A Flexible, Code-Free, and Scalable Workflow for Structured Data Extraction from CO_2 Electrochemical Reduction Literature

Xiong Ying[*1], Dai Haiwen[*1, 2], MahsaSadat Adel Rastkhiz[1], Danny Ren Zekun[2], Kedar Hippalgaonkar[1, 2]
[*]*Equal contribution*

[1] *School of Material Science and Engineering, 50 Nanyang Avenue, Nanyang Technological University, Singapore 639798*

[2]*The Berkeley Education Alliance for Research in Singapore, #11-00, 1 Create Way, Singapore 138602*

Correspondence to: Kedar Hippalgaonkar

1. Introduction

Data-driven approaches are increasingly central to scientific discovery in the era of artificial intelligence. Unlike images or natural language corpora, experimental research data are inherently costly and difficult to curate, owing to their high dimensionality, hierarchical organization, and multimodal presentation. Electrochemical CO_2 reduction (CO_2RR) is a rapidly expanding research field motivated by carbon neutrality, yet the associated experimental data remain largely inaccessible in machine-readable form.

Comprehensive structured datasets in CO_2RR must capture catalyst composition and structure, synthesis and processing details, electrochemical testing conditions, and performance metrics, often exceeding 100 distinct fields per publication. Automated extraction of such information from the literature remains challenging. Prior approaches, ranging from rule-based systems and classical natural language processing models (e.g., BERT, CLSTM, and named entity recognition pipelines) to more recent agent-based and large language model (LLM) methods, have been limited by extraction accuracy (typically 80–90% F1), restricted flexibility, and substantial computational cost, often requiring expert intervention, model retraining, or dedicated GPU infrastructure [1], [2], [3], [4]. Moreover, existing methods generally fail to recover quantitative information embedded in figures and plots, leading to systematic data loss.

Here, we present a flexible, code-free, and scalable workflow for structured data extraction from CO_2 electrolysis literature (Fig. 1). The workflow combines an object-oriented data schema encompassing catalyst, process, and performance attributes with carefully engineered prompt sets, few-shot examples, model selection, and a multi-model voting strategy. Without explicit annotation or model retraining, the approach achieves near-supervised extraction performance. The workflow reaches approximately 95% accuracy across key structured fields, at a per-document cost of USD 0.01–0.05 and a processing time of 30–60 seconds.

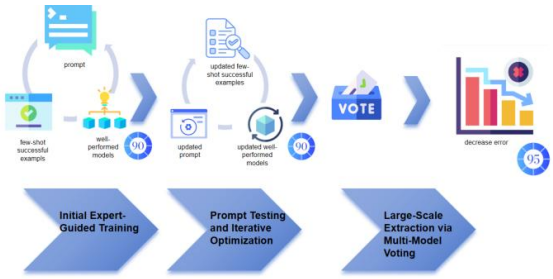

Fig.1 End-to-end data extraction workflow from schema engineering, expert-in-the-loop prompt optimization, to multi-model voting.

2. Substantial section

2.1 Data Collection and Preprocessing

Representative CO_2RR research articles, including associated supplementary information, were collected in PDF format. An initial set of ten publications was selected to span a broad range of catalyst materials, electrochemical cell configurations, and reporting styles. All documents were converted into machine-readable Markdown format using Docling OCR. Graphical elements, including figures, schemes, and tables, were separated from the main text to reduce noise during text-based extraction.

2.2 Schema Definition

An object-oriented data schema was designed to standardize information extraction across publications (Fig. 2). **The schema comprises primary classes: catalyst**, **testing environment**,

catalyst process, and **performance,** with shared sub-classes such as geometry and material attributes. This hierarchical structure enables consistent representation of complex experimental information while maintaining flexibility across diverse reporting formats.

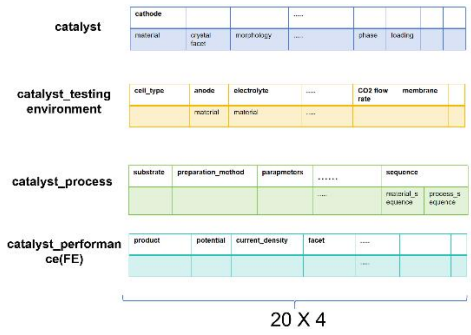

Fig.2 Object-oriented data schema for structured extraction from CO_2RR literature.

2.3 Expert-in-the-Loop Prompt and Example Optimization

Initial experiments confirmed that a single, generic LLM prompt is insufficient for reliable extraction of complex experimental data. A naïve baseline prompt, which simply instructed the model to populate the predefined schema, exhibited systematic errors, including (i) hallucination of information not present in the source text, (ii) incorrect material names or stoichiometric ratios, and (iii) cross-contamination between textual descriptions and graphical content. We adopted an iterative, expert-in-the-loop optimization strategy combining prompt refinement with carefully selected few-shot examples (Fig. 3). Prompts were progressively refined to explicitly define extraction boundaries, normalization rules, and exclusion criteria, until the performance converged on the initial ten-paper development set.

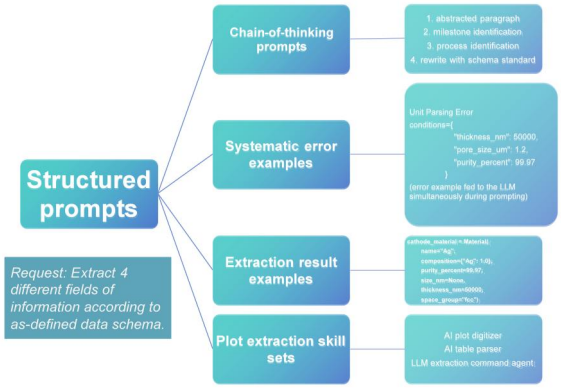

Fig.3 Prompt and example structure for accurate data extraction.

2.4 Extraction Performance

Using single-model best model (MiniMax 2.0), the extraction pipeline achieves an overall accuracy of ~87%. Field-specific accuracies for catalyst, process, and electrochemical testing data are 85.95%, 87.48%, and 88.40%, respectively, at a cost of USD ~0.03 per paper. Incorporating a multi-model voting strategy further improves performance. By default, two inexpensive or free models (GPT-OSS 120B and Nemotron 3 Nano) are used, while two higher-cost models (MiniMax 2.0 and GLM 4.6) are invoked only when initial model outputs disagree. This approach increases overall accuracy to ~95%, corresponding to absolute gains of 6–18% field-wise, without a significant change in per-document cost (USD 0.01–0.05).

References

[1] W. Ning, M. Li, J. R. Reimers, and R. Kobayashi, “Optimizing data extraction from materials science literature: a study of tools using large language models,” *Digital Discovery*, p. 10.1039.D5DD00482A, 2026, doi: 10.1039/D5DD00482A.

[2] X. Wang, A. Raj, M. Luebbe, H. Wen, S. Xu, and K. Lu, “Reliable End-to-End Material Information Extraction from the Literature with Source-Tracked Multi-Stage Large Language Models,” 2025, *arXiv*. doi: 10.48550/ARXIV.2510.05142.

[3] L. Wang *et al.*, “A corpus of CO2 electrocatalytic reduction process extracted from the scientific literature,” *Sci Data*, vol. 10, no. 1, p. 175, Mar. 2023, doi: 10.1038/s41597-023-02089-z.

[4] J. Choi *et al.*, “Deep learning of electrochemical CO_2 conversion literature reveals research trends and directions,” *J. Mater. Chem. A*, vol. 11, no. 33, pp. 17628–17643, 2023, doi: 10.1039/D3TA02780E.