PERELMAN: Pipeline for scientific literature meta-analysis

Daniil Sherki [1 2] **Daniil Merkulov** [1 2] **Alexandra Savina** [2] **Artem Abakumov** [2] **Ekaterina Muravleva** [1 2]
[1] *Sberbank of Russia, AI4S Center, Moscow, Russian Federation* [2] *Skolkovo Institute of Science and Technology, Moscow, Russian Federation*. Correspondence to: Daniil Sherki, Daniil Merkulov {dssherki,dmmerkulov}@sberbank.ru.

1. Introduction

The volume of scientific publications is growing at an unprecedented rate. Preprint platforms such as arXiv now receive tens of thousands of new submissions per month [1], and the projected number of arXiv submissions in 2025 is roughly twice that of 2019 [2]. In parallel with the rapid growth of preprint platforms, the Crossref DOI registry, which covers journal and conference publishing, reports metadata for ~178 million records, including over 120 million journal-article DOIs and nearly 10 million conference-proceedings DOIs (updated Dec 25, 2025) [3].

Many research areas involve rapidly evolving domain-specific knowledge, making the construction of comprehensive and reproducible meta-analyses increasingly challenging across chemistry, medicine, materials science, physics, and machine learning.

We present a pipeline for agentic processing of large amounts of data from scientific papers. We call our pipeline *PERELMAN* (PipEline foR sciEntific Literature Meta-ANalysis). This agentic pipeline aims to extract target data from large collections of domain-specific articles to support expert-driven meta-review writing.

2. PERELMAN

2.1 Architecture

Agentic System Architecture The agentic system consists of three components: a subsystem for domain knowledge extraction, an article parsing module, and an agent that uses image processing tools to extract data from texts and figures.

The overall architecture is shown in Fig. 1.

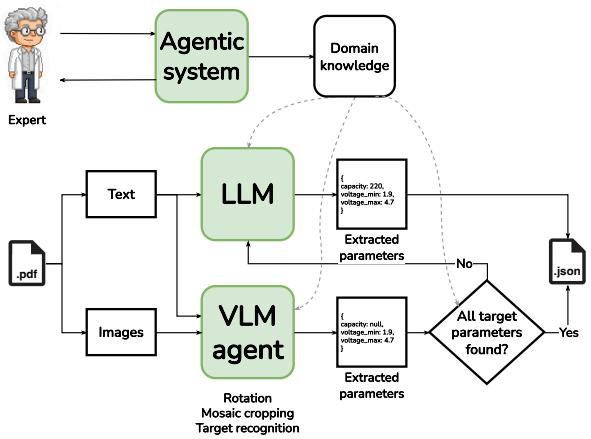

Fig. 1: Principal *PERELMAN* architecture scheme.

2.2 Domain Knowledge

This component acquires domain knowledge through dialogue with a human expert and applies it to the analysis of scientific literature. The knowledge guides the entire pipeline, from problem formulation and target selection to prompt design and downstream reasoning. Its integration into the system is shown in Fig. 1.

2.3 Article Parsing Module

We use an OCR system that converts tables and layouts into a text-only large language model (LLM) readable format. In our setup, `Docling`[4, 5] provided the best performance. Each PDF is transformed into structured Markdown, while figure images are stored separately with captions, textual descriptions, and all mentions of the figure in the article text preserved.

2.4 VLM Agent with Tools

Figure Data Extraction Extraction is multimodal. For each figure, we build a context from the caption and textual mentions, which is passed to the VLM to ground the extraction in local descriptions.

Mosaic Normalization The VLM predicts a strict JSON schema describing the mosaic layout (example in Fig. 2), panel labels, target panel, and required image rotation. The grid is corrected if needed, after which the image is split into panels and oriented consistently. The agent then normalizes the image (crops the target panel, rotates it if needed) using available agent tools.

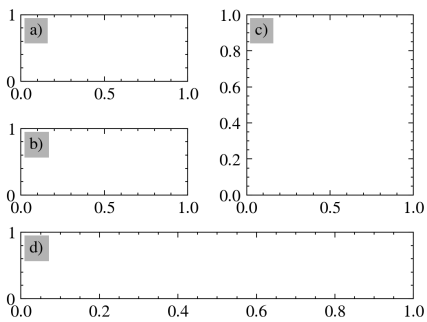

Fig. 2: Example of a subplot mosaic representation.

The VLM model represents Fig. 2 as a structured `mosaic` encoding, internally perceived as follows:

```
ac
bc
dd
```

Multimodal Metric Extraction After normalization, a large VLM extracts target parameters in JSON format using the system prompt and the acquired domain knowledge. If visual extraction fails, we fall back to a text-only LLM that processes the LLM-readable Markdown format text in overlapping

chunks.

Outputs The pipeline produces a JSON file with the extracted metrics, which is stored in the output database for expert review and meta-analysis.

3. Experiments

Layered $LiNi_xMn_yCo_zO_2$ (NMC) oxides are key cathode materials for lithium-ion batteries. The Ni-rich NMC811 delivers high specific capacity but suffers from complex degradation, leading to fragmented and inconsistent reports of its performance. A recent meta-analysis by A. Savina and A. Abakumov [6] consolidated this literature into benchmark distributions of core electrochemical metrics. In this work, we reuse their article corpus and take these benchmarks as the basis for our analysis.

4. Results

4.1 NMC811 Meta-analysis Validation

We validated our framework on the article set from [6] and compared it against a VLM-based baseline that performs page-level extraction. As shown in Table 1, our pipeline consistently outperforms the baseline in extracting key electrochemical metrics.

We further analyzed the impact of system configurations. Increasing the size of the VLM used for chart annotation leads to a systematic gain in overall accuracy (Table A1). In the current setup, the image normalization module achieves an effective accuracy of **62.5%**, providing a robust foundation for downstream extraction.

Source-type analysis reveals that about **70%** of target parameters originate from charts (Table A2), while charts remain the most challenging source modality. This makes chart-centric question answering a critical capability of the system.

Table 1: Extraction accuracy by metric type for different systems (in %).

System	Cap. (1st)	Coul. eff.	$V_{\max}$	$V_{\min}$
Ours	**58.33**	**33.33**	**95.83**	70.80
GPT-5.2 Pro Extended Thinking	25.00	0.00	79.20	**79.20**
GPT-5.2 Thinking Heavy	8.30	8.30	58.30	58.30

4.2 Chart-based Question Answering

We evaluated a range of state-of-the-art VLMs on extracting information from scientific charts. Consistent with existing benchmarks, baseline performance remains far below human-level accuracy (Table A3).

A decisive improvement comes from prompt tuning with expert domain knowledge. Instead of generic instructions, we provide precise metric definitions, typical visual contexts, and expert-like reading strategies. This leads to an order-of-magnitude increase in extraction accuracy from images, raising performance from 4.55% to **80.95%** for first-cycle discharge capacity (Table A4).

5. Discussion and Limitations

The main limitation of our system lies in the strong dependence of extraction quality on domain knowledge: even large VLMs perform poorly on images when such knowledge is scarce, while expert input yields a sharp improvement. In addition, the end-to-end performance is bounded by the weakest component, particularly the target-image classifier. Finally, the complex structure of scientific figures often violates the assumptions behind our current mosaic-based processing.

Future work will focus on improving mosaic handling, automating domain knowledge acquisition from expert interactions, and adding watermark removal for scientific figures.

6. Related Work

Automation of evidence synthesis has progressed most visibly in biomedicine, where systems such as RobotReviewer and ASReview partially automate systematic reviews [7, 8]. Scalable literature processing is enabled by infrastructure resources, including machine-readable corpora and PDF parsers such as S2ORC, GROBID, and PubLayNet [9, 10, 11].

In scientific NLP, the focus remains on text: models and benchmarks such as SciBERT and SciERC/SciIE, together with domain-specific datasets like POLYIE, underpin knowledge extraction [12, 13, 14]. In materials science, however, key results often appear in charts, and even specialized ChartQA-style and in-domain benchmarks highlight the difficulty of multimodal extraction [15, 16].

Historically, chemistry and materials extraction has relied on rule-based and hybrid systems, including ChemDataExtractor [17, 18]. Recent evaluations of LLMs show that robust extraction still requires explicit schemas and provenance [19]. In this setting, agentic tool-use approaches such as ReAct and Toolformer provide a promising framework for modular and interpretable scientific pipelines [20, 21]. Memory-augmented Transformer approaches such as RATE [22] and ELMUR [23] represent a promising direction for extending this framework, as they enable tool-using agents to retain and revise key context, such as experimental conditions and metric definitions, across long scientific documents.

7. Conclusions

We introduced *PERELMAN*, an agentic framework for multimodal literature meta-analysis, and validated it on the NMC811 benchmark, where domain-guided tool use substantially improves extraction accuracy over VLM-only baselines. The main remaining bottleneck is robust chart understanding, motivating improved figure handling and scalable domain knowledge acquisition.

References

[1] arXiv. arxiv sets new record for monthly submissions, again, 2024. Accessed: 2025-01-26.

[2] Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, et al. Surveyx: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776*, 2025.

[3] Crossref. Crossref status report (crossref stats page). Website, 2025. Updated Dec 25, 2025.

[4] Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikolaos Livathinos, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Fabian Lindlbauer, Kasper Dinkla, Lokesh Mishra, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. Docling technical report, 2024.

[5] Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. Docling: An efficient open-source toolkit for ai-driven document conversion, 2025.

[6] Aleksandra A. Savina and Artem M. Abakumov. Benchmarking the electrochemical parameters of the lini0.8mn0.1co0.1o2 positive electrode material for li-ion batteries. *Heliyon*, 9(12):e21881, 2023.

[7] Iain J. Marshall, Joël Kuiper, and Byron C. Wallace. Robotreviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23(1):193–201, 2015.

[8] Rens van de Schoot, Jonathan de Bruin, Raoul Schram, Parisa Zahedi, Jan de Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, Albert Harkema, Joukje Willemsen, Yongchao Ma, Qixiang Fang, Sybren Hindriks, Lars Tummers, and Daniel Oberski. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2):125–133, 2021.

[9] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[10] Patrice Lopez. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27–October 2, 2009, Proceedings*, pages 473–474. Springer, 2009.

[11] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis, 2019.

[12] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019.

[13] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

[14] Jerry Cheung, Yuchen Zhuang, Yinghao Li, Pranav Shetty, Wantian Zhao, Sanjeev Grampurohit, Rampi Ramprasad, and Chao Zhang. POLYIE: A dataset of information extraction from polymer material scientific literature. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2370–2385, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[15] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022.

[16] Defne Circi, Miles Bradley, Sam Blouir, Boris Wilthan, Antonis Anastasopoulos, Amarda Shehu, L. Catherine Brinson, and Bhuwan Dhingra. Information extraction from diverse charts in materials science, 2025. Workshop paper at COLM 2025.

[17] Matt E. Swain and Jacqueline M. Cole. Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904, 2016.

[18] Shu Huang and Jacqueline M. Cole. A database of battery materials auto-generated using chemdataextractor. *Scientific Data*, 7(1):260, 2020.

[19] Luca Foppiano, Guillaume Lambard, Toshiyuki Amagasa, and Masashi Ishii. Mining experimental data from materials science literature with

large language models: an evaluation study. *Science and Technology of Advanced Materials: Methods*, 2024.

[20] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2022.

[21] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Luke Lombardo, Luke Zettlemoyer, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools, 2023.

[22] Egor Cherepanov, Alexey Staroverov, Dmitry Yudin, Alexey K Kovalev, and Aleksandr I Panov. Recurrent action transformer with memory. *arXiv preprint arXiv:2306.09459*, 2023.

[23] Egor Cherepanov, Alexey K Kovalev, and Aleksandr I Panov. Elmur: External layer memory with update/rewrite for long-horizon rl. *arXiv preprint arXiv:2510.07151*, 2025.

Appendix A. Resulting tables

Table A1: Extraction quality by VLM configuration (in %).

VLM	Text model	Cap.	Coul.	$V_{\max}$	$V_{\min}$
—	gpt-oss-120b	19.23	11.54	42.31	30.77
granite-2b	gpt-oss-120b	50.00	19.23	80.77	69.23
granite-8b	gpt-oss-120b	53.85	30.77	84.62	73.08

Table A2: Distribution of target-parameter source types.

Source	Fraction (%)
Text	29
Table	1
Chart	70

Table A3: VLM evaluation on chart-based information extraction (in %).

Model	Overall	Cap.	Coul.	$V_{\max}$	$V_{\min}$
Claude 3.7 Sonnet (Vision)	30.68	0.00	0.00	72.73	50.00
Claude 4.5	0.00	0.00	0.00	0.00	0.00
Gemma 3 27B IT	0.00	0.00	0.00	0.00	0.00
GLM 4.5V	7.95	0.00	0.00	18.18	13.64
GPT-5 Mini	9.09	0.00	0.00	13.64	22.73
Grok-4	3.57	0.00	0.00	7.14	7.14
Nemotron-Nano-12B-V2-VL	15.28	0.00	0.00	27.78	33.33
Qwen 3 VL 235B	17.05	4.55	0.00	31.82	31.82

Table A4: Effect of domain knowledge on extraction accuracy (Qwen3-VL-235B, in %).

Approach	Cap.	Coul.	$V_{\max}$	$V_{\min}$
Default prompt w/o domain knowledge	4.55	0.00	31.82	31.82
After prompt tuning and normalization	80.95	36.36	95.65	73.91