SC-HyDE: Mitigating Hallucinations in Chinese Legal Question Answering via Self-Corrected Hypothetical Document Embeddings

Haoran Li

School of mathematics and statistics, Northeastern University at Qinhuangdao, Qinhuangdao, China

lanshengliri@gmail.com

Abstract. Retrieval-Augmented Generation (RAG) has been cited as a potential technique for reducing hallucinations in Large Language Models (LLM). However, applying RAG in the Chinese legal landscape remains a challenge due to the large semantic gap between informal user queries and professional legal terminology, and the risk of retrieval noise and false information generation. To address these concerns, this paper proposes SC-HyDE, a novel, training-free system that blends generative expansion with self-reflective filtering. First, the domain based Hypothetical Document Embeddings (HyDE) system is applied to generate an imaginative legal analysis, effectively filling out the semantic gap and restoring retrieval accuracy. Second, since HyDE can produce hallucinating legal reference points within its latent space, there is a simple module incorporated for Self-Correction Critic. This module compares the relevance of retrieved documents to the original query by using a zero-shot prompting strategy while excluding noise before the second generation stage. The proposed framework is evaluated against a sample of 588 cases from official legal datasets. The experimental results show that SC-HyDE performs well above standard RAG and vanilla HyDE levels. In particular, the method drops the Noise Rate from 88.33% to 25.06% and the Hallucination Rate from 25.68% to 13.61%, offering an effective approach for those high stakes legal consultancies where factual accuracy is a major concern.

Keywords: Retrieval-Augmented Generation, Legal Question Answering, Hallucination Mitigation, Hypothetical Document Embeddings, Self-Correction.

1. Introduction

Large Language Models (LLM) have proved to be extremely effective in understanding and producing natural language in a wide variety of domains. However, their use in knowledge-intensive and demanding work environments such as legal consultation is significantly impeded by the phenomenon of hallucination. Hallucination occurs when a model generates plausible but factually incorrect information, which in the legal context can result in wrong advice or wrong statutory references. To avoid such risk, ground LLM responses were assessed using Retrieval-Augmented Generation (RAG) within external, non-parametric knowledge bases [1]. RAG models theoretically avoid the use of model internal weights by retrieving the relevant data prior to their generation, thereby maintaining more consistency in fact.

Although RAG has gotten state-of-the-art results in its general open-domain tasks, its direct application to the Chinese legal domain is unique and gravely problematic. The first is the semantic gap. Law firms use colloquialism to describe their legal concerns, whereas courts use rigid professional terms. As Ma et al. point out in the LeCaRD benchmark [2], conventional dense retrieval models lack the scaling to these heterogeneous texts, which leads to poor recall. This mismatch is particularly pronounced in a civil and criminal law system legally applicable where legal applicability depends on precisely defined text. To address this, a previous paper proposed Hypothetical Document Embeddings where an LLM is used to generate a hypothetical answer to guide retrieval [3]. While HyDE increases recall, this research suggests that there is a corresponding side effect in the legal realm where the generative model tends to have incorrect article numbers or sentences within the hypothetical document itself.

The second problem is locating noise and the hallucinations resulting from this noise. Retrieving irrelevant statutes is not simply inadvertent but actively harmful in legal situations. Previous studies of legal LLMs showed that models are easily misled by irrelevant retrieval results, in other words, citation hallucinations in which the model confidently refers to non-existing or inapplicable law because it existed in the collection context. Even though recent developments, such as Self-RAG and Corrective RAG (CRAG) have introduced critique and corrective mechanisms for handling noise [4,5], these require computationally expensive training of specialized models of critics.

This paper proposes a non-training, dual-stage RAG framework, SC-HyDE, to bridge this semantic gap and strictly control the hallucination risk with self-Corrected HyDE for the Chinese legal domain. This approach is based upon a fundamental insight that, in order to be effective at legal reasoning one needs to balance two conflicting ends of semantic flexibility and factual rigidity. The models must also be flexible enough to anticipate the colloquial elements of a layman's question (by enriching colloquial queries), but they must also be faithful enough to adhere strictly to law. Other retrieval methods are unable to meet these conflicting goals simultaneously.

So, SC-HyDE decouples semantic interpretation from factual verification and mimics the cognitive processes of a legal professional brainstorming potential legal grounds and following up strictly with checking for errors. In particular, this framework first utilizes a domain-specific HyDE mechanism to construct hypothesized legal studies. This step deals more with recall, casting a large net through the reasoning capabilities of the model. Yet there is a critical paradox, in which this mechanism that increases the retrieval by generative expansion also generates toxic noise. If an article is numbered in hallucination in a hypothetical document, this represents a detrimental error, creating a distorted query yields totally irrelevant statutes. This is countered by the inclusion of a small Self-Correction Critic module. This module operates on a zero shot prompting strategy to test for relevance of retrieved documents with the original question, acting as a filtering system to eliminate hallucination noise prior to the final generation phase. This framework, by combining broad recall expansion with strict filter verification, ensures high retrieval accuracy without sacrificing factual consistency. The remainder of this paper is organized as follows.

Section 2 examines other work on legal RAG as well as the inherent limitations of current legal AI systems. The SC-HyDE methodology is presented in Section 3, with the knowledge base construction using statutory file integrated and the self-correction logic. Section 4 validates the proposed framework through large-scale experiments in 588 cases. Section 5 discusses the limitations and an envisioned solution. Section 6 concludes the study, describing the research achievements and summarizes results.

2. Related Work

2.1. Retrieval-Augmented Generation in Specialized Domains

RAG has become a dominant model for legal and medical AI. RAG evolved from a relatively dense retrieval process to a more iterative, self-reflective architecture. A standard RAG system uses encoders such as BERT or RoBERTa to produce vector representations of documents. But these systems often fail when the query is not specified or is semantically disconnected from the target document. To combat this, HyDE was developed to exploit the generative power of LLMs to create a bridge between the query and document space [3]. Although HyDE has been successful in many things, the application of HyDE in law is still not well understood, particularly with regard to the noise produced by the hypothetical generation.

2.2. Limitations of Current Legal AI and Reasoning

Despite the proliferation of general purpose LLMs, their effectiveness in court remains limited. Benchmarks like LawBench and LeCaRD have provided evidence that even the most advanced models often fail in the face of complex statutory reasoning [2,6]. The primary reasons are threefold. First, the long-tail nature of the distribution of statutes: models fit well to common crimes such as

theft, but ignore nuances of rare offenses. Second, the strict logical legal application: The judge requires absolute deductive consistency, while LLMs are merely probabilistic machines. Third, the dynamism of the courts' interpretations: general models on static datasets can often lack accurate legal knowledge. These have been addressed by methods of fine-tuning such as in DISC-LawLLM, but a high probability of hallucinating particular article numbers has not been achieved for actual deployment, with the likelihood of an article number being hallucinated in certain articles [7].

2.3. Hallucination Mitigation and Self-Correction

LLM research has been focused on reducing hallucinations. Such methods range from supervised fine-tuning to reinforcement learning. In RAG, this is the logical direction to take. Reflection tokens were introduced by Self-RAG to allow models to critique their own retrieval and generation [4]. But, since specialized training is necessary to make it generalizable it is limited. As discussed in the present study, prompt-based self-correction is a more flexible alternative, since advanced LLMs, like GLM-4.7 or GPT-4, can use the reasoning capabilities to serve as a zero-shot filter.

3. Methodology

The SC-HyDE model organizes the legal question answering process as a multi-step mapping and verification process. The overall system is designed to translate squabbled descriptions of a case into professional judgements by the courts.

3.1. Integrated Legal Knowledge Base Construction

Rather than common RAG systems that store multiple small files, the proposed solution utilizes an integrated Statutory Corpus. The entire Chinese Criminal Law is written as a single integrated document so that the laws remain consistent within and organized in a hierarchy. Document is grouped into context aware chunks during preprocessing and uses a hierarchical chunking strategy. Each chunk represents a statute or logical set of provisions, and the identifiers for the articles (e.g. Article 12) are injected into the content so that the embedded model retains the unique identity of each statute. This processed together prevents the semantic splintering that occurs in multi-file systems.

3.2. Generative Expansion via Domain-Adapted HyDE

To bridge the semantic gap, a generation function G(·) is defined. The hypothetical legal document d_{hypo} is generated as follows:

$$d_{hypo} = G(q, p_{hyde}; \theta_{gen}) \tag{1}$$

Where q is the original user query, p_{hyde} is the domain specific prompt to the model to do as a legal analyst, and θ_{gen} is the parameter of the generator model GLM-4.7 for producing the hypothetical text d_{hypo}.

3.3. Dense Vector Retrieval and Embedding

The framework maps the generated analysis into a high-dimensional vector space. The retrieval process seeks a subset D_{raw} of size k that maximizes the cosine similarity:

$$D_{raw} = \arg\max_{d \in \mathcal{K}}^{k} \left(\frac{E(d_{hypo}) \cdot E(d)}{|E(d_{hypo})||E(d)|} \right) \tag{2}$$

In this equation, $E(\cdot)$ is the embedding function (BGE-M3) that converts texts into vectors, $\mathcal{K}$ is the combined knowledge base with all statutory chunks, and k represents the number of candidates retrieved. The fraction represents the cosine similarity score of the hypothetical document vector and the candidate document vector.

3.4. Self-Correction Filtering Logic

The raw retrieval set D_{raw} frequently contains statutes that are semantically similar but legally inapplicable. A Critic function $C(\cdot)$ is implemented to filter these results. The filtered context D_{sc} is constructed as:

$$D_{sc} = \{d_i \in D_{raw} \mid C(q, d_i; \theta_{critic}) = 1\} \tag{3}$$

Here, d_i represents the i-th document in the raw retrieval set, and θ_{critic} refers to the parameters of the critic model. The function C outputs 1 if the document is deemed relevant and factual relative to the query q, and 0 otherwise.

To handle cases where the critic might be overly aggressive and filter out all documents, a fallback mechanism is implemented:

$$D_{final} = \begin{cases} D_{sc} & \text{if } D_{(sc)} \neq \emptyset \\ \{d_{(1)}\} & \text{if } D_{(sc)} = \emptyset \end{cases} \tag{4}$$

Where D_{sc} is the filtered set from Equation (3), $\emptyset$ denotes the empty set, and $d_{(1)}$ represents the single statute with the highest similarity score in D_{raw}, ensuring the final generation is always grounded in at least one document.

3.5. Prompt Engineering and Logic-Based Filtering

This model is instructed in the judge prompt in the critic module to ignore the article number from Stage 1 and focus only on consistency in content between the case facts and the retrieved statutes. It is essential that the model performs a factor analysis, which, for example, equates the subjective intent and objective act before deciding to keep or discard the document. This separation of concerns helps to avoid hallucinations from the stage of generative retrieval to the stage of final decision.

4. Experiments and Evaluation

4.1. Experimental Setup

The framework was tested in the GLM-4.7 model. The Statutory Corpus is a combination of the integrated Chinese Criminal Law. 588 cases were utilized for a statistically significant assessment.

For the baseline comparison, a Zero-shot Self-RAG baseline is implemented, where prompt-based critique and naive retrieval are used as mirrors to the filtering logic of the Self-RAG model [4].

4.2. Evaluation Metrics and Results

To quantitatively assess the performance, two specific metrics are defined.

The Noise Rate (NR) measures the proportion of queries where the retrieved context contains at least one irrelevant statute. It is calculated as

$$NR = \frac{\sum_{i=1}^{N} II\left(noise \in D_{retrieved}^{(i)}\right)}{N} \times 100\% \tag{5}$$

Where N is the total number of test cases (588), $D_{retrieved}^{(i)}$ is the set of retrieved documents for the i-th query, and $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the set contains irrelevant information (noise) and 0 otherwise.

The Hallucination Rate (HR) evaluates the factual accuracy of the final generation. It is defined as:

$$HR = \frac{\sum_{i=1}^{N} \mathrm{II}\left(ans^{(i)} \notin S_{ground\_truth}^{(i)}\right)}{N} \times 100\% \tag{6}$$

Where $ans^{(i)}$ represents the citations generated by the model for the i-th query, and $S^{(i)}_{ground\_truth}$ represents the set of correct statutory articles identified by legal experts. If the model fails to cite the correct statute (citation hallucination), the indicator function returns 1.

Table 1. Experimental Results with Standard Deviation (N=588)

Method	Noise Rate ↓	Hallucination Rate ↓	Filter Error ↓
HyDE Only	88.33% ± 0.059	25.68% ± 0.437	0.0000
Self-RAG (Zero-shot)	33.90% ± 0.430	54.76% ± 0.498	N/A
SC-HyDE (Ours)	25.06% ± 0.353	13.61% ± 0.343	0.2398

Table 1 summarizes the results based on these metrics. The experimental results suggest there is a significant inter-recall deficit between retrieval recall and accuracy across approaches. The Noise Rate on the HyDE Only baseline is alarmingly high at 88.33%. This confirms the hypothesis that generative expansion provides a semantic bridge, but it is a double-edged sword; it introduces not-existent or unhelpful legal concepts and retrieves unrelated statutes. While the generation model can tolerate some noise, the 25.68% Hallucination Rate suggests that the model is misled by this irrelevant context.

Interesting is that Zero-shot Self-RAG reduces noise compared to HyDE but dramatically increases the Hallucination Rate (54.76%). This is odd, and suggests that within the absence of domain-specific calibration, generic critique prompts may be overly aggressive, focusing on incorrect but semantically nuanced statutes (false negatives) or that the model is unable to keep instruction followed by legal reasoning in a zero shot setting.

SC-HyDE matches that optimum balance. It reduces noise by over 63 percentage points compared to baseline vanilla HyDE by dissolving the generative retrieval from strict logic-based criticism. Crucially, the very low Filter Error (0.2398) suggests that the domain-adapted Critic module will maintain ground truth properly while removing noise. These broad-recall, Strict-Filter constraints directly translate into a 47% relative improvement in factual accuracy (13.61% HR) over the strongest baseline, confirming the necessity of the Broad-Recall, Strict-Filter paradigm in answer to high-stakes legal questions.

4.3. Mechanism Analysis and Generalizability

To show the quantitative progress, a qualitative analysis of the success cases was conducted in the following table 2. The superior performance of SC-HyDE can be attributed to the complementary characteristics in two stages, which effectively separate semantic recall from logical verification.

Table 2. Qualitative Analysis of the SC-HyDE Reasoning Process

Stage	Core Content / Reasoning Logic	Role
User Query	Defendant Gan privately carved corporate seals without authorization to sign construction agreements and later surrendered.	Layman Input
HyDE Draft	Identified "Forgery of Corporate Seals" (**Article 280**). Focuses on unauthorized carving rather than financial fraud.	Semantic Bridge
Raw Retrieval	[ID:0] Art. 272 (Embezzlement); **[ID:1] Art. 280 (Forgery)**; [ID:2] Art. 163 (Bribery).	Broad Recall
Critic Logic	**Discard ID:0**: No personal use of funds found. **Retain ID:1**: Directly matches the act of forging seals. **Discard ID:2**: No "intent of illegal possession" for fraud.	Strict Filter
Final Answer	Guilty of **Forgery of Corporate Seals**. (Matches Ground Truth: 280)	Precise Output

To further define the quantitative gains, a qualitative analysis of success cases was conducted. In fact, the two stage complementary nature of SC-HyDE confers superior performance because they effectively divorce semantic recall from logic verification. In stage one, even when the HyDE Draft successfully identifies the correct statute (e.g., Article 280 in Table 2), the Raw Retrieval stage often

introduces "semantic noise." It frequently returns topically related but legally irrelevant articles, such as “Article 272” (Embezzlement) or “Article 163” (Bribery), because they share similar keywords like "corporate" or "illegal gain." This confirms the hypothesis that the hypothetical document is a strong semantic anchor. Even if the statute in question is factually incorrect, the descriptive text satisfies the gap between the layman’s query and the professional database so that the correct statute is found in Raw Retrieval. One might hardly expect a dense retriever without this generative expansion to offer a clear picture of these detail-based connections [8].

The Self-Correction Critic is the key to developing this noisy set in the second stage. The result of this analysis is that the Critic creates a discriminatory boundary between only topically related and legally applicable statutes. For example, in embezzlement vs forgery cases, the raw retrieval frequently returns both statutes because they are semantically similar. The generative critic, who is instructed to factor-based argument, succeeds in seeing that the characteristic of unauthorized carving of seals is the one that distinguishes the statute of forgery. This suggests that LLMs do possess a specialized ability to engage in zero-shot legal reasoning that is easily engaged through prodding without expensive parameter updates [9].

Plus, SC-HyDE is a training-free compound that illustrates high generalizability. In contrast to supervised methods that overfit specific legal datasets such as criminal law, the framework is based on the general reasoning power of the LLM and semantic matching of the retriever. This means that SC-HyDE could be immediately moved onto other dynamic legal domains, like Civil Code or Administrative Law, or even to other vertical areas like medical guidelines, simply by merely changing the underlying knowledge base and adapting the prompt instructions. This flexibility is especially useful in real-world environments where statutes are often amended and retraining costs prohibitive.

5. Discussion

Despite the success of SC-HyDE, there are several issues and technical bottlenecks requiring further scrutiny. The first one is the computational overhead and cost. This multi-stage prompting design necessitates three calls to the LLM, one for each of the hypothetical analysis, document filtering, and the final verdict generation. This inevitably leads to increased consumption of tokens and slower inference latency. In real-life judicial scenarios where people need to obtain rapid responses with minimal cost, this squandering of tokens due to the chain-of-thought process might hinder the feasibility of deploying the designed prompt at scale.

Second, while a low hallucination rate (HR) this time measured 13.61%, the strict hallucination mitigation concerns for an absolute-precision domain such as law may still hold. This hallucination rate means that 1 in 7 judicial consultations may still contain erroneous citations of law. Most of the remaining hallucinations are due to the semantic ambiguity of law, wherein the model errs in classifying a particular crime as under the jurisdiction of a similar, but distinct article. Further reducing this rate without fine-tuning is challenging.

Plus, the current study is limited by the scale of experimental and document complexity. While 588 cases are analyzed, it is relatively small in comparison to the millions of cases in national legal databases. The self-correction critic’s performance with super-big data sets or extremely long documents with hundreds of complex cross-references is yet to be fully proven. As the length of the document increases, the lost-in-the-middle phenomenon inherent in LLMs is likely to amp up the noise problem and could degrade the filtering logic.

Future directions include research into model distillation methods to address such challenges. The ability of the judge-role to reason would be reduced to a smaller, locally applied model (e.g. 7B or 14B parameters), significantly reducing the token costs and inference time. Plus, the use of structured graphs of legal knowledge could extend the vector-based RAG system and become a second layer of symbolic logic validation that can be used to further decompose the rest of the hallucination space [10].

6. Conclusion

In this paper, the SC-HyDE framework was proposed to address the double problem of the semantic gap and retrieval noise in the Chinese legal domain. The paper has integrated a domain-adapted Hypothetical Document Embeddings mechanism with a self-reflective filtering module to bridge the gap between colloquial descriptions and professional statutes without the need for additional model training. This study developed a coherent pipeline of statutory processing to protect the structural integrity of the Chinese Criminal Law in a single cohesive knowledge base.

The proposed architecture was successfully tested in 588 cases through experimental analyzes. The results indicate that SC-HyDE can successfully prune retrieval noise, reducing the Noise Rate from 88.33% in vanilla HyDE to 25.06%. The framework also exhibited a final Hallucination Rate of 13.61%, which was significantly higher than the naive RAG and HyDE-only baselines. These findings show that an "Out-of-the-Box" Broad-Recall, Strict-Filter approach is an indispensable component to maintaining factual grounding in high-stakes professional consultations.

This is despite the results, but this study identifies ongoing challenges related to computational efficiency and the cost of multiple-stage API consumption. The latency and token costs of triple-prompting are not substantial obstacles. Also, the scaling of this training-free approach to ultra-long documents and multi-jurisdictional databases requires more extensive testing.

In the short run, the overall outlook for SC-HyDE is that for a reduction in cost and response speed, LLM calls should move away from general purpose LLM calls and to locally generated distilled local models. The end goal is to refine the logic of self-correction using expertly defined graphs of legal reasoning to make automated legal assistants able to provide almost no hallucinations to the modern justice system.

References

[1] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020, pp. 9459–9474J.

[2] C. Xiao et al., "CAIL2018: A large-scale dataset for legal judgment prediction," in Proc. 27th Int. Joint Conf. Artificial Intelligence (IJCAI-18), 2018, pp. 4487–4493

[3] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels," in Proc. 61st Annu. Meeting Assoc. Computational Linguistics (Vol. 1: Long Papers), 2023, pp. 1762–1777

[4] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in Proc. 12th Int. Conf. Learning Representations (ICLR 2024), 2024.

[5] S. Yan, J.-C. Gu, Y.-Z. Jiang, and Z.-H. Ling, "Corrective retrieval augmented generation," in Proc. 2024 Conf. North American Chapter Assoc. Computational Linguistics (NAACL 2024),

[6] Z. Fei et al., "LawBench: Benchmarking legal knowledge of large language models," in Proc. 2024 Conf. Empirical Methods in Natural Language Processing (EMNLP 2024), 2024, pp. 7933–7962.

[7] S. Yue et al., "DISC-LawLLM: Fine-tuning large language models for intelligent legal services," arXiv:2309.11325. 2023.

[8] A. B. Hou, W. Jurayj, N. Holzenberger, A. Blair-Stanek, and B. Van Durme, "Gaps or hallucinations? Gazing into machine-generated legal analysis for fine-grained text evaluations," arXiv:2409.09947. 2024.

[9] V. Noël, E. Y. Seidou, C. K. Capo-Chichi, and G. Amari, "HalluGraph: Auditable hallucination detection for legal RAG systems via knowledge graph alignment," arXiv:2512.01659. 2025.

[10] P. Trivedi et al., "Self-rationalization improves LLM as a fine-grained judge," arXiv:2410.05495. 2024.