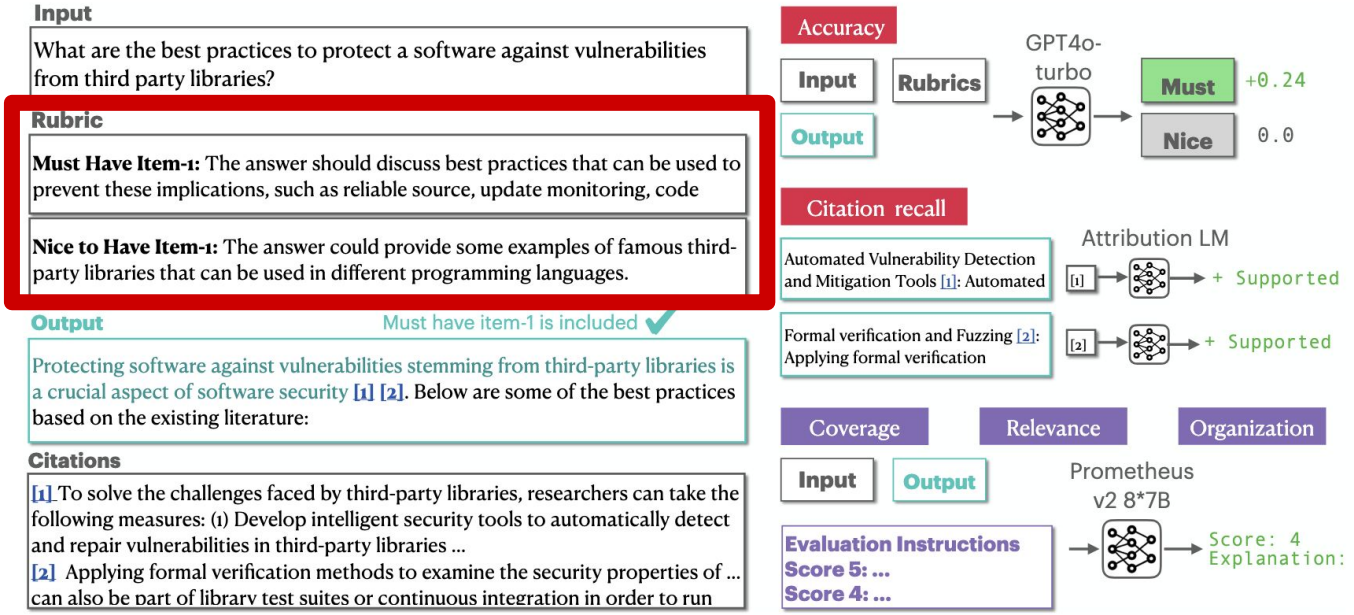
1) pair-wise filtering, compare the quality of model outputs at the the final step and the initial step, and retain the output that is judged to be higher quality at the final step.

2) Rate the selected response on a 5-point scale for two criteria: **organization**, and **factual precision/citation accuracy**. To be valid, the model's output must score at least 4.5 in both.

[1] Asai et al., Nature 2026. Synthesizing scientific literature with retrieval-augmented language models.

ScholarQABench

- ScholarQA-CS
  - 100 questions + detailed answer rubrics
  - Expert annotators (expert annotators holding Ph.D.s in the field)



[1] Asai et al., Nature 2026. Synthesizing scientific literature with retrieval-augmented language models.

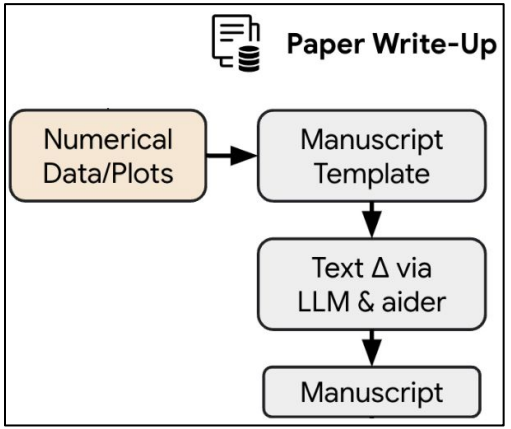
Agenda

- Short Text Generation with Citation Grounding
- Automatic Research Paper Writing
- Automated Survey & Deep Research Generation
- Meta-analysis Table Generation and Comparative Literature Synthesis
- Summary

The AI Scientist Paper Writing-up Process

Step 1: Per-section text generation

1. Prompt Aider to fill in a blank conference latex paper template section by section
2. All previous sections of the paper has already written are in the context of the language model
3. Each section is initially refined with one round of self-reflection
4. Aider is prompted to not include any citations in the text at this stage, and fill in only a skeleton for the related work



Paper Writing Aider Prompt

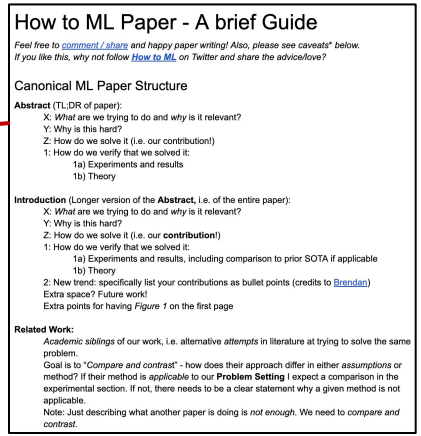
```
We've provided the `latex/template.tex` file to the project. We will be
filling it in section by section.

First, please fill in the {section} section of the writeup.

Some tips are provided below:
{per_section_tips}

Before every paragraph, please include a brief description of what you plan
to write in that paragraph in a comment.

Be sure to first name the file and use *SEARCH/REPLACE* blocks to perform
these edits.
```



How to ML Paper - A brief Guide

Feel free to comment / share and happy paper writing! Also, please see caveats below.*
If you like this, why not follow How to ML on Twitter and share the advice/love?

Canonical ML Paper Structure

Abstract (TL;DR of paper):
- X: *What* are we trying to do and *why* is it relevant?
- Y: Why is this hard?
- Z: How do we solve it (i.e. our contribution!)
- 1: How do we verify that we solved it:
  - 1a) Experiments and results
  - 1b) Theory

Introduction (Longer version of the **Abstract,** i.e. of the entire paper):
- X: *What* are we trying to do and *why* is it relevant?
- Y: Why is this hard?
- Z: How do we solve it (i.e. our **contribution**!)
- 1: How do we verify that we solved it:
  - 1a) Experiments and results, including comparison to prior SOTA if applicable
  - 1b) Theory
- 2: New trend: specifically list your contributions as bullet points (credits to Brendan)
- Extra space? Future work!
- Extra points for having *Figure 1* on the first page

Related Work:
Academic siblings of our work, i.e. alternative *attempts* in literature at trying to solve the same problem.
Goal is to "*Compare and contrast*" - how does their approach differ in either *assumptions* or method? If their method is *applicable* to our **Problem Setting** I expect a comparison in the experimental section. If not, there needs to be a clear statement why a given method is not applicable.
Note: Just describing what another paper is doing is *not enough*. We need to *compare and contrast*.

[Ref] Lu et al., arXiv 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery.

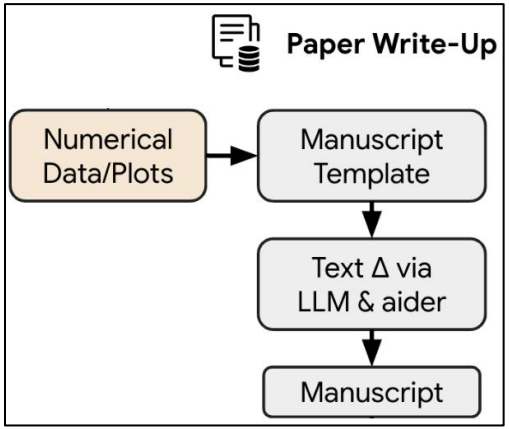
The AI Scientist Paper Writing-up Process

Step 2: Web search for references

1. 20 rounds to poll the Semantic Scholar API looking for the most relevant sources to compare and contrast the near-completed paper against for the **related work section**

2. Select papers for discussion and complete missing citations elsewhere in the paper

3. For each selected paper, **a brief note on where/how to cite it is sent to Aider**, and its BibTeX is automatically added to the LaTeX file



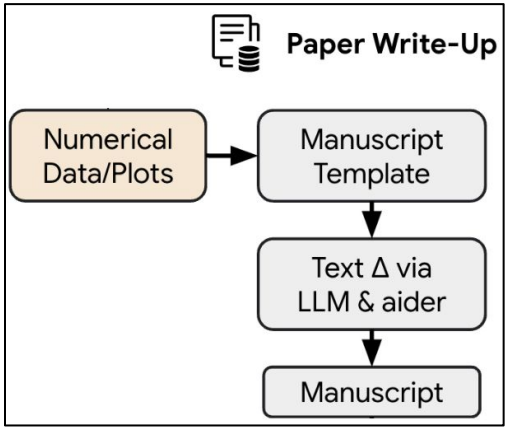
The AI Scientist Paper Writing-up Process

Step 2: Web search for references

1. 20 rounds to poll the Semantic Scholar API looking for the most relevant sources to compare and contrast the near-completed paper against for the **related work section**

2. Select papers for discussion and complete missing citations elsewhere in the paper

3. For each selected paper, **a brief note on where/how to cite it is sent to Aider**, and its BibTeX is automatically added to the LaTeX file



Step 3: Refinement

One final round of self-reflection section-by-section (aiming to remove any duplicated information and streamline the arguments)

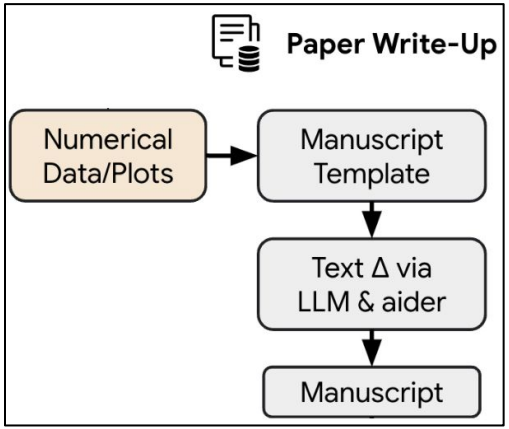
The AI Scientist Paper Writing-up Process

Step 2: Web search for references

1. 20 rounds to poll the Semantic Scholar API looking for the most relevant sources to compare and contrast the near-completed paper against for the **related work section**

2. Select papers for discussion and complete missing citations elsewhere in the paper

3. For each selected paper, **a brief note on where/how to cite it is sent to Aider**, and its BibTeX is automatically added to the LaTeX file



Step 3: Refinement

One final round of self-reflection section-by-section (aiming to remove any duplicated information and streamline the arguments)

Step 4: Compilation

After the Latex paper template is populated, it is compiled, and any errors are sent back to Aider for automatic correction

Agent Skills: ML-Paper-Writing

- An AI agent skill that transforms a research repo with code & experimental results into a publication-ready LaTeX paper, targeting top ML/AI and systems venues.

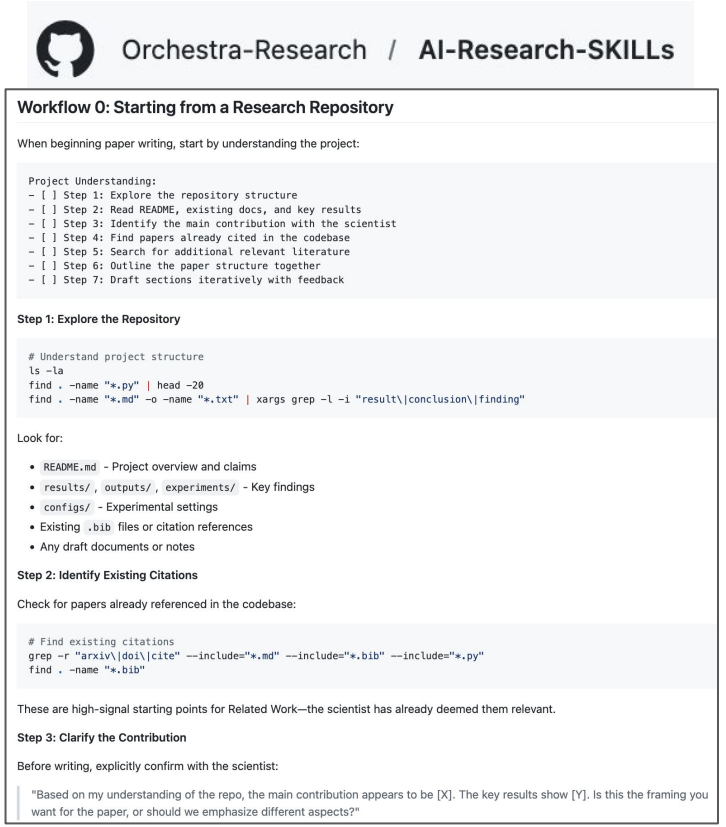
- Workflow

  ➔ **Understand**: Explore the repo, identify contribution, search literature

  ➔ **Draft**: Write proactively section by section, with the scientist giving feedback at each step

  ➔ **Cite**: Never hallucinate citations, always fetch BibTeX programmatically via Semantic Scholar/DOI; mark unverified as [placeholder]

  ➔ **Format & Submit**: Use conference LaTeX templates, check page limits, run checklists



Orchestra-Research / AI-Research-SKILLs

Workflow 0: Starting from a Research Repository

When beginning paper writing, start by understanding the project:

```
Project Understanding:
- [ ] Step 1: Explore the repository structure
- [ ] Step 2: Read README, existing docs, and key results
- [ ] Step 3: Identify the main contribution with the scientist
- [ ] Step 4: Find papers already cited in the codebase
- [ ] Step 5: Search for additional relevant literature
- [ ] Step 6: Outline the paper structure together
- [ ] Step 7: Draft sections iteratively with feedback
```

Step 1: Explore the Repository

```
# Understand project structure
ls -la
find . -name "*.py" | head -20
find . -name "*.md" -o -name "*.txt" | xargs grep -l -i "result\|conclusion\|finding"
```

Look for:

- `README.md` - Project overview and claims
- `results/`, `outputs/`, `experiments/` - Key findings
- `configs/` - Experimental settings
- Existing `.bib` files or citation references
- Any draft documents or notes

Step 2: Identify Existing Citations

Check for papers already referenced in the codebase:

```
# Find existing citations
grep -r "arxiv\|doi\|cite" --include="*.md" --include="*.bib" --include="*.py"
find . -name "*.bib"
```

These are high-signal starting points for Related Work—the scientist has already deemed them relevant.

Step 3: Clarify the Contribution

Before writing, explicitly confirm with the scientist:

> "Based on my understanding of the repo, the main contribution appears to be [X]. The key results show [Y]. Is this the framing you want for the paper, or should we emphasize different aspects?"

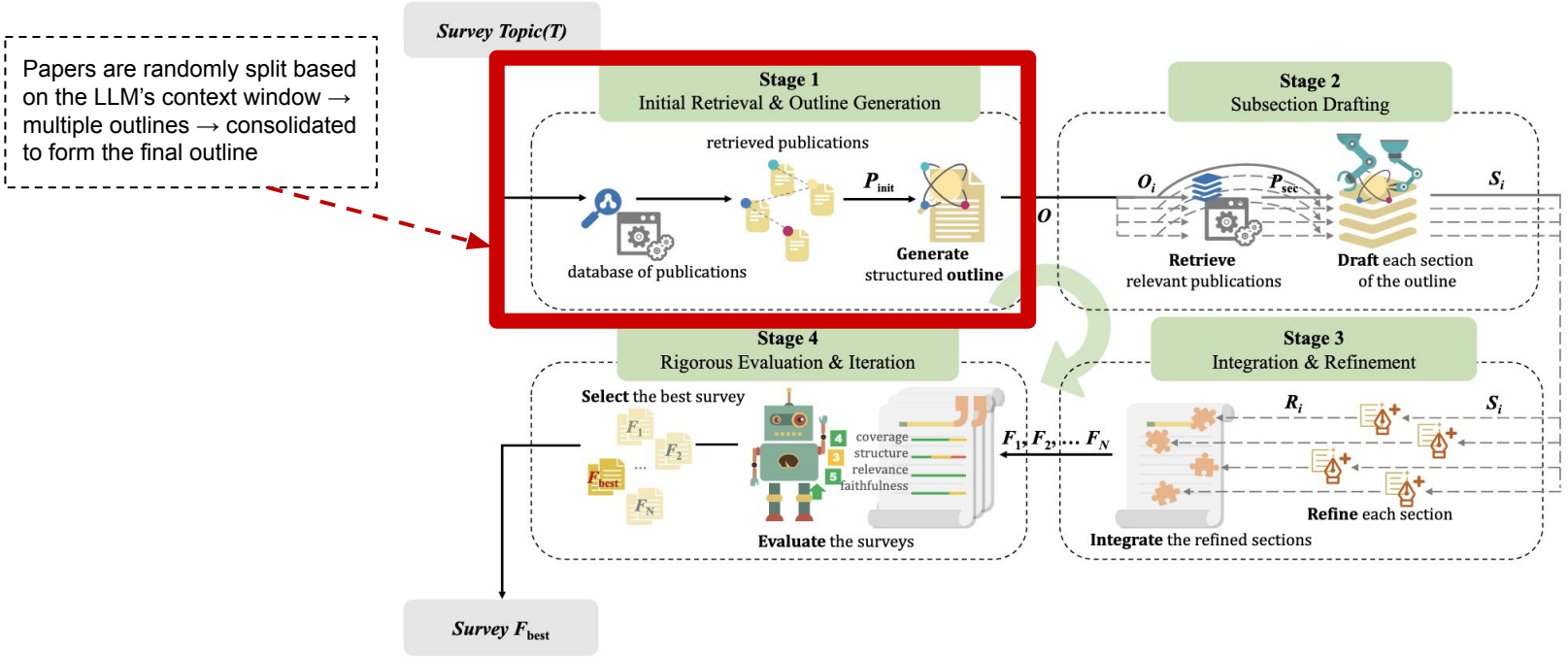
[Ref] https://github.com/Orchestra-Research/AI-Research-SKILLs/tree/main/20-ml-paper-writing.

Agenda

- ❑ Short Text Generation with Citation Grounding
- ❑ Automatic Research Paper Writing
- ❑ Automated Survey & Deep Research Generation
- ❑ Meta-analysis Table Generation and Comparative Literature Synthesis
- ❑ Summary

AutoSurvey

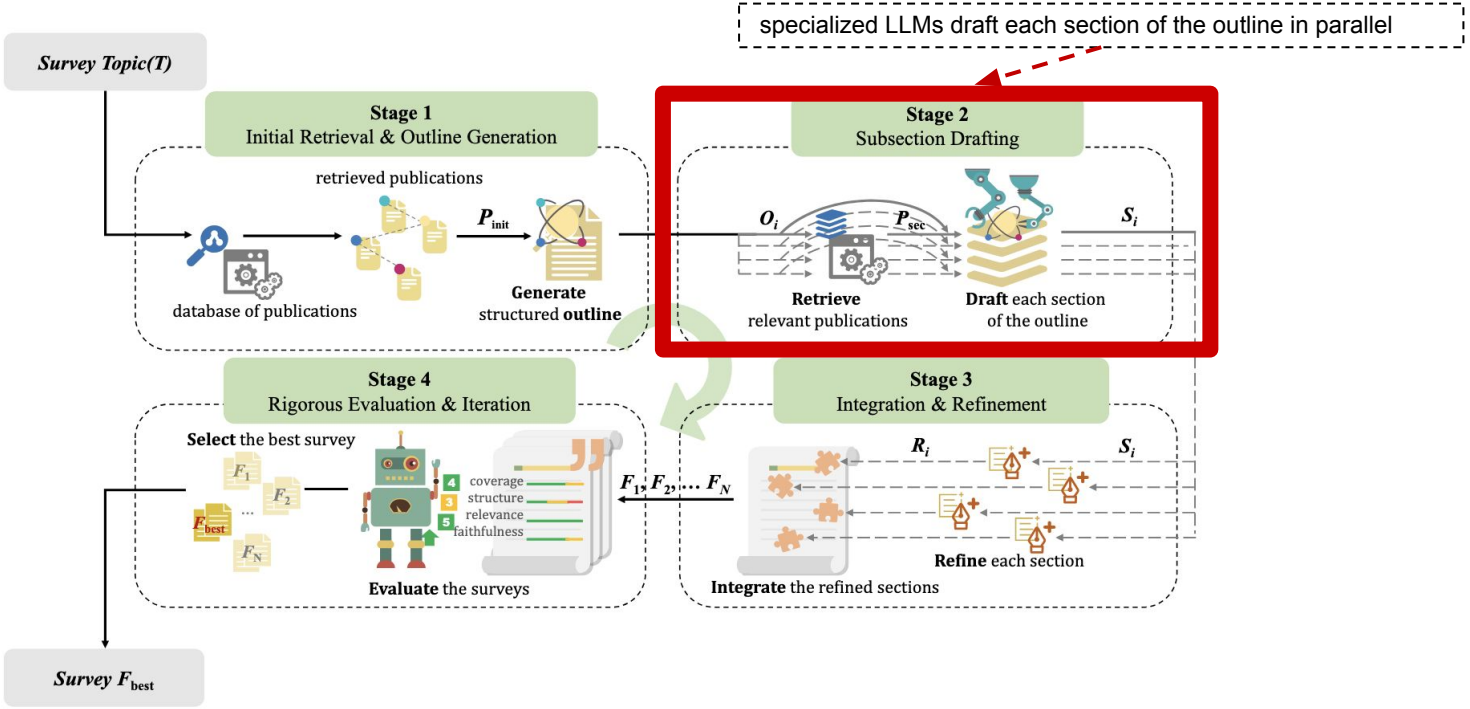
- AutoSurvey pipeline - a cost of $1.2 and 3 minutes per survey (with Claude-haiku)



[Ref] Wang et al., NeurIPS 2024. AutoSurvey: Large Language Models Can Automatically Write Surveys.

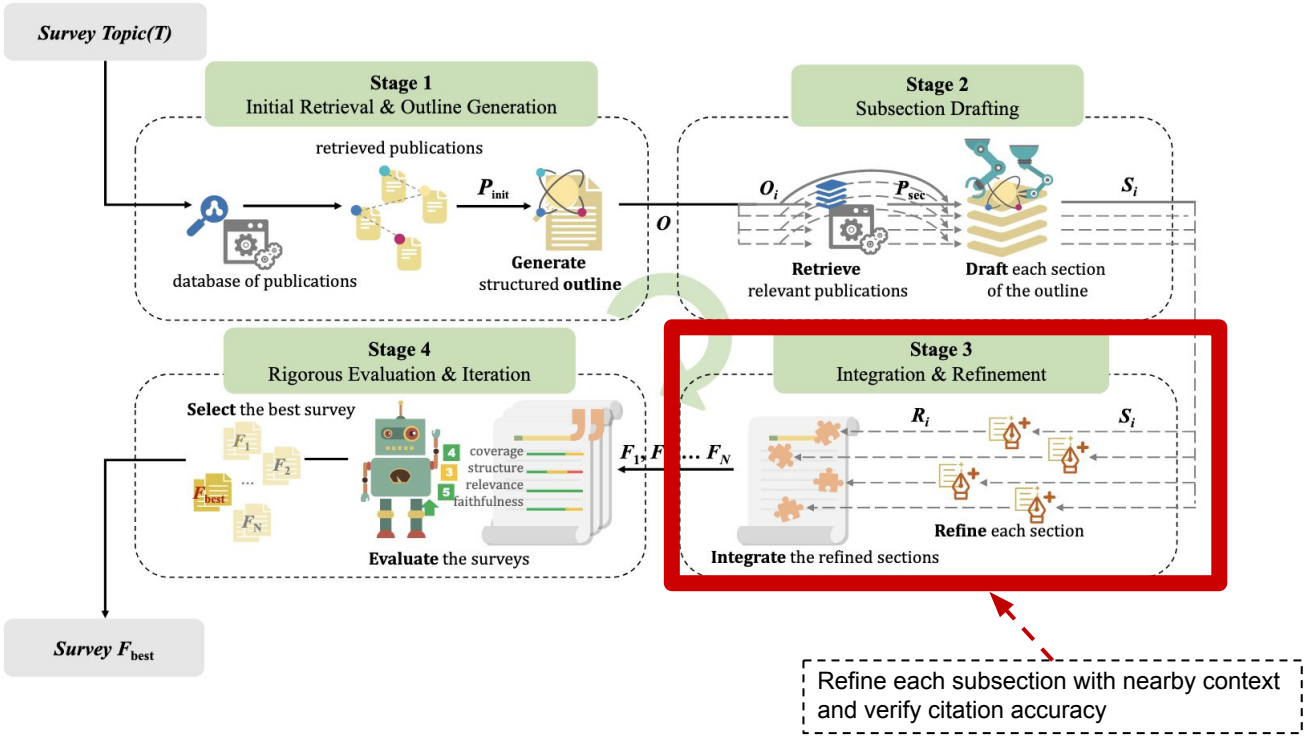
AutoSurvey

- AutoSurvey pipeline - a cost of $1.2 and 3 minutes per survey (with Claude-haiku)



AutoSurvey

- AutoSurvey pipeline - a cost of $1.2 and 3 minutes per survey (with Claude-haiku)



AutoSurvey

- Evaluation on 20 survey papers about LLMs (Claude-haiku as the writer)

Survey Title	Citations
A survey for in-context learning	323
A Survey on Large Language Models for Recommendation	55
A Survey of Detecting LLM-Generated Texts	42
Explainability for Large Language Models	25
A Survey on Evaluation of Large Language Models	183
A Survey on Large Language Model based Autonomous Agents	101
A Survey of Large Language Models in Medicine	234
Domain Specialization as the Key to Make Large Language Models Disruptive	14
Practical and Ethical Challenges of Large Language Models in Education	53
Aligning Large Language Models with Human	53
A Survey on ChatGPT and Beyond	144
Instruction Tuning for Large Language Models	45
Large Language Models for Information Retrieval	22
Towards Safer Generative Language Models: Safety Risks, Evaluations, and Improvements	17
A Survey of Chain of Thought Reasoning	13
A Survey on Hallucination in Large Language Models	116
Bias and Fairness in Large Language Models	12
Large-scale Multi-Modal Pre-trained Models	61
A Survey on Model Compression and Acceleration for Pretrained Language Models	22
Large Language Models for Software Engineering	49

AutoSurvey

- Evaluation on 20 survey papers about LLMs (Claude-haiku as the writer)

Survey Length (#tokens)	Methods	Speed	Citation quality		Content Quality			
			Recall	Precision	Coverage	Structure	Relevance	Avg.
8k	Human writing	0.16	80.00	87.50	4.50	4.16	5.00	4.52
	Naive RAG-based LLM generation	79.67	$78.14_{\pm 5.23}$	$71.92_{\pm 6.83}$	$4.40_{\pm 0.48}$	$3.86_{\pm 0.71}$	$4.86_{\pm 0.33}$	4.33
	AutoSurvey	107.00	$82.48_{\pm 2.77}$	$77.42_{\pm 3.28}$	$4.60_{\pm 0.48}$	$4.46_{\pm 0.49}$	$4.8_{\pm 0.39}$	4.61
16k	Human writing	0.14	88.52	79.63	4.66	4.38	5.00	4.66
	Naive RAG-based LLM generation	43.41	$71.48_{\pm 12.50}$	$65.31_{\pm 15.36}$	$4.46_{\pm 0.49}$	$3.66_{\pm 0.69}$	$4.73_{\pm 0.44}$	4.23
	AutoSurvey	95.51	$81.34_{\pm 3.65}$	$76.94_{\pm 1.93}$	$4.66_{\pm 0.47}$	$4.33_{\pm 0.59}$	$4.86_{\pm 0.33}$	4.60
32k	Human writing	0.10	88.57	77.14	4.66	4.50	5.00	4.71
	Naive RAG-based LLM generation	22.64	$79.88_{\pm 4.35}$	$65.03_{\pm 8.39}$	$4.41_{\pm 0.64}$	$3.75_{\pm 0.72}$	$4.66_{\pm 0.47}$	4.23
	AutoSurvey	91.46	$83.14_{\pm 2.44}$	$78.04_{\pm 3.14}$	$4.73_{\pm 0.44}$	$4.26_{\pm 0.69}$	$4.8_{\pm 0.54}$	4.58
64k	Human writing	0.07	86.33	77.78	5.00	4.66	5.00	4.88
	Naive RAG-based LLM generation	12.56	$68.79_{\pm 11.00}$	$61.97_{\pm 13.45}$	$4.4_{\pm 0.61}$	$3.66_{\pm 0.47}$	$4.66_{\pm 0.47}$	4.19
	AutoSurvey	73.59	$82.25_{\pm 3.64}$	$77.41_{\pm 3.84}$	$4.73_{\pm 0.44}$	$4.33_{\pm 0.47}$	$4.86_{\pm 0.33}$	4.62

LLM as a judge

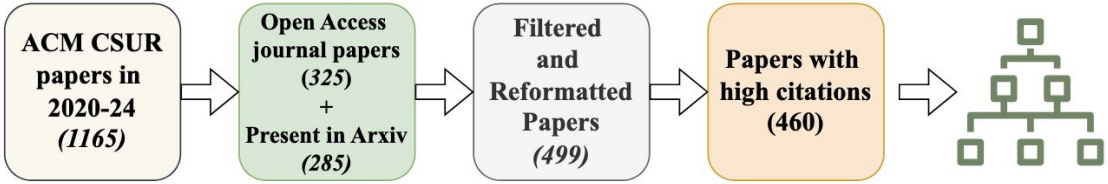
73.59 survey papers per hour

NLI model to assess the claim grounding coverage and relevance of citations

"It achieves near-human levels of coverage, relevance, and citation quality while maintaining a significantly lower time cost."

TaxoAlign

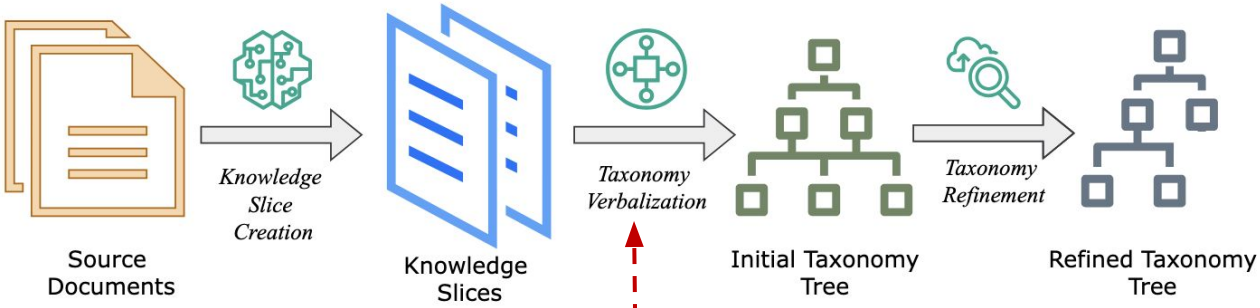
- Goal: Automated taxonomy creation that can bridge the gap between human-generated and automatically-created taxonomies



[Ref] Lahiri et al., EMNLP 2025. TaxoAlign: Scholarly Taxonomy Generation Using Language Models.

TaxoAlign

- Goal: Automated taxonomy creation that can bridge the gap between human-generated and automatically-created taxonomies



[Ref] Lahiri et al., EMNLP 2025. TaxoAlign: Scholarly Taxonomy Generation Using Language Models.

Agenda

- Short Text Generation with Citation Grounding
- Automatic Research Paper Writing
- Automated Survey & Deep Research Generation
- Meta-analysis Table Generation and Comparative Literature Synthesis
- Summary

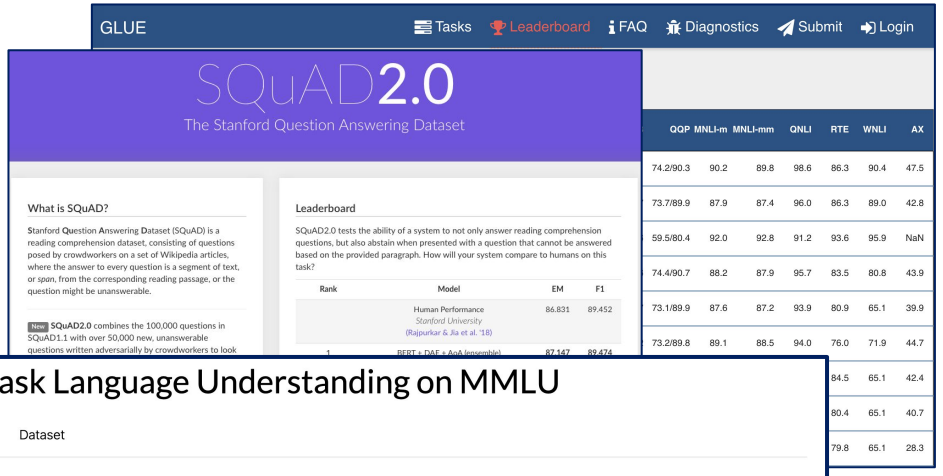
Meta-analysis for NLP/ML Literature: Scientific Leaderboard Construction

EACL 2026 RABAT·MOROCCO Mars · March 24-29, 2026 · مارس

- A **scientific leaderboard** is a ranked list of methods/models evaluated on the same task using a shared benchmark and metric.

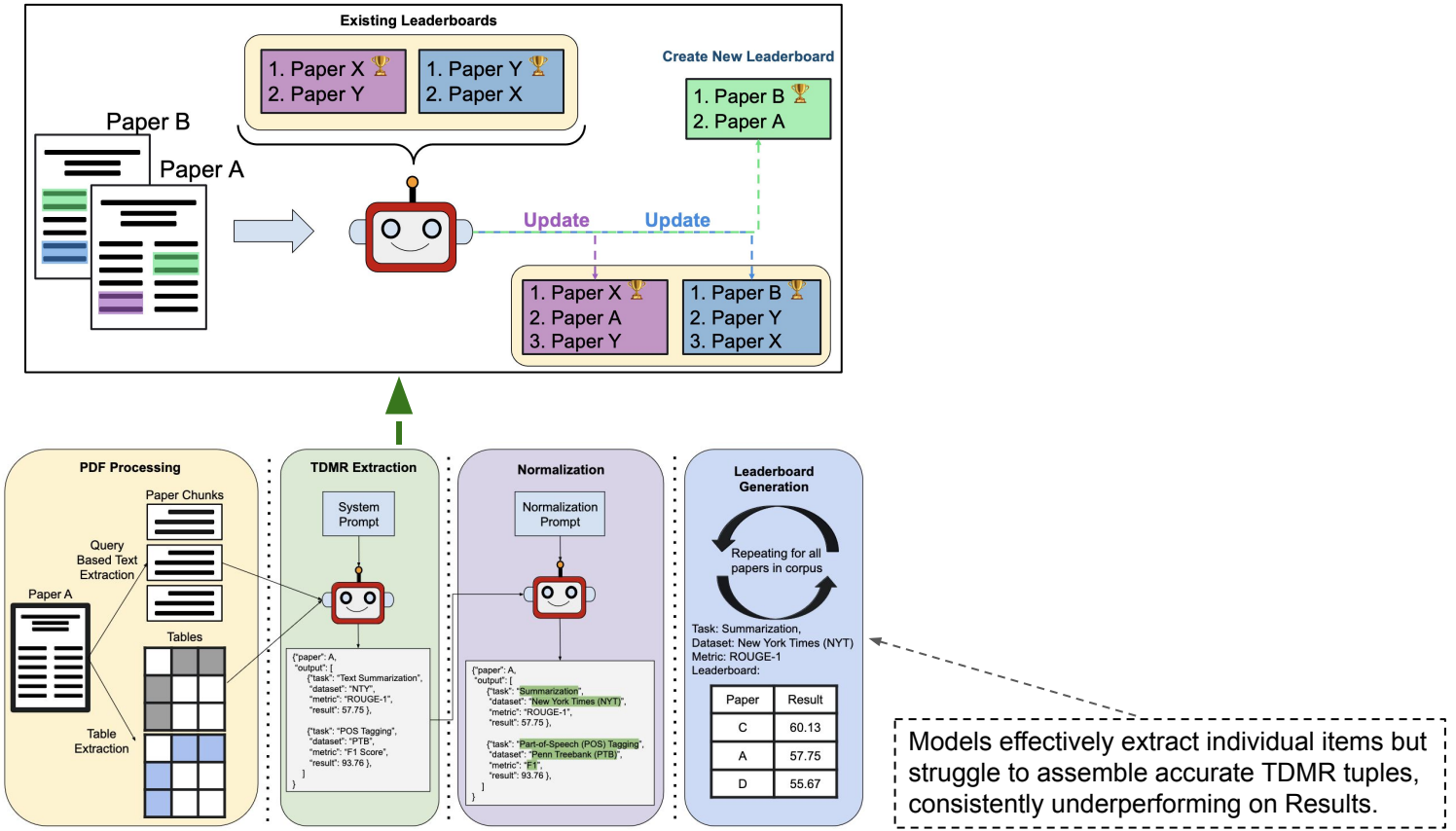
?

Automate the Whole Process



[Ref] Hou et al., ACL 2019. Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction.

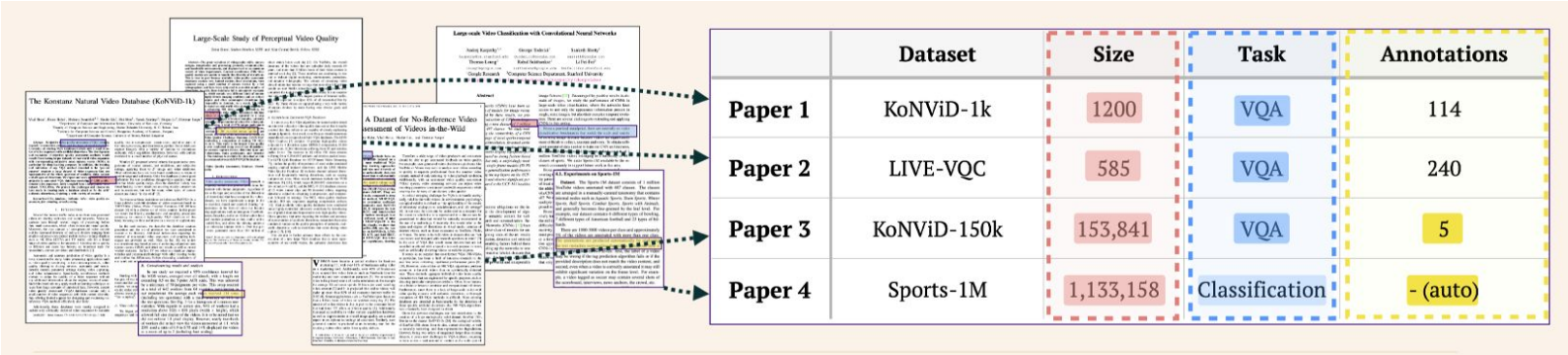
Meta-analysis for NLP/ML Literature: Scientific Leaderboard Construction



[Ref] Şahinuç et al., EMNLP 2024. Efficient Performance Tracking: Leveraging Large Language Models for Automated Construction of Scientific Leaderboards.

ArxivDIGESTables: Automatically Generate Literature Review Tables

- **Task: literature review table generation**
  - Rows are a set of papers
  - Columns are a set of aspects that the papers share

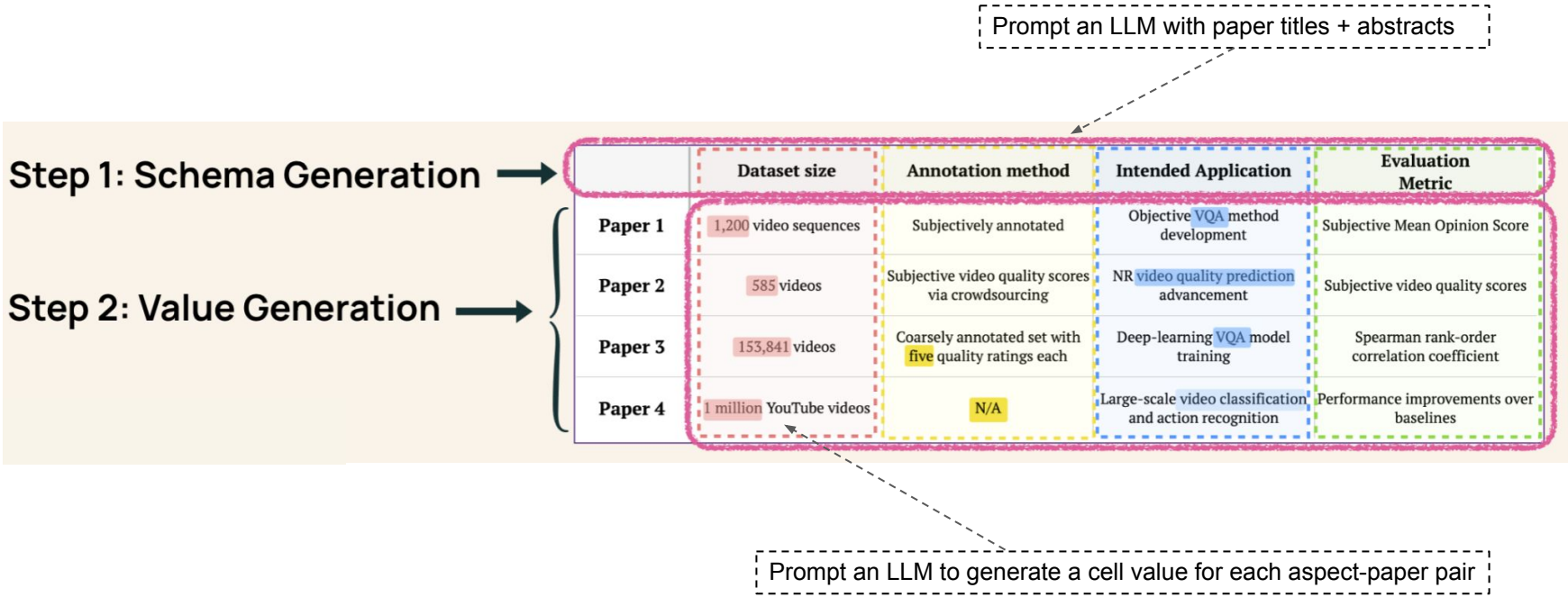


	Dataset	Size	Task	Annotations
Paper 1	KoNViD-1k	1200	VQA	114
Paper 2	LIVE-VQC	585	VQA	240
Paper 3	KoNViD-150k	153,841	VQA	5
Paper 4	Sports-1M	1,133,158	Classification	- (auto)

[Ref] Newman et al., EMNLP 2024. ArxivDIGESTables: Synthesizing Scientific Literature into Tables using Language Models.

ArxivDIGESTables: Automatically Generate Literature Review Tables

- **Two stage decomposition generation**



[Ref] Newman et al., EMNLP 2024. ArxivDIGESTables: Synthesizing Scientific Literature into Tables using Language Models.

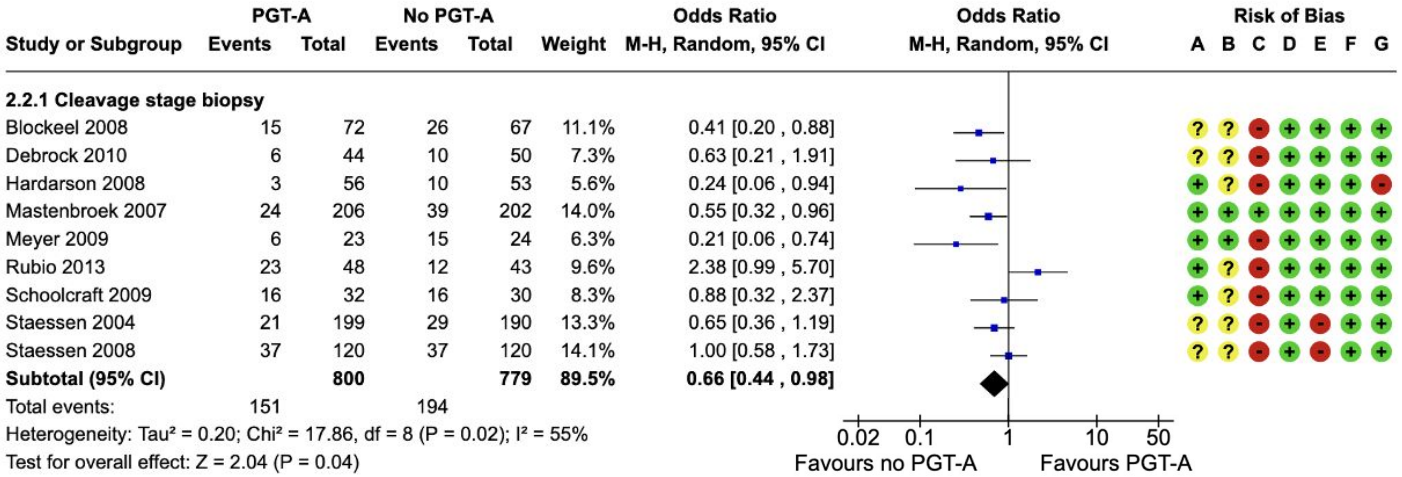
Meta-analysis for Biomedical Literature: Forest-plot Generation

EACL 2026
RABAT · MOROCCO
Mars · March 24-29, 2026 · مارس

- A forest plot is a chart that shows the results of several studies in one place.
- It helps you compare each study and see the overall result when all studies are combined.

Meta-analysis for Biomedical Literature: Forest-plot Generation

Research question: Whether the intervention PGT-A (formerly known as preimplantation genetic screening) leads to fewer live births than in the control group for women undergoing an IVF (in vitro fertilization) treatment?

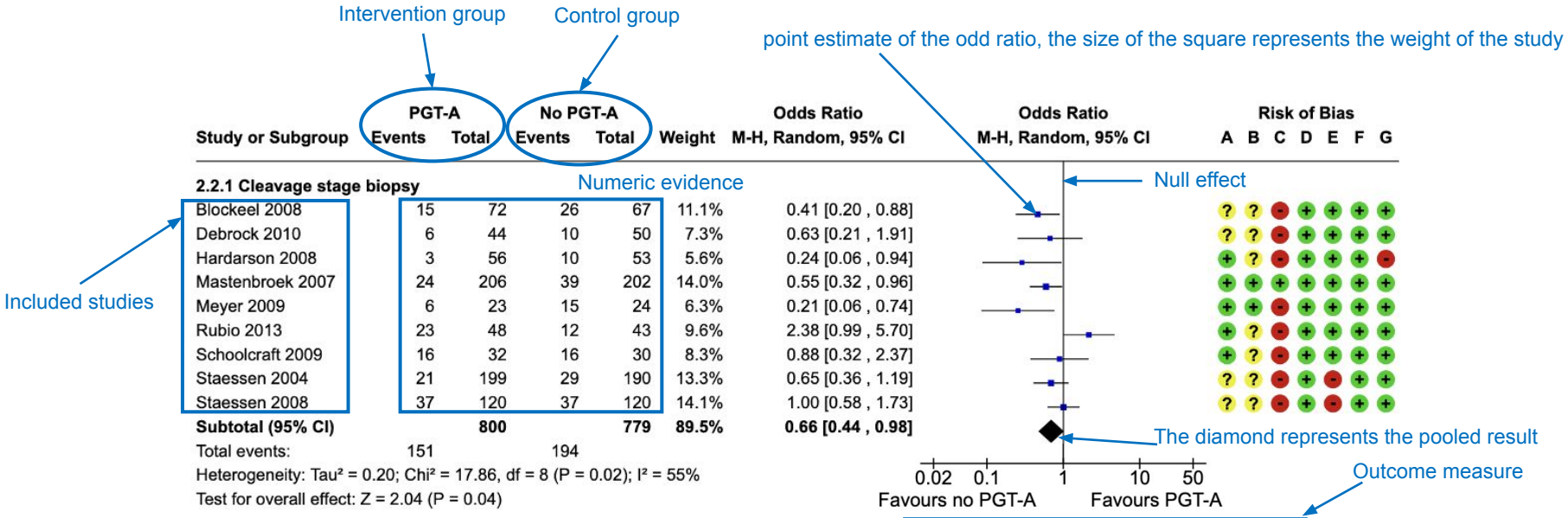


IVF with PGT-A versus IVF without PGT-A with the use of FISH for the genetic analysis, outcome: live birth rate after the first embryo transfer per woman randomised.

Conclusions: The currently available evidence is insufficient to support PGT-A in routine clinical practice.

Meta-analysis for Biomedical Literature: Forest-plot Generation

Research question: Whether the intervention PGT-A (formerly known as preimplantation genetic screening) leads to fewer live births than in the control group for women undergoing an IVF (in vitro fertilization) treatment?



Study or Subgroup	PGT-A Events	PGT-A Total	No PGT-A Events	No PGT-A Total	Weight	Odds Ratio M-H, Random, 95% CI
2.2.1 Cleavage stage biopsy						
Blockeel 2008	15	72	26	67	11.1%	0.41 [0.20 , 0.88]
Debrock 2010	6	44	10	50	7.3%	0.63 [0.21 , 1.91]
Hardarson 2008	3	56	10	53	5.6%	0.24 [0.06 , 0.94]
Mastenbroek 2007	24	206	39	202	14.0%	0.55 [0.32 , 0.96]
Meyer 2009	6	23	15	24	6.3%	0.21 [0.06 , 0.74]
Rubio 2013	23	48	12	43	9.6%	2.38 [0.99 , 5.70]
Schoolcraft 2009	16	32	16	30	8.3%	0.88 [0.32 , 2.37]
Staessen 2004	21	199	29	190	13.3%	0.65 [0.36 , 1.19]
Staessen 2008	37	120	37	120	14.1%	1.00 [0.58 , 1.73]
Subtotal (95% CI)		**800**		**779**	**89.5%**	**0.66 [0.44 , 0.98]**
Total events:	151		194			

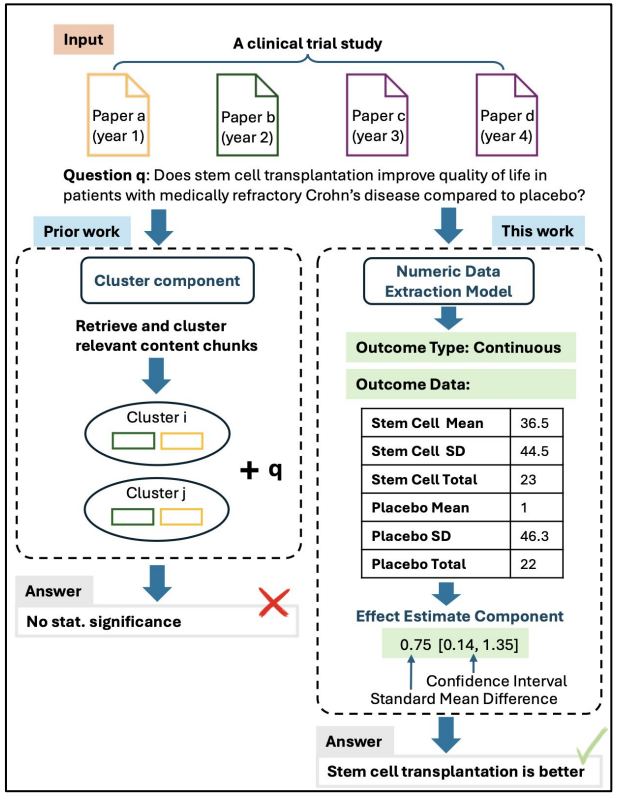
Heterogeneity: Tau² = 0.20; Chi² = 17.86, df = 8 (P = 0.02); I² = 55%
Test for overall effect: Z = 2.04 (P = 0.04)

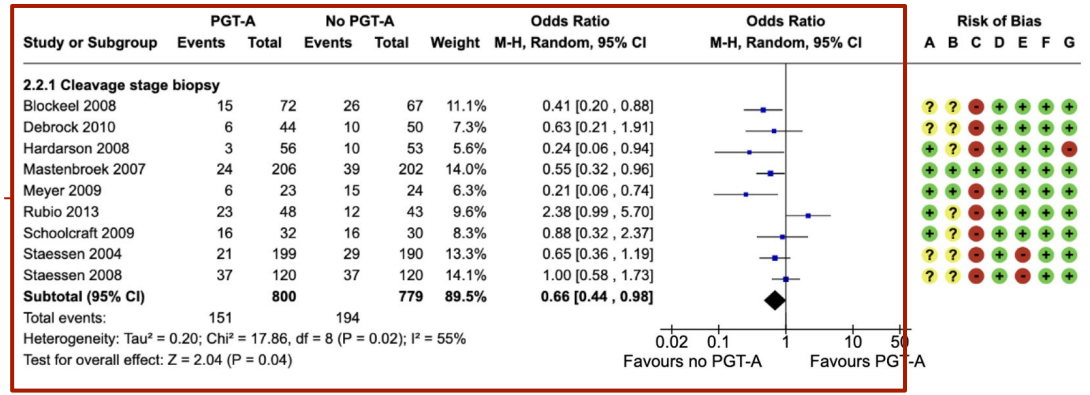
IVF with PGT-A versus IVF without PGT-A with the use of FISH for the genetic analysis, outcome: live birth rate after the first embryo transfer per woman randomised.

Conclusions: The currently available evidence is insufficient to support PGT-A in routine clinical practice.

Meta-analysis for Biomedical Literature: Forest-plot Generation

- A numeric reasoning approach to estimate study effect

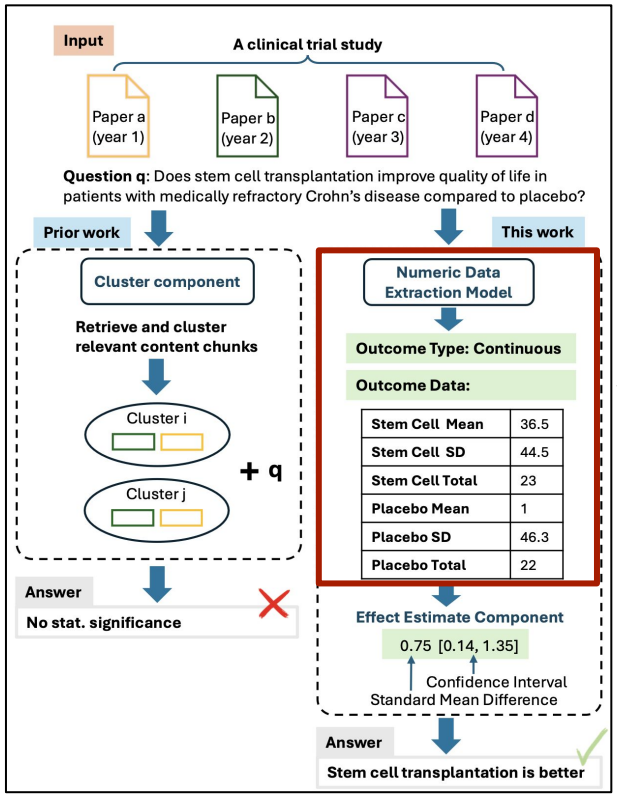


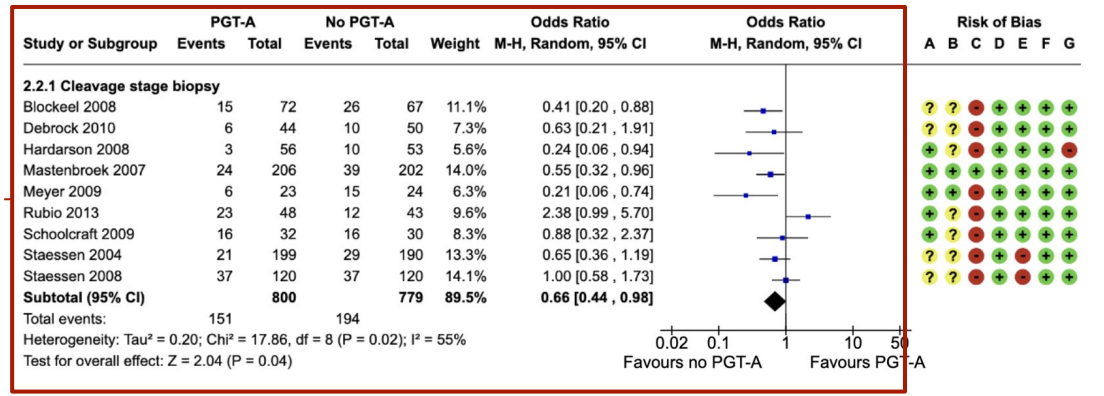


[Ref] Pronesti et al., EMNLP 2025. Enhancing Study-Level Inference from Clinical Trial Papers via RL-based Numeric Reasoning.

Meta-analysis for Biomedical Literature: Forest-plot Generation

- A numeric reasoning approach to estimate study effect





Fine-tune a numeric evidence extraction model using RL with GRPO

- **Correctness Reward (CR)**:

$$R_{\text{CR}} = \frac{1 + \sum_{j=1}^{n} \mathbb{1}\{v_j \approx \hat{v}_j\}}{1+n}$$

- **Format Reward (FR)**:

$$R_{\text{FR}} = \begin{cases} 1 & \text{if } \pi_\theta(x) \in \mathcal{F} \\ 0 & \text{otherwise} \end{cases}, \quad \text{with } \mathcal{F} \text{ set of valid formats}$$

- **Thought Format Reward (TFR)**:

$$R_{\text{TFR}} = \begin{cases} 1 & \text{if } \pi_\theta(x) \text{ matches thought pattern} \\ 0 & \text{otherwise} \end{cases}$$

- **Final reward**:

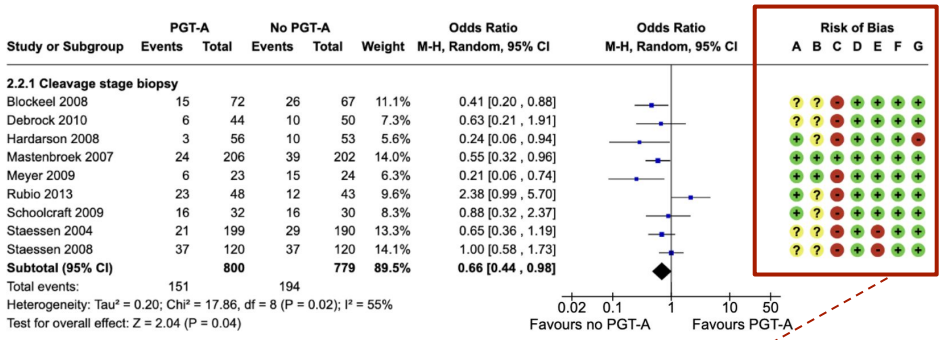
$$R = 0.8 \cdot R_{\text{CR}} + 0.1 \cdot R_{\text{FR}} + 0.1 \cdot R_{\text{TFR}}$$

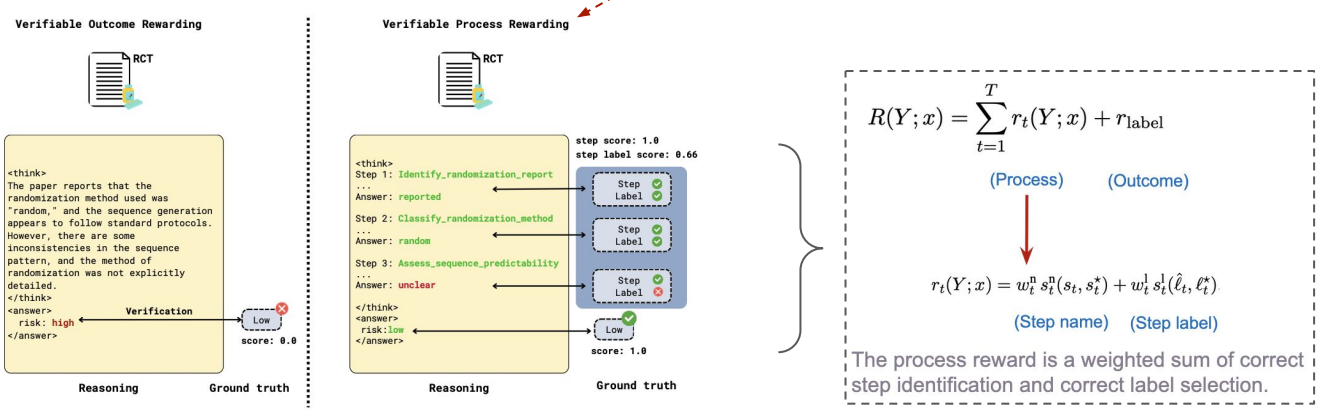
Existing reasoning models are insufficient for complex domain-specific tasks.

[Ref] Pronesti et al., EMNLP 2025. Enhancing Study-Level Inference from Clinical Trial Papers via RL-based Numeric Reasoning.

Meta-analysis for Biomedical Literature: Forest-plot Generation

- Verifiable process reward models for risk of bias analysis

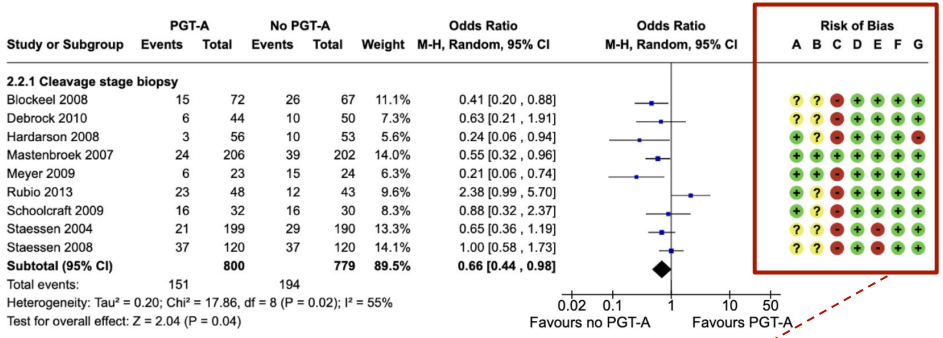




[Ref] Pronesti et al., arXiv 2026. Beyond Outcome Verification: Verifiable Process Reward Models for Structured Reasoning.

Meta-analysis for Biomedical Literature: Forest-plot Generation

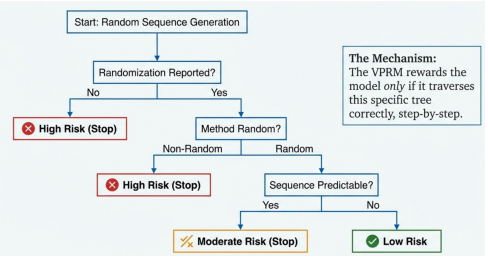
- Verifiable process reward models for risk of bias analysis

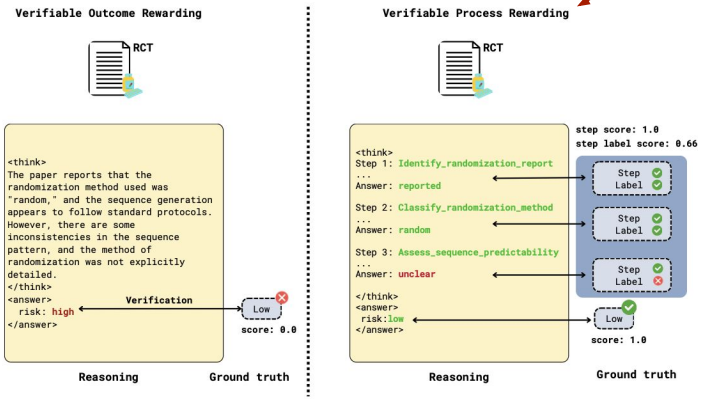


Error caught immediately during the process. Model penalized.

Algorithm 1 RoB A Macro

```
procedure PredictLabel-A(steps)
    if steps[IdentifyRandomizationReport] = NotReported then
        return Moderate
    end if
    if steps[ClassifyRandomizationMethod] = NonRandom then
        return High
    end if
    if steps[AssessSequencePredictability] = Predictable then
        return Moderate
    end if
    if steps[BaselineImbalance] = Likely then
        return High
    end if
    return Low
end procedure
```





$$R(Y;x) = \sum_{t=1}^{T} r_t(Y;x) + r_{\text{label}}$$

(Process) (Outcome)

$$r_t(Y;x) = w_t^{\text{n}}\, s_t^{\text{n}}(s_t, s_t^\star) + w_t^{\text{l}}\, s_t^{\text{l}}(\hat{\ell}_t, \ell_t^\star)$$

(Step name) (Step label)

The process reward is a weighted sum of correct step identification and correct label selection.

[Ref] Pronesti et al., arXiv 2026. Beyond Outcome Verification: Verifiable Process Reward Models for Structured Reasoning.

Agenda

- Short Text Generation with Citation Grounding
- Automatic Research Paper Writing
- Automated Survey & Deep Research Generation
- Meta-analysis Table Generation and Comparative Literature Synthesis
- Summary

Summary

	Pre-LLM Era	LLM Era
Key tasks	Related work generation Citation text generation Cite-worthiness detection Citation recommendation Citation intent classification Citation analysis	LLM citation generation and attribution Agentic deep research • Planning • Information gathering • Tool usage • Synthesis and reporting
Core challenges	Coherent multi-doc synthesis	Verifiable, non-hallucinated attribution
Generation quality	Often disfluent or extractive	Fluent but may hallucinate or post-rationalize
Evaluation	Rouge, F1, BERTScore	LLM-as-a-judge, citation quality based on NLI models (citation precision/citation recall)

Summary

- **The verification bottleneck:** Generation has outpaced verification!
  - The field can now produce fluent, cited text at scale
  - Automated attribution verification (NLI-based, LLM-as-judge) is still unreliable that human checking remains necessary.

Summary

- **The verification bottleneck:** Generation has outpaced verification!

- **AI-assisted scientific writing**
  - Fluent AI-generated text with seemingly grounded citations can create a false sense of rigor.
  - Without reliable automated verification, AI paper-writing tools may accelerate scientific output while **quietly eroding the trust and reliability on which researchers depend**.

References

1. Wright & Augenstein, ACL 2021 Findings. CiteWorth: Cite-Worthiness Detection for Improved Scientific Document Understanding.
2. Çelik & Tekir, EMNLP 2025. CiteBART: Learning to Generate Citations for Local Citation Recommendation.
3. Jurgen et al., TACL 2018. Measuring the Evolution of a Scientific Field through Citation Frames.
4. Li et al., NAACL 2022. CORWA: A Citation-Oriented Related Work Annotation Dataset.
5. Pramanick et al., arXiv 2026. ClaimFlow: Tracing the Evolution of Scientific Claims in NLP.
6. Chen et al., ACL 2021. Capturing Relations between Scientific Papers: An Abstractive Model for Related Work Section Generation.
7. Lu et al., EMNLP 2020. Multi-XScience: A Large-scale Dataset for Extreme Multi-document Summarization of Scientific Articles.
8. Xing et al., ACL 2020. Automatic Generation of Citation Texts in Scholarly Papers: A Pilot Study.
9. Funkquist et al., EMNLP 2023. CiteBench: A Benchmark for Scientific Citation Text Generation.
10. Şahinuç et al., ACL 2024. Systematic Task Exploration with LLMs: A Study in Citation Text Generation.
11. Li et al., arXiv 2023. A Survey of Large Language Models Attribution.
12. Schreieder et al., arXiv 2025. Attribution, Citation, and Quotation: A Survey of Evidence-based Text Generation with Large Language Models.
13. Asai et al., Nature 2026. Synthesizing scientific literature with retrieval-augmented language models.
14. Lu et al., arXiv 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery.
15. Wang et al., NeurIPS 2024. AutoSurvey: Large Language Models Can Automatically Write Surveys.
16. Lahiri et al., EMNLP 2025. TaxoAlign: Scholarly Taxonomy Generation Using Language Models.
17. Hou et al., ACL 2019. Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction.
18. Şahinuç et al., EMNLP 2024. Efficient Performance Tracking: Leveraging Large Language Models for Automated Construction of Scientific Leaderboards.
19. Newman et al., EMNLP 2024. ArxivDIGESTables: Synthesizing Scientific Literature into Tables using Language Models.
20. Pronesti et al., EMNLP 2025. Enhancing Study-Level Inference from Clinical Trial Papers via RL-based Numeric Reasoning.
21. Pronesti et al., arXiv 2026. Beyond Outcome Verification: Verifiable Process Reward Models for Structured Reasoning.