Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Demonstrations Track

PERELMAN: Pipeline for Scientific Literature Meta-Analysis

Daniil Sherki[1,2], **Daniil Merkulov**[1,2], **Alexandra Savina**[3], **Dzhantemir Kikov**[4], **Dmitry Parpulov**[4], **Alexander Ivanov**[4], **Artem Abakumov**[3], **Ekaterina Muravleva**[1,2]
[1]AI4Science Center, Moscow, Russia
[2]Applied AI institute, Moscow, Russia
[3]Center for Energy Science and Technology, Moscow, Russia
[4]SB AI Lab, Moscow, Russia
{dssherki,dmmerkulov}@sberbank.ru

Abstract

We present PERELMAN (PipEline foR sciEntific Literature Meta-ANalysis), an agentic framework designed to extract specific information from a large corpus of scientific articles to support large-scale literature reviews and meta-analyses. Our central goal is to reliably transform heterogeneous article content into a unified, machine-readable representation. PERELMAN first elicits its domain knowledge-including target variables, inclusion criteria, units, and normalization rules-through a structured dialogue with a subject-matter expert. This domain knowledge is then reused across multiple stages of the pipeline and guides coordinated agents in extracting evidence from narrative text, tables, and figures, enabling consistent aggregation across studies. In order to assess reproducibility and validate our implementation, we evaluate the system on the task of reproducing the meta-analysis of layered Li-ion cathode properties $LiNi_{0.8}Mn_{0.1}Co_{0.1}O_2$ (NMC811). We describe our solution, which has the potential to reduce the time required to prepare meta-analyses from months to minutes.

1 Introduction

Scientific literature is growing faster than researchers can synthesize it: the number of published papers in 2025 is roughly twice that of 2019 [Liang *et al.*, 2025], while Crossref already indexes on the order of 10^8 DOI records [Crossref, 2025]. In many experimental domains, this growth hinders reproducible evidence synthesis because key metrics must be aggregated across many partially inconsistent studies.

We introduce PERELMAN (PipEline foR sciEntific Literature Meta-ANalysis), an agentic pipeline that extracts target variables from domain papers into auditable structured records for expert meta-reviews. PERELMAN combines (i) expert-guided domain knowledge elicitation, (ii) layout-aware PDF parsing into an LLM-readable representation (text + figure assets), and (iii) figure-centric multimodal extraction with image normalization and a text fallback. The system outputs a per-article `json` record with provenance for selective human verification.

The approach builds on automated evidence-synthesis tools [Marshall *et al.*, 2015; van de Schoot *et al.*, 2021], scientific document infrastructure [Lo *et al.*, 2020; Lopez, 2009; Zhong *et al.*, 2019], text-focused scientific IE [Beltagy *et al.*, 2019; Luan *et al.*, 2018; Cheung *et al.*, 2024], and chart benchmarks that expose persistent numerical-reasoning gaps in VLMs [Masry *et al.*, 2022; Wang *et al.*, 2024b; Circi *et al.*, 2025]. In materials and chemistry, this is aligned with hybrid extraction pipelines that combine models, schemas, and normalization [Swain and Cole, 2016; Huang and Cole, 2020; Foppiano *et al.*, 2024]. At the workflow level, PERELMAN follows tool-using agentic designs [Yao *et al.*, 2022; Schick *et al.*, 2023]; memory-augmented architectures such as RATE and ELMUR are relevant for long-horizon context tracking [Cherepanov *et al.*, 2023; Cherepanov *et al.*, 2025].

We evaluate PERELMAN on a subset of the NMC811 benchmark meta-analysis dataset [Savina and Abakumov, 2023]. The study focuses on extracting first-cycle discharge capacity, first-cycle coulombic efficiency, and voltage range, and shows that domain-knowledge-enhanced prompting improves over a page-wise VLM baseline in this single-domain demo setting.

A video demonstrating its use is available at[1].

2 Pipeline

2.1 Agentic System Architecture

The architecture of the agentic system consists of three major components:
(i) agentic system for extracting domain knowledge,
(ii) an article-parsing tool,
(iii) an agent for extracting data from the parsed articles, using domain knowledge.

The principal architecture is presented in Figure 1.

2.2 Domain knowledge

Domain knowledge is represented as a structured object via a three-step pipeline. A clarification agent first resolves ambiguity by asking 2–5 focused questions, with expert answers appended to the research context. A constraints agent then converts the enriched description into a typed schema {`query_understanding`, `entity_type`,

[1]https://youtu.be/RL9MOHb3uD0

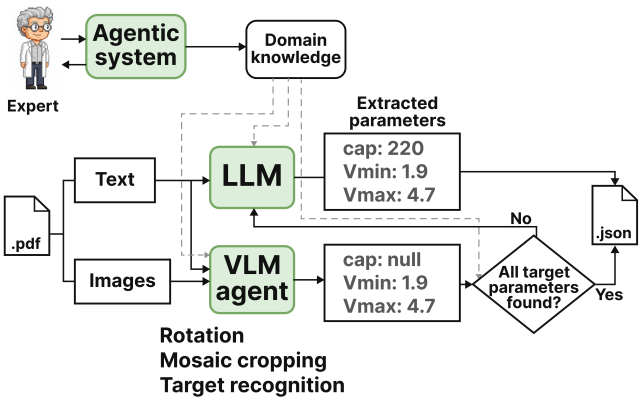

Figure 1: Principal PERELMAN Architecture scheme

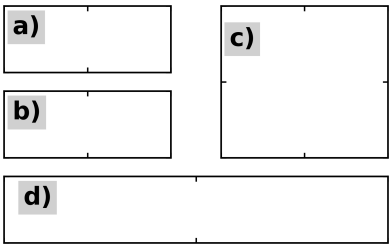

Figure 2: Subplot mosaic representation. The VLM encodes the panel layout as `ac/bc/dd`; each letter maps to a subplot identifier.

`entity_description, limitations, search_queries`}, where `limitations` are verifiable conditions and queries specify domain and recency.

2.3 Data Collection

The module runs parallel `Tavily` searches, extracts typed candidates with provenance using LLM batch analysis, deduplicates them by normalized names plus an LLM `DedupPlan`, and verifies candidates with 1–2 combined queries before PDF metric extraction.

2.4 Article Parsing

The core parser can be any OCR system that preserves table structure and produces LLM-readable output. In our setup, the best results were obtained with `Docling` [Auer *et al.*, 2024; Livathinos *et al.*, 2025]. We convert each PDF into layout-aware Markdown and extract figure images into a dedicated folder, preserving captions when available.

2.5 VLM Agent

Figure data extraction

Data extraction is multimodal. For each figure, we build extra context: the caption plus a ± 1000 character window around figure mentions in Markdown. This context is passed to the vision model to ground extraction in nearby textual descriptions.

We use a VLM to predict a strict JSON schema that includes mosaic grid (rows/cols), panel labels, optional target panel, relevance flag, and required rotation. If the model outputs a 1x1 grid but labels indicate multiple panels, we correct the grid; if labels are absent, we apply a simple aspect-ratio heuristic and split the image into panels. The same invocation predicts rotation ($0°$, $90°$, $180°$, $270°$), and the figure is rotated if needed.

Multi-modal Metric Extraction

After image normalization, a larger VLM processes the target image together with the full **extra content** from Section 2.5. Target parameters are extracted in `json` format according to the expert-defined **domain knowledge** (Section 2.2). If image extraction fails, we fall back to a text-only LLM; in our default setup, the extractor uses `GPT-oss 120B` [OpenAI *et al.*, 2025]. Text extraction runs on overlapping Markdown chunks, and first non-null values are merged into the final record.

Outputs

At the end of the pipeline, we produce a `json` file with extracted target metrics for downstream expert analysis and meta-review writing. These records are candidate extractions, and fields with lower reliability, especially chart-derived Coulombic efficiency, remain targets for human verification. In this setup, the cropping module reaches 70.83% accuracy. Given correctly cropped images, the target-image classifier reaches 88.25%. The effective Image Normalizer accuracy is therefore $70.83\% \cdot 88.25\% = 62.5\%$.

Table 3 shows that most target information is located in figures. At the same time, chart extraction remains substantially weaker than text/table extraction. In an earlier pipeline version, extraction accuracy was 86% for text, 100% for tables, but only 34.6% for charts.

Chart-based Question Answering Specialized chart-QA benchmarks show that modern VLMs still do not reach human-level performance on scientific charts [Wang *et al.*, 2024b]. We evaluated top models from this benchmark, together with several recent VLMs, on retrieval of target chart values. Results are presented in Table 4. After prompt tuning, extraction was run only on expert-selected, cropped, and normalized target images. The main tuning change was explicit injection of expert domain knowledge: metric definitions, typical reporting locations, and chart-reading heuristics. As shown in Table 5, this improves target-parameter extraction by up to an order of magnitude (from 4.55% to 80.95% for one metric).

3 Experiments

Layered oxides of the form $\mathrm{LiNi}_x\mathrm{Mn}_y\mathrm{Co}_z\mathrm{O}_2$ (NMC, $x + y + z = 1$) are key Li-ion cathode materials. The Ni-rich composition $\mathrm{LiNi}_{0.8}\mathrm{Mn}_{0.1}\mathrm{Co}_{0.1}\mathrm{O}_2$ (NMC811) provides high specific capacity ($\approx 200\ \mathrm{mAh\,g^{-1}}$) and is widely studied for high-energy cells [Li *et al.*, 2015]. At the same time, NMC811 suffers from complex degradation mechanisms, which leads to partially inconsistent metrics across the literature [Jung *et al.*, 2017]. Savina and Abakumov addressed this with a large-scale manual meta-analysis of NMC811, aggregating 548 articles and more than 950 records [Savina and Abakumov, 2023]. Their benchmark highlights

System	First Discharge Capacity	Coulombic Efficiency	Voltage (max)	Voltage (min)
Ours	**58.33%**	**33.33%**	**95.83%**	70.80%
Ours (MinerU parser)	54.17%	**33.33%**	87.50%	**87.50%**
Ours (DoclingOCR)	53.85%	30.77%	84.62%	73.08%
Ours (DeepSeekOCR)	53.85%	23.08%	84.62%	73.08%
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%

Table 1: Accuracy by type for different systems

VLM	Text	First cycle Discharge Capacity	Coulombic efficiency	Voltage (max)	Voltage (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 2: Extraction quality by type (%)

Source	Text	Table	Chart
Fraction	29%	1%	70%

Table 3: Target parameters source type distribution

Model	Avg.	Cap.	CE	$V_{\max}$	$V_{\min}$
Claude 3.7 Sonnet	30.7	0.0	0.0	72.7	50.0
Gemma 3 27B	0.0	0.0	0.0	0.0	0.0
GLM 4.5V	8.0	0.0	0.0	18.2	13.6
GPT-5 Mini	9.1	0.0	0.0	13.6	22.7
Grok-4	3.6	0.0	0.0	7.1	7.1
Nemotron-12B-V2	15.3	0.0	0.0	27.8	33.3
Qwen 3 VL 235B	17.1	4.6	0.0	31.8	31.8

Table 4: VLM evaluation results (default prompt, no domain knowledge)

Approach	Cap.	CE	$V_{\max}$	$V_{\min}$
Default	4.55%	0.00%	31.82%	31.82%
Tuned + normalized	80.95%	36.36%	95.65%	73.91%

Table 5: Prompt-tuning accuracy for qwen3-vl-235B (3.0% tolerance)

both the value of systematic synthesis and the high manual effort required. We reproduce a 24-article expert-grounded subset: each target value is linked to its source location, including the specific figure or panel and extraction procedure. For each record, the reference meta-analysis reports: first-cycle discharge capacity ($\mathrm{mAh\,g^{-1}}$), first-cycle Coulombic efficiency (%), upper cut-off voltage and C-rate.

4 Results

Evaluation metric. A prediction is counted as correct if it falls within $\pm 3\%$ relative tolerance of the benchmark value in [Savina and Abakumov, 2023]. For voltage ranges, both endpoints must satisfy this criterion. Table 1 reports $N = 24$ results. On a larger $n = 200$ run with the same setup, accuracies were 56.00%, 49.50%, 86.70%, and 85.64% for capacity, Coulombic efficiency, voltage maximum, and voltage minimum, respectively.

In Table 1, "Ours" denotes the full pipeline with `granite-8b` for figure annotation and `GPT-oss 120B` for text extraction. The `MinerU` [Wang *et al.*, 2024a] parser variant is close to `Docling` on the main metrics, but was more than $2\times$ slower in our sampled runs. "DoclingOCR" and "DeepSeekOCR" replace the figure-annotation VLM, while GPT-5.2 rows are page-wise baselines without pipeline components. We also analyze configuration effects; for example, Table 2 shows that larger chart-annotation VLMs improve end-to-end extraction accuracy.

We observed three main challenges: chart extraction with minimal domain context, component coupling through target-image classification, and complex multi-panel figures that violate near-uniform mosaic splitting. The low Coulombic-efficiency scores also show that reliability is uneven across scientific parameters.

5 Conclusions

We introduced PERELMAN, an agentic pipeline for multimodal literature meta-analysis. Combining layout-aware parsing with figure-centric extraction, it outperforms VLM-only baselines on an expert-grounded NMC811 benchmark with figure/panel-level source annotations. PERELMAN reduces manual effort by producing auditable candidate records for expert verification; future work will improve complex multi-panel figure handling and test schema-based adaptation beyond NMC811.

References

[Auer *et al.*, 2024] 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. arXiv preprint arXiv:2408.09869, 2024.

[Beltagy *et al.*, 2019] 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)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.

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

[Cherepanov *et al.*, 2025] 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.

[Cheung *et al.*, 2024] 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.

[Circi *et al.*, 2025] 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.

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

[Foppiano *et al.*, 2024] 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*, 4(1), 2024. Article 2356506.

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

[Jung *et al.*, 2017] Roland Jung, Michael Metzger, Filippo Maglia, Christoph Stinner, and Hubert A. Gasteiger. Oxygen release and its effect on the cycling stability of linixmnycozo2 (nmc) cathode materials for li-ion batteries. *Journal of The Electrochemical Society*, 164(7):A1361–A1377, 2017.

[Li *et al.*, 2015] Jing Li, Laura E. Downie, Lin Ma, Wenda Qiu, and Jeff R. Dahn. Study of the failure mechanisms of lini0.8mn0.1co0.1o2 cathode material for lithium ion batteries. *Journal of The Electrochemical Society*, 162(7):A1401–A1408, 2015.

[Liang *et al.*, 2025] 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.

[Livathinos *et al.*, 2025] 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. arXiv preprint arXiv:2501.17887, 2025.

[Lo *et al.*, 2020] 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*, pages 4969–4983, Online, July 2020. Association for Computational Linguistics.

[Lopez, 2009] 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.

[Luan *et al.*, 2018] 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*, pages 3219–3232, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.

[Marshall *et al.*, 2015] 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.

[Masry *et al.*, 2022] 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. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[OpenAI *et al.*, 2025] OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, et al. gpt-oss-120b & gpt-oss-20b model card. arXiv preprint arXiv:2508.10925, 2025.

[Savina and Abakumov, 2023] 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.

[Schick *et al.*, 2023] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom.

Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761, 2023.

[Swain and Cole, 2016] 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.

[van de Schoot *et al.*, 2021] 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.

[Wang *et al.*, 2024a] Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. Mineru: An open-source solution for precise document content extraction. arXiv preprint arXiv:2409.18839, 2024.

[Wang *et al.*, 2024b] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. arXiv preprint arXiv:2406.18521, 2024.

[Yao *et al.*, 2022] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629, 2022.

[Zhong *et al.*, 2019] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. arXiv preprint arXiv:1908.07836, 2019.