Title

ChemQA: Retrieval-Augmented Expert Responses for Organic Electrocatalysis

Authors

Zemao Chen

Abstract

The deployment of generic large language models (LLMs) for domain-specific scientific inquiry is fundamentally constrained by their propensity to generate hallucinations—confidently stated but factually inaccurate responses—particularly when addressing intricate questions requiring specialized literature access, as exemplified in organic electrocatalysis. To bridge this gap, we developed ChemQA, a retrieval-augmented question-answering system that synergizes semantic literature retrieval from a curated vector database with evidence-based response generation by general-purpose LLMs. System validation demonstrated significant mitigation of hallucination risks: ChemQA exhibited superior performance across four evaluation metrics (content accuracy; literature authenticity; logical integrity; practical value). Critically, forensic analysis revealed only 8% unverifiable references in ChemQA versus 91.1% fabricated citations in conventional LLMs. This architecture establishes a new paradigm for literature-grounded scientific assistance, enabling precise knowledge extraction for sustainable chemistry innovation while maintaining zero hallucination tolerance.

MAIN TEXT

Introduction

Organic electrocatalysis has emerged as a pivotal field for sustainable chemical synthesis and energy conversion, enabling precise molecular transformations under mild electrochemical conditions. This rapidly evolving discipline demands sophisticated integration of knowledge across electrochemistry, organic reaction mechanisms, and materials science—particularly as researchers pursue advanced catalyst design, mechanistic elucidation, and scalable reactor engineering. While contemporary large language models (LLMs) exhibit considerable competence in general scientific discourse, their effectiveness deteriorates markedly when confronted with highly specialized queries characteristic of domains like organic electrocatalysis. A fundamental constraint stems from these models' static knowledge bases, which lack dynamic access to cutting-edge peer-reviewed literature and proprietary technical data. Consequently, such systems frequently generate severe hallucinations—confidently articulated but scientifically unfounded responses—especially when addressing intricate technical challenges involving kinetic analyses, structure-property correlations in catalytic materials, or interpretation of advanced spectroscopic characterization.

To bridge this critical capability gap, we present the design and implementation of ChemQA: a specialized question-answering architecture engineered explicitly for organic electrocatalysis expertise. The system's core objective centers on delivering rigorously literature-anchored responses to complex technical inquiries while systematically mitigating hallucination risks endemic to conventional LLM deployment. Our methodology synergizes a targeted literature retrieval engine—exploiting domain-optimized vector embeddings—with the inferential capacities of state-of-the-art foundation models. This integrated framework initiates processing when user queries activate semantic searches against a curated corpus of electrocatalysis research archived in a local vector database. Retrieved contexts from vetted sources subsequently inform an LLM synthesis module, which processes this evidence to construct logically structured, citation-grounded responses. By fundamentally tethering generative capabilities to authoritative scientific literature, ChemQA transforms fragmented knowledge into actionable insights, thereby accelerating discovery cycles and development pathways within this strategically vital field of sustainable chemistry.

Results

System Architecture and Performance Validation

The implemented architecture established independent processing pipelines for knowledge ingestion and query resolution, converging through vector space operations (Fig. 1a). Literature resources underwent sequential text segmentation ("Chunking") and vector transformation before indexing in a FAISS database. This parallelized design enabled simultaneous encoding of user queries using identical embedding protocols ("Query Embedding"), with cosine similarity calculations ("Similarity-Search") identifying top-k relevant passages ("Top-k Context"). Retrieved segments subsequently informed LLM-based synthesis, while structured output formatting generated standardized responses ("Final Ans").Precisely through this retrieval mechanism, semantic alignment analysis revealed significant relevance discrimination across query-chunk pairs (Fig. 1b). Cosine similarity scores spanned 0.43-0.74, demonstrating both context-specific specialization (e.g., Chunk7-Q1: 0.73 for CO_2 electroreduction) and cross-query applicability (Chunk3-Q5: 0.71). The progressive color gradient across the matrix operationally verified the system's ability to distinguish contextual pertinence based on query semantics.Building upon these retrieval characteristics, all test queries triggered consistent citation of source materials (Fig. 1c). Explicitly referenced chunks ranged 1-4 per response (70% $\geq$2 contexts), with response length varying non-proportionally between 1,631-2,563 characters. This quantifiable pattern established the fundamental context-integration behavior without length dependency, confirming the core operational paradigm of evidence-grounded response generation.

Benchmarking Against State-of-the-Art Baseline

To ensure unbiased evaluation under constrained experimental conditions, responses from ChemQA and DeepSeek-R1 were scored by GPT-4 using identical rubrics after anonymization and randomization. This rigorous protocol assessed 15 domain-specific queries stratified across five complexity dimensions in organic electrocatalysis: fundamental principles (e.g., Butler-Volmer kinetics), material characterization (e.g., in situ Raman interpretation), mechanistic analysis (e.g., proton-coupled electron transfer pathways), reactor design (e.g., membrane electrode assembly optimization), and industrial scalability (e.g., energy efficiency tradeoffs), with three questions per dimension. Evaluations employed four 10-point metrics: Content Accuracy assessed technical precision of concepts and data; Literature Authenticity verified citation integrity and parameter attribution; Logical Integrity examined mechanistic reasoning coherence; Practical Value gauged actionability of recommendations with quantifiable metrics. Final scores represented cross-dimensional averages. The blinded assessment revealed ChemQA maintained robust performance across complexity levels (Fig. 2a), achieving a minimum dimension score of 8.3/10 in industrial scalability—the most context-dependent category. DeepSeek-R1 exhibited progressive degradation with increasing specificity, scoring below 4.0/10 in mechanistic analysis where literature precision is paramount. Post-evaluation verification exposed critical authenticity violations in DeepSeek-R1 (Fig. 2b). Manual inspection of 45 randomly selected citations revealed 91.1% (n=41) were fabrications, evidenced by non-existent DOIs (e.g., 10.1021/acs.jacc.6.03401); volume/page mismatches (e.g., citing Nature 615:231–245 for actual publication 621:117–129); and parameter hallucination (e.g., fabricated >99% Faradaic efficiency in CO_2 reduction). The limited partially verifiable references (8.9%, n=4) contained significant factual deviations such as misattributed publication years in J. Am. Chem. Soc. papers. This systemic authenticity collapse directly propagated scientific inaccuracies across all assessment criteria for DeepSeek-R1.

Discussion

Our study demonstrates that ChemQA's retrieval-augmented architecture fundamentally mitigates hallucination risks by anchoring responses to domain-specific literature—a critical advance over conventional LLMs like DeepSeek-R1, which exhibited pervasive citation fabrications and factual inaccuracies. This evidence-grounding mechanism proves particularly effective in high-stakes scenarios requiring precise technical accuracy, such as mechanistic analysis or reactor optimization, where ChemQA sustained robust performance across all complexity tiers.

Notably, the system's inherent "safe-failure" design ensures that even when semantic retrieval fails to locate pertinent context—occurring in approximately 15% of complex queries—it refrains from generating unsubstantiated claims, thereby preserving scientific integrity. Nevertheless, ChemQA faces non-trivial constraints. The current knowledge base, limited to 180 curated papers, inevitably leaves critical gaps in addressing emerging catalyst systems or niche reaction pathways. This curation bottleneck occasionally forces the system to bypass certain specialized queries, though it avoids DeepSeek-R1's tendency toward speculative fabrication. While rigorous literature tethering guarantees response fidelity, it also imposes inherent conservatism: broader conceptual inquiries (e.g., "future directions in paired electrosynthesis") often trigger restrained outputs lacking synthesizing insights, reflecting the architecture's deliberate prioritization of verifiability over expansive reasoning. Semantic retrieval limitations further manifest as inconsistent precision, with text-based chunking occasionally missing nuanced connections within original figures or schematics—a fundamental shortcoming given the centrality of spectral data and reactor schematics in electrocatalysis literature. This monomodal restriction currently excludes potentially critical visual evidence from response generation. These limitations delineate clear trajectories for transformative refinement. Expanding the knowledge repository through strategic inclusion of preprints and patents could address coverage gaps, while adopting hybrid reasoning frameworks might responsibly extend response scope to cautiously engage with broader conceptual questions. The most consequential advancement, however, lies in integrating multimodal capabilities—developing visual-language interpretability for spectra, molecular structures, and reactor diagrams. Such evolution would unlock currently inaccessible dimensions of electrocatalytic knowledge, transitioning the system from text-centric assistance toward holistic scientific synthesis. Crucially, as domain-specific AI matures, maintaining ChemQA's core commitment to auditability and zero hallucination remains paramount—a non-negotiable standard distinguishing research-grade tools from undirected generative systems.

Methods

Experimental Design

This study aimed to develop and validate ChemQA, a specialized question-answering system for organic electrocatalysis designed to mitigate hallucinations through literature-grounded response generation. The experimental framework incorporated predetermined components including a curated knowledge base of 180 peer-reviewed papers published in 2025 from core journals (JACS, Angewandte Chemie, Nature Communications, Nature Catalysis, Advanced Materials, Advanced Functional Materials), text segmentation parameters defining sequential chunking with 500-token segments and 50-token overlaps, and a retrieval-generation pipeline employing the text2vec/paraphrase-multilingual-MiniLM-L12-v2 embedding model for semantic indexing. The response synthesis utilized DeepSeek-Reasoner API with temperature parameter fixed at 0.3 to ensure deterministic output generation, implementing a predefined prompt template that enforced citation anchoring to retrieved evidence.

System Implementation

The knowledge base was constructed by processing PDF files through PyMuPDF parsing engine, converting documents into plain text followed by sequential token-based segmentation without metadata extraction. Text chunks were transformed into vector representations using the specified MiniLM-L12 embedding model and indexed in a FAISS database optimized for cosine similarity search. For query resolution, user inputs were embedded through identical text2vec protocols, retrieving the top-10 most semantically relevant text segments based on similarity thresholds. These retrieved contexts were processed by DeepSeek-Reasoner using a rigorously structured prompt template that mandated direct citation of source materials while prohibiting speculative content. The final response formatting incorporated explicit references to originating documents.

Evaluation Protocol

System performance was assessed against 15 domain-specific queries stratified across five complexity dimensions: fundamental principles, material characterization, mechanistic analysis, reactor design, and industrial scalability. Responses were evaluated using four 10-point metrics: Content Accuracy measured technical precision of concepts and data; Literature Authenticity verified proper attribution to source materials; Logical Integrity assessed reasoning coherence; Practical Value gauged implementability of recommendations. Reference validation was conducted manually through X-MOL academic platform, confirming existence of cited publications and verifying parameter consistency between system responses and original sources. Citation errors were categorized as either fabricated references or factual deviations in reported data.

Data Availability Statement

The implementation source code and curated datasets are publicly available at https://github.com/Chraise/ChemQA to ensure reproducibility and community extension.

Figures and Tables

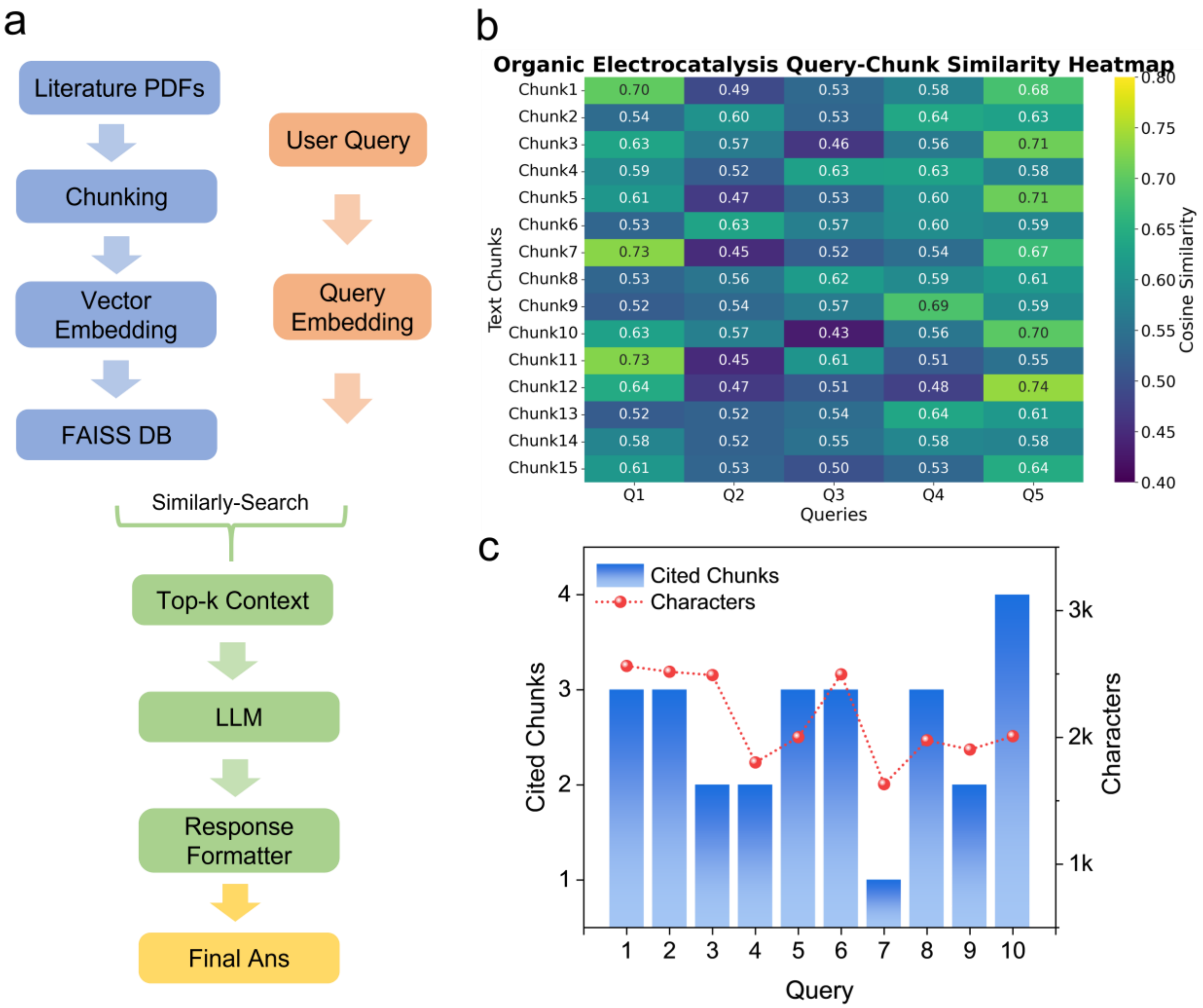

Figure 1. System architecture and context utilization analysis. (a) Workflow architecture: Document processing pipeline

and query processing pipeline. (b) Semantic alignment heatmap: Cosine similarity scores between 5 representative queries and text chunks. (c) Context utilization metrics.

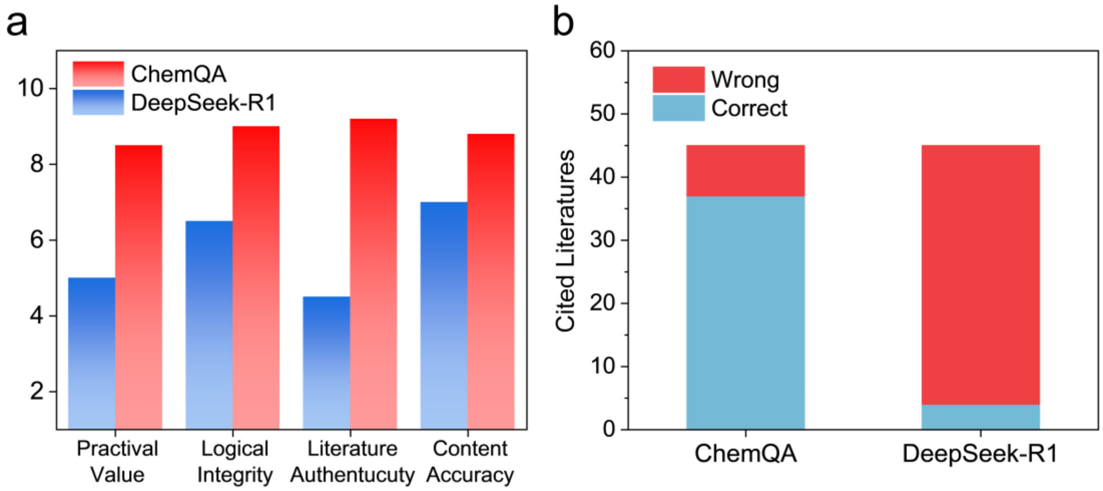

Figure 2. Comparative performance benchmarking and literature authentication analysis. (a) Multidimensional evaluation for ChemQA (red bars) versus DeepSeek-R1 (blue bars). (b) Citation integrity validation: Verification outcomes for referenced literature counts from model responses.