STORM and Co-STORM

A deep research dive into the Stanford pre-writing system and its 2025 follow-up.

Internal briefing · June 2026 · 28 slides

Source: paper-brief § 1, 2. arXiv 2402.14207, 2408.15232. stanford-oval/storm. storm.genie.stanford.edu.

Pre-writing is the bottleneck.

LLMs can write. They cannot research. STORM is the system that closes the gap.

Source: paper-brief § 1. arXiv 2402.14207 abstract.

The shape of the deck

1. The problem and the paper
2. The method in detail
3. The dataset (FreshWiki)
4. The automatic evaluation
5. The human evaluation
6. What the editors flagged as wrong
7. Co-STORM, the 2025 follow-up
8. The system today
9. Adopters, related work, limitations
10. Mapping to Parallax
11. References

Source: paper-brief § 1-14.

1. The problem and the paper

Source: paper-brief § 3. arXiv 2402.14207 § 1.

The problem

Long-form grounded writing is hard.

Prior work has either:

- Assumed reference documents are supplied in advance (WikiSum, Liu et al., 2018)
- Assumed an outline is given (Fan and Gardent, 2022)
- Focused on paragraph-level rather than article-level (Balepur et al., 2023)

STORM is the first system to study full article-from-scratch with **no outline or references supplied**.

Source: paper-brief § 3. arXiv 2402.14207 Table 1.

Why pre-writing matters

D. Gordon Rohman, 1965, *College Composition and Communication*:

> "Pre-writing: The Stage of Discovery in the Writing Process."

Even experienced writers spend more time on research and outline than on drafting. The pre-writing stage is:

- Discover what to research (perspective discovery)
- Collect and curate references (`R`)
- Build a multi-level outline (`O`)
- Then write, section by section

Source: paper-brief § 3. arXiv 2402.14207 § 1. Rohman, 1965.

The paper

Shao, Y., Jiang, Y., Kanell, T. A., Xu, P., Khattab, O., Lam, M. S. (2024). "Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models." *NAACL 2024 Main Conference*.

- 27 pages, main conference.
- arXiv:2402.14207. Submitted 22 Feb 2024, v2 8 Apr 2024.
- MIT-licensed open source at github.com/stanford-oval/storm.
- 28,475 stars, 2,601 forks as of June 18 2026.

Source: paper-brief § 1, 2, 8. arXiv 2402.14207. GitHub API.

2. The method in detail

Source: paper-brief § 4. arXiv 2402.14207 § 3, Figure 2.

The STORM method, in one diagram

Stage A: Perspective discovery

Given topic `t`:

1. LLM generates a list of related topics
2. Fetch tables of contents of their Wikipedia articles (via the `Wikipedia-API` PyPI package)
3. Concatenate the tables of contents into a context
4. Ask the LLM to identify `N` perspectives that "can collectively contribute to a comprehensive article on `t`"
5. Always add baseline perspective `p0`: "basic fact writer focusing on broadly covering the basic facts"

The paper sets `N = 5` in the main experiments.

Source: paper-brief § 4.1. arXiv 2402.14207 § 3.1.

Stage B: Simulated conversations

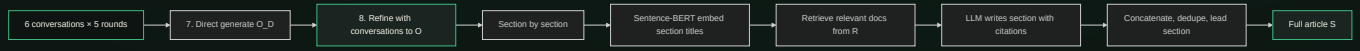
For each perspective `p` in `P = {p0, p1, ..., pN}`:

```
for round in range(M):
    q_i = llm_writer(topic, perspective, history)
    queries = split_queries(q_i)
    sources = search_engine.search(queries)
    sources = filter_reliable(sources)  # Wikipedia's Reliable sources guideline
    a_i = llm_expert(synthesize, sources)
    references.update(sources)
    history.append(q_i, a_i)
```

`M = 5` rounds. Full loop runs `N + 1 = 6` parallel conversations, up to 30 question-answer pairs, up to 30+ retrieved sources per topic.

Source: paper-brief § 4.2. arXiv 2402.14207 § 3.2, Algorithm 1.

Stage C: Outline curation and writing



The outline is generated twice: first as a draft from the topic alone, then refined using the topic, the draft, and the conversation transcripts.

Source: paper-brief § 4.3. arXiv 2402.14207 § 3.3, § 3.4.

Implementation

- **Framework:** zero-shot prompting in DSPy (Khattab et al., 2023, arXiv:2310.03714)
- **Search backend:** You.com (the pipeline is search-engine-agnostic)
- **Embeddings:** Sentence-BERT `paraphrase-MiniLM-L6-v2`
- **NER:** FLAIR (Akbik et al., 2019) for entity recall
- **Main LLM:** GPT-3.5-turbo for question-asking, GPT-3.5-turbo-instruct for the rest
- **Article generation LLM:** GPT-4 (because GPT-3.5 is not faithful to sources when generating text with citations)
- **Temperature:** 1.0, top_p 0.9

No fine-tuning. Pure prompting.

Source: paper-brief § 4.3, 6.1. arXiv 2402.14207 § 4.4.

3. The dataset: FreshWiki

Source: paper-brief § 5. arXiv 2402.14207 § 2.1.

FreshWiki

The dataset that avoids data leakage.

- Top-100 most-edited Wikipedia pages per month, Feb 2022 to Sep 2023
- B-class quality or above (per ORES). About 3% of Wikipedia pages meet this.
- Excludes list articles and articles with no subsections
- Plain text only (no tables, no images)

Controlled experiment: 100 articles under 3,000 words.

Source: paper-brief § 5. arXiv 2402.14207 § 2.1, Appendix A.

FreshWiki statistics

STATISTIC	MEAN
Sections per article	8.4
All-level headings per article	15.8
Average section length (words)	327.8
Average total article length (words)	2,159.1
Average number of references	90.1

Note: 90.1 references per article. This is the "well-revised" baseline. STORM is a first draft, not a final article.

Source: paper-brief § 5. arXiv 2402.14207 Table 7.

4. The automatic evaluation

Source: paper-brief § 6.1. arXiv 2402.14207 § 4, 5.

Four metrics

METRIC	WHAT IT MEASURES
ROUGE-1, ROUGE-L	Overlap with human-written article (Lin, 2004)
Entity Recall	Percentage of named entities from the human article that appear in the generated article (FLAIR NER)
Prometheus rubric	LLM-as-judge on 5 criteria, 1-5 scale (Kim et al., 2023)
Citation Recall / Precision	Whether cited passages entail the generated sentence (Mistral-7B-Instruct judge)

Outline quality is measured separately by **heading soft recall** (Sentence-BERT cosine) and **heading entity recall** (FLAIR NER).

Article quality (Table 2)

CONDITION	ROUGE-1	ROUGE-L	ENTITY RECALL	INTEREST	ORG	COVERAGE
Direct Gen	25.62	12.63	5.08	2.87	4.60	4.16
RAG	28.52	13.18	7.57	3.14	4.22	4.08
oRAG	44.26	16.51	12.57	3.90	4.79	4.70
STORM	**45.82**	**16.70**	**14.10**	**3.99**	**4.82**	**4.88**
w/o Outline	26.77	12.77	7.39	3.33	4.87	3.35

GPT-4 used for final article generation. STORM significantly beats the strongest baseline (oRAG) on ROUGE-1, entity recall, and the rubric criteria.

Outline quality (Table 3, GPT-3.5)

CONDITION	HEADING SOFT RECALL	HEADING ENTITY RECALL
Direct Gen	80.23	32.39
RAG / oRAG	73.59	33.85
RAG-expand	74.40	33.85
STORM	**86.26**	**40.52**
w/o Perspective	84.49	40.12
w/o Conversation	77.97	31.98

Reference coverage (Table 5)

CONDITION	UNIQUE REFERENCES COLLECTED
STORM full pipeline	**99.83**
w/o Perspective	54.36
w/o Conversation	39.56

STORM collects twice the unique references of the perspective-only ablation, and 2.5x the conversation-only ablation. The full pipeline is what produces the breadth.

Source: paper-brief § 6.1. arXiv 2402.14207 Table 5.

Citation quality (Table 4)

	RECALL	PRECISION
STORM	84.83	85.18

Citation quality is judged by Mistral-7B-Instruct on whether the cited passage entails the generated sentence. The paper notes that unsupported sentences are primarily "improper inferences and inaccurate paraphrasing, not hallucinated content".

Source: paper-brief § 6.1. arXiv 2402.14207 Table 4.

5. The human evaluation

Source: paper-brief § 6.2. arXiv 2402.14207 § 6.

Setup

- 10 experienced Wikipedia editors
- Minimum 500 edits, more than 1 year of experience
- 20 topics randomly sampled from FreshWiki
- Each topic: one STORM article, one oRAG article (the best baseline)
- Each pair rated by 2 editors on 1-7 scale across 5 criteria
- Plus pairwise preference and a 1-5 usefulness survey

Inter-annotator agreement (Krippendorff’s alpha) ranges from 0.221 to 0.388 (fair to moderate).

Source: paper-brief § 6.2. arXiv 2402.14207 § 6.

The 25% and 10% numbers (Table 6)

ASPECT	ORAG % ≥ 4	STORM % ≥ 4	P-VALUE
Interest Level	57.5%	**70.0%**	0.077
Organization	45.0%	**70.0%**	0.005
Relevance	62.5%	65.0%	0.347
Coverage	57.5%	**67.0%**	0.084
Verifiability	67.5%	67.5%	0.843
Pairwise preferred	14	**26**	

The unanimity finding (Figure 3)

The 10 editors were asked whether STORM is useful for their pre-writing stage.

100% agreed. 70% strongly agreed, 30% somewhat agreed.

80% said it would help them edit a Wikipedia article for a new topic. 70% said it could be useful for the Wikipedia community, with only 10% disagreeing.

This is the strongest qualitative signal in the paper.

Source: paper-brief § 6.2. arXiv 2402.14207 Figure 3.

6. What the editors flagged as wrong

Source: paper-brief § 6.3, 11. arXiv 2402.14207 § 6, § 8.

Four documented failure modes

1. **Red herring fallacy.** The model introduces connections between unrelated facts in the retrieved sources. The cited passage supports each fact individually, but the *connection* the model makes is not in the source. Hurts Verifiability.
2. **Overspeculation.** The model adds details not in the sources. Also hurts Verifiability.
3. **Neutrality transfer.** Seven of ten editors note that STORM-generated articles "sound emotional or unneutral". The model absorbs the tone of the most prominent internet sources. This is a bias issue, not a factual one.
4. **Quality gap to well-revised human articles.** STORM is a first draft. Human Wikipedia articles average 90 references and have been edited by many humans over months. STORM does not get there.

7. Co-STORM (2025 follow-up)

Source: paper-brief § 7. arXiv 2408.15232.

The follow-up paper

Jiang, Y., Shao, Y., Ma, D., Semnani, S. J., Lam, M. S. (2024). "Into the Unknown Unknowns: Engaged Human Learning through Participation in Language Model Agent Conversations." *ICLR 2025*.

- arXiv:2408.15232. v2 dated 17 Oct 2024.
- 5 authors, 4 from Stanford OVAL, 1 from Yale (Dekun Ma).
- The same STORM demo URL (`storm.genie.stanford.edu`) is now branded Co-STORM.

Source: paper-brief § 7. arXiv 2408.15232.

The shift: from QA to discourse

STORM is a QA system. The user has to know the right question. Co-STORM inverts this: the user is a *participant*, not a questioner.

Three roles:

- **Moderator.** Asks questions on the user's behalf to steer the discourse.
- **Expert agents.** Each holds a perspective (the same perspective discovery from STORM, repurposed as a live conversation).
- **Dynamic mind map.** The user-facing artifact. A live, branching outline that grows as the agents talk.

The user observes by default, can take a turn, can ask for a topic shift, or click to dive deeper on a mind-map branch.

Source: paper-brief § 7. arXiv 2408.15232 § 1, Figure 1.

WildSeek dataset

The Co-STORM analog of FreshWiki. Real information-seeking records with explicit user goals, collected for the evaluation. This is a contribution of the paper, not just a benchmark.

The original STORM paper studied *article generation* (a writing task). Co-STORM studies *information seeking* (a learning task). The datasets are designed for the different problems.

Source: paper-brief § 7. arXiv 2408.15232.

Co-STORM human eval (Table 4)

COMPARISON	ASPECT	BASELINE	CO-STORM	WIN% (LOSE%)	P-VALUE
vs Search Engine	Serendipity	2.70	**3.90**	70% (10%)	0.030
vs Search Engine	Breadth	3.60	**4.10**	50% (10%)	0.096
vs Search Engine	Depth	3.10	**4.00**	60% (10%)	0.081
vs RAG Chatbot	Serendipity	2.78	**3.78**	67% (0%)	0.009

8. The system today

Source: paper-brief § 8. github.com/stanford-oval/storm. storm.genie.stanford.edu.

Live system status (June 18 2026)

COMPONENT	STATUS
GitHub repo `stanford-oval/storm`	28,475 stars, 2,601 forks, 238 commits, MIT
Last push	30 Sep 2025
Last commit	"Merge pull request #400 from stanford-oval/patch/requirement-change"
Open issues	97
Separate `co-storm` repo	None (verified via GitHub API 404)

Common community uses

The 2,601 forks and 97 open issues suggest the following usage patterns:

- **RAG pipeline integration.** Use STORM as a backbone for chat-with-your-knowledge-base products.
- **Search backend swap.** Replace You.com with Bing, Brave, Tavily, SerpAPI, or a local retriever.
- **Model swap.** Replace GPT-3.5 / GPT-4 with open-weight models (Llama, Mistral, Qwen). Several open issues are about the system degrading when the LLM is changed.
- **Knowledge backbone.** Use STORM as the research layer for content products.

Source: paper-brief § 8. github.com/stanford-oval/storm/issues.

9. Adopters, related work, limitations

Source: paper-brief § 9, 10, 11. arXiv 2402.14207, 2408.15232.

Known adopters

From primary sources, the verified users of STORM are:

- **Stanford OVAL Lab** (Shao, Jiang, Lam, Khattab, Kanell, Xu). The developers.
- **10 experienced Wikipedia editors** in the human study. Anonymized in the paper.
- **Yijia Shao's PhD research.** Built on STORM and its descendants.
- **Yucheng Jiang.** Continues to publish follow-up work in the Co-STORM line.
- **Nav Toor.** The most visible popularizer. His June 17 2026 X Article (archived in this same folder) is a paraphrase of the paper into a 4-prompt Claude workflow.

No verified public commercial deployment found. Marked as: needs follow-up survey.

Source: paper-brief § 9. arXiv 2402.14207 § 6. X Article 2067194761446920264.

Related work: same lab

- **WikiChat** (Semnani et al., EMNLP Findings 2023, arXiv:2305.14292, `stanford-oval/WikiChat`, 1,602 stars). RAG chatbot grounded on English Wikipedia to stop hallucination. Same lab, same hallucination problem, complementary technique.
- **DSPy** (Khattab et al., 2023, arXiv:2310.03714). The framework STORM is built on. Same lab. STORM is one of the canonical DSPy programs.
- **Re3, DOC** (Yang et al., 2022, 2023). Recursive reprompting and detailed outline control for long story generation. Cited by STORM as related work on outline-driven generation.

Source: paper-brief § 10. arXiv 2402.14207 § 7.

Related work: other labs

The paper's related-work section (Section 7) groups prior work into three buckets. Key citations:

- **RAG foundations:** Lewis et al., 2020 (NeurIPS); Ram et al., 2023; Izacard et al., 2023 (Atlas); Bohnet et al., 2023.
- **Active retrieval:** Jiang et al., 2023 (EMNLP); Parisi et al., 2022 (TALM); Shuster et al., 2022; Yao et al., 2023 (ReAct).
- **Expository writing:** Balepur et al., 2023 (EMNLP); Shen et al., 2023.
- **Citation-grounded generation:** Menick et al., 2022; Gao et al., 2023 (EMNLP).
- **Wikipedia generation:** Liu et al., 2018 (WikiSum, ICLR); Fan and Gardent, 2022 (ACL); Sauper and Barzilay, 2009 (ACL).
- **Question asking:** Ram, 1991; Qi et al., 2020 (EMNLP Findings); Press et al., 2023 (EMNLP Findings).

Limitations (combined)

From the paper (Section 8):

- Quality gap to well-revised human articles
- Neutrality transfer (tone of dominant sources)
- Verifiability beyond factual hallucination (red herring, overspeculation)
- English-only
- Retrieval dependence (no content sifting)

From observed community use:

- Citation accuracy on long-tail topics
- Prompt sensitivity when the LLM is swapped
- No fact-checking module in the generation pipeline (only in evaluation)

The popular 4-prompt Claude workflow (Nav Toor's article) trades depth for cost. The full STORM loop is expensive; the 4-prompt version is a cheap

10. Mapping to Parallax

Source: paper-brief § 12. summary § What this means for Parallax.

The mapping

STORM CONCEPT	PARALLAX ANALOG
Long-form Wikipedia article	Shippable game spec
Perspective discovery (5 expert views)	Player, designer, art, monetisation, retention
Simulated conversation (5 rounds)	Spec refinement pass before code generation
Outline curation	The brief that the build pipeline consumes
Section-by-section writing with citations	Section-by-section code generation with reference to design doc

The 4-prompt Claude workflow

Nav Toor's popularization (June 17 2026, archived in this same folder) is a 4-prompt Claude re-implementation:

1. **Multi Perspective Scan.** Simulate 5 expert views.
2. **Contradiction Map.** Where do the 5 voices fight?
3. **Synthesis.** Research briefing with 5 key findings.
4. **Peer Review.** Confidence scores, bias check, missing perspectives.

5 minutes of human time. Trades the structural pre-writing loop of STORM for a cheap prompt chain.

Source: article § Phases 3-7. summary § The four prompts.

Three open questions for Assaf

1. **Where in the roadmap is "Deep Brainstorm"?** v1 is fast and forgiving. v3 might compete on depth. The four-prompt loop is a v3 feature, not a v1 one, unless the unit economics change.
2. **Does the contradiction map help non-expert creators?** Game design has a "designer" perspective most users have not internalised. Running the perspective loop might confuse them, not help. Worth a usability test before shipping.
3. **Should we cite the four-prompt workflow in the investor deck?** 25 percent more organized, 10 percent broader are concrete, externally validated numbers. A single sentence in the "Why now" slide is enough.

Source: paper-brief § 12. summary § Open questions for Assaf.

References

Source files

PATH	USED FOR
`raw data/storm-method/paper-brief.md`	All § 1-12, the academic layer
`raw data/storm-method/article.md`	§ Phases 3-7, the popular 4-prompt body
`raw data/storm-method/summary.md`	§ What STORM is, § The four prompts, § Mapping to Parallax
`raw data/storm-method/how-to-fill.md`	Provenance record of the archive

Get in touch

hello@thriv.es · Set a call

Source: paper-brief § 1-14. arXiv 2402.14207, 2408.15232.

Feedback