RESEARCH ARTICLE

Journal of Terminology and Knowledge Engineering
1: 20260018, 2026
https://doi.org/10.3724/jtke-20260018

Artificial Intelligence

Zero-shot terminology extraction via multi-perspective prompting and coarse-to-fine verification

Lu Xiang*, Yaping Zhang & Yu Zhou

State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

*Corresponding author (email: lu.xiang@nlpr.ia.ac.cn)

Received 12 February 2026; Revised 9 May 2026; Accepted 18 May 2026; Published online 15 June 2026

Abstract: Automatic term extraction (ATE) constitutes a core component in a range of downstream applications, including knowledge graph construction, machine translation, and information retrieval. Although large language models (LLMs) have demonstrated substantial performance gains across diverse natural language processing tasks, their effectiveness in zero-shot ATE settings remains constrained by two principal issues: (1) reliance on single-perspective prompting often leads to insufficient recall, as the generated candidates fail to adequately capture the full semantic space of domain-specific terms; (2) LLM outputs tend to exhibit elevated false-positive rates, primarily due to hallucinated content and the inclusion of semantically irrelevant or overly general lexical items. To overcome these limitations, the present study proposes MP-CoVe, a training-free framework that decomposes the ATE process into two distinct stages: a multi-perspective candidate generation phase and a coarse-to-fine verification phase. During the generation stage, a multi-perspective prompting strategy is employed to elicit term candidates from the LLM under diverse semantic viewpoints, thereby enhancing coverage and mitigating the biases associated with single-query formulations. In the subsequent verification stage, a stratified filtering mechanism is introduced to reduce the impact of model stochasticity. Rather than uniformly subjecting all candidates to semantic validation, the framework leverages corpus-level statistical signals to differentiate high-confidence terms from evident noise, reserving LLM-based verification exclusively for ambiguous instances. Empirical evaluations conducted on the English ACTER dataset and the Chinese medical entity extraction (CMeEE) dataset demonstrate that MP-CoVe consistently outperforms competitive baselines. Notably, it achieves a more favorable trade-off between precision and recall, without requiring domain-specific fine-tuning.

Keywords: automatic terminology extraction, LLMs, multi-perspective prompting, coarse-to-fine verification

INTRODUCTION

Automatic term extraction (ATE) seeks to identify domain-specific terminology from specialized corpora and serves as a foundational component in a variety of downstream tasks, including ontology learning, machine translation, and information retrieval [1]. Conventional approaches to ATE have largely relied on statistical heuristics, such as term-frequency-inverse-document-frequency (TF-IDF) and C-Value [2], or on supervised sequence labeling frameworks that necessitate substantial volumes of annotated data [3]. Although these supervised paradigms can achieve strong performance, their reliance on costly manual annotation renders them less adaptable in low-resource settings and limits their generalizability across domains.

© The Author(s) 2026. Published by Science Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The advent of large language models (LLMs) has introduced new methodological possibilities for ATE through in-context learning. Recent work has explored their application in ontology learning, demonstrating an ability to capture domain-specific linguistic regularities [4]. Nevertheless, the direct deployment of generative models in this context remains constrained by intrinsic stochasticity and a propensity for hallucination. To alleviate these issues, several inference-time strategies have been proposed. Notably, self-consistency [5] aggregates outputs from multiple stochastic generations to reduce variance, while self-verification [6] prompts the model to evaluate and refine its own predictions. Despite their effectiveness in general settings, such strategies do not adequately address challenges that are specific to terminology extraction.

(1) **Prompt sensitivity and complementarity:** LLMs exhibit pronounced sensitivity to prompt formulation, which directly affects both recall and the diversity of extracted terms. Empirical observations on the heart failure (HTFL) subset of the ACTER dataset [7] indicate that F1 scores may vary by more than ten points across different zero-shot prompting strategies. More critically, distinct prompts tend to retrieve complementary subsets of terminology. For example, in the context of cardiac therapy, a definition-oriented prompt may identify abstract concepts such as heart failure, whereas a syntax-oriented prompt is more likely to extract complex nominal expressions such as cardiac resynchronization therapy. Consequently, relying on a single prompt or repeatedly sampling under the same prompt inevitably leads to incomplete coverage of the terminologies and systematic recall deficiencies.

(2) **Stochasticity in uniform verification:** Existing generate-then-verify paradigms typically apply uniform verification across all candidate terms, without accounting for differences in confidence levels. This approach is suboptimal in the context of ATE, where valid terms are often characterized by consistent recurrence and stable usage patterns within the corpus [2,8]. Such corpus-level regularities frequently provide more reliable evidence than a single-pass model judgment. Subjecting high-confidence candidates to redundant LLM-based verification not only incurs unnecessary computational overhead but also exposes them to stochastic variation, potentially resulting in the erroneous rejection of valid terms. An effective framework should therefore differentiate between candidates supported by strong statistical signals and those that genuinely require semantic disambiguation.

To address these limitations, we introduce MP-CoVe, a training-free framework that integrates multi-perspective generation with coarse-to-fine verification. In contrast to prior approaches that depend on repeated stochastic sampling and indiscriminate validation, MP-CoVe combines diverse prompting strategies with a stratified selection mechanism. Specifically, the framework decomposes the extraction process into two stages. The first stage employs multi-perspective prompting to elicit a diverse and comprehensive set of candidate terms from complementary semantic viewpoints. The second stage implements a coarse-to-fine verification procedure: corpus-level statistical indicators, such as extraction frequency, extraction rate, and inter-prompt consensus, are first used to distinguish high-confidence candidates from obvious noise. Subsequently, only those candidates deemed ambiguous are subjected to LLM-based semantic verification. By grounding the verification process in empirical statistical evidence, this design mitigates the instability associated with excessive reliance on generative models. The main contributions of this work are threefold.

(1) It identifies prompt sensitivity and the stochasticity induced by uniform verification as key bottlenecks in zero-shot ATE, thereby clarifying the limitations of existing prompting and validation strategies.

(2) It proposes MP-CoVe, a novel framework that jointly leverages multi-perspective prompting for

enhanced recall and a stratified verification mechanism for improved precision.

(3) Extensive experiments conducted on the English ACTER and CMeEE (Chinese medical entity extraction) datasets demonstrate that MP-CoVe consistently outperforms strong baselines, achieving a more balanced trade-off between precision and recall without requiring domain-specific fine-tuning.

RELATED WORK

This section reviews prior research from three complementary perspectives. We first present an overview of ATE methods. We then examine self-consistency techniques that enhance the reliability of LLMs through multiple sampling. Finally, we discuss self-verification approaches that exploit the asymmetry between generation and evaluation. The proposed framework integrates insights from these strands by combining multi-sample generation with selective verification to achieve robust term extraction.

Automatic term extraction

ATE focuses on identifying domain-specific terminology from specialized corpora. Following established taxonomies in recent surveys [4,9], existing approaches can be broadly classified into four categories.

Traditional methods. Early ATE research predominantly relied on linguistic and statistical heuristics, without incorporating machine learning. Linguistic approaches utilized part-of-speech patterns and syntactic structures to identify candidate terms [8], whereas statistical methods employed measures such as TF-IDF, C-value, and domain specificity to rank candidate expressions [2]. Hybrid approaches combined these paradigms by applying linguistic filters for candidate identification and statistical metrics for subsequent ranking [10]. Although these methods offer interpretability, their reliance on manually designed rules limits their adaptability and hampers cross-domain generalization.

Machine learning-based methods. Subsequent work formulated ATE as a supervised learning problem, typically framed as either classification or sequence labeling. Models based on conditional random fields (CRF), support vector machines (SVM), and random forests leveraged handcrafted features to distinguish terms from non-terms [11,12]. While these approaches improved performance relative to rule-based systems, they remained dependent on extensive feature engineering and domain-specific annotated data.

Neural network-based methods. The adoption of deep learning marked a substantial shift, enabling models to learn feature representations automatically. Early neural approaches employed recurrent architectures, such as bidirectional long short-term memory-conditional random field (BiLSTM-CRF), for sequence labeling tasks [13]. More recently, pre-trained language models (PLMs), including bidirectional encoder representations from Transformers (BERT) and XLM-RoBERTa, have achieved state-of-the-art performance through task-specific fine-tuning on annotated datasets [14,15]. For instance, Lang *et al*. [15] explored transformer-based formulations encompassing token classification, sequence classification, and neural machine translation paradigms. The TermEval 2020 shared task [1] provided a comprehensive evaluation on the ACTER benchmark, demonstrating the superiority of transformer-based models over traditional techniques. However, these approaches typically require substantial labeled data and task-specific adaptation.

LLM-based methods. More recent advances have leveraged LLMs within an in-context learning

framework, thereby reducing reliance on supervised training. Yuan *et al*. [16] proposed knowledge-guided prompting strategies that incorporate topic information derived from knowledge graphs. Other studies have explored the application of chat generative pre-trained Transformer (ChatGPT) and GPT-4 to ATE, demonstrating promising performance without additional fine-tuning [9]. Nonetheless, as highlighted in recent surveys, LLM-based methods remain susceptible to output variability and prompt sensitivity. The present work addresses these limitations by integrating self-consistency with a selective verification mechanism, thereby improving both reliability and precision in zero-shot settings.

Self-consistency in LLMs

Self-consistency was introduced by Wang *et al*. [5] as a decoding strategy aimed at enhancing the robustness of chain-of-thought (CoT) reasoning. The central premise is that complex reasoning tasks often permit multiple valid solution paths that converge on the same correct outcome. By sampling a diverse set of reasoning trajectories and selecting the most frequently occurring answer via majority voting, self-consistency has been shown to substantially improve performance across arithmetic, commonsense, and symbolic reasoning benchmarks. Subsequent research has extended this paradigm in several directions. For instance, Aggarwal *et al*. [17] proposed adaptive consistency, which dynamically adjusts the number of samples required based on task complexity. Raj *et al*. [18] introduced semantic consistency as a means of assessing model reliability through invariance under paraphrasing. In a different domain, Chen *et al*. [19] applied self-consistency to code generation, employing execution-based validation to identify correct solutions among multiple candidates. Collectively, these studies demonstrate the general applicability of consistency-based strategies. The present work extends this line of research to structured extraction tasks, where the objective is to generate a set of terms rather than a single deterministic output.

Self-verification in LLMs

Self-verification builds on the observation that verifying a candidate is often easier than generating one, a principle grounded in computational complexity theory. Weng *et al*. [6] introduced backward verification for reasoning tasks, wherein the model assesses whether its generated outputs satisfy the constraints of the original problem. Gero *et al*. [20] extended this idea to clinical information extraction, prompting LLMs to provide supporting evidence for their outputs and to validate extracted information. Their findings indicate that self-verification can yield consistent accuracy improvements across different models in extraction settings.

More recent studies have examined both the capabilities and limitations of self-verification. Huang *et al*. [21] explored iterative self-refinement, demonstrating that models can enhance output quality through successive rounds of self-critique under certain conditions. However, Stechly *et al*. [22] highlighted inherent limitations in complex reasoning and planning tasks, showing that verification performance is highly contingent on task structure. Shinn *et al*. [23] proposed reflexion, a framework in which language agents leverage verbal reinforcement learning to iteratively reflect on and refine their outputs. Building on these insights, the approach proposed in this work integrates verification with consistency-based scoring and applies verification selectively to uncertain candidates, rather than uniformly across all outputs. This design enables a more effective balance between accuracy and computational efficiency.

PROPOSED METHODOLOGY

In this section, we present MP-CoVe, a training-free framework for ATE. Our approach decouples the extraction process into two distinct phases: a multi-perspective generation phase that utilizes diverse prompting strategies to maximize candidate coverage, and a coarse-to-fine verification phase that employs a statistics-guided mechanism to filter noise. Figure 1 illustrates the overall architecture.

Problem definition

Given a domain-specific corpus $\mathcal{D} = \{s_1, s_2, \cdots, s_N\}$ and a domain label d, zero-shot ATE aims to identify a set of valid domain terms $T = \{t_1, t_2, ..., t_M\}$ without parameter updates. When using an LLM for extraction, for a given input $s \in \mathcal{D}$, we query the model with an instruction prompt $p(s, d, E)$ to generate terms $\mathcal{C}_s$:

$$\mathcal{C}_s = \text{LLM}(p(s, d, E); \tau), \tag{1}$$

where $p(s, d, E)$ is a prompt comprising the input content s, the domain description d, and a set of optional demonstrations E (where $E = \varnothing$ in zero-shot settings). τ is the sampling temperature.

Nonetheless, relying solely on a single prompt or sampling pass is insufficient due to the model's high sensitivity to instruction phrasing and inherent stochasticity. Such a monolithic approach often fails to capture

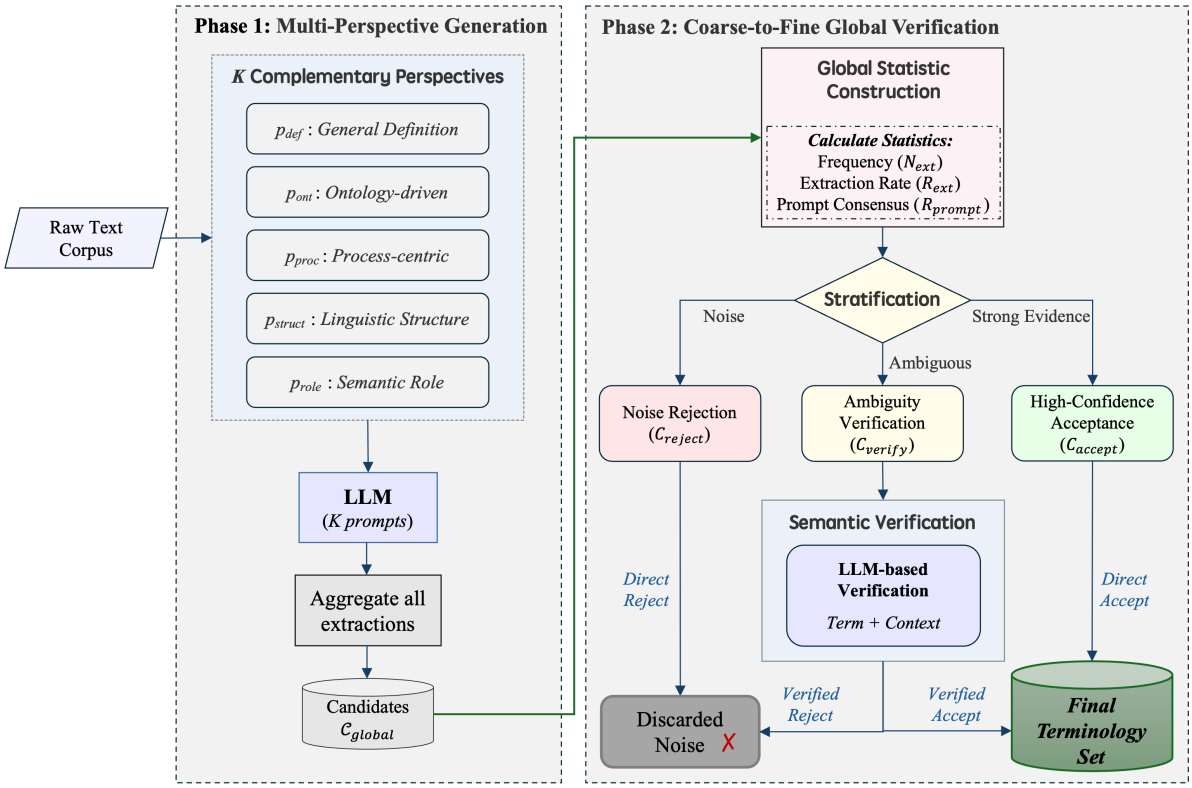

Figure 1 Overview of the MP-CoVe framework. (1) Multi-perspective generation: candidate terms are aggregated from diverse prompt perspectives to ensure comprehensive terminological coverage. (2) Coarse-to-fine global verification: a statistical mechanism utilizes global metrics (N_{ext}, R_{ext}, R_{prompt}) to stratify candidates into high-confidence, noise, and ambiguous sets, reserving fine-grained semantic verification strictly for the ambiguous instances.

the full spectrum of domain terminology while introducing significant noise. To address this issue, we introduce MP-CoVe, a framework that reformulates the extraction task into two cascading phases: (1) Multi-perspective generation; (2) Coarse-to-fine verification.

Phase 1: Multi-perspective candidate generation

Existing studies demonstrate repeated sampling of a single prompt to explore the model's output. However, such an approach is inevitably constrained by the specific bias of the chosen instruction. To address this issue, we employ a multi-perspective prompting strategy. In this respect, we design a set of K complementary prompts $\mathcal{P} = \{p_1, p_2, ..., p_K\}$. Each prompt instructs the LLM to help analyze the input sentence from a distinct semantic perspective, thereby retrieving complementary subsets of terminology to maximize overall coverage.

•**General definition perspective (p_{def}):** Instructing the model to extract terms based on a broad definition of domain terminology. This captures the most salient and unambiguous concepts in domain d.

•**Ontology-driven perspective (p_{ont}):** Guiding the model to identify terms fitting specific taxonomic categories relevant to the domain. By explicitly enumerating key sub-types, this prompt ensures the coverage of distinct semantic classes that might be overlooked by a generic query.

•**Process-centric perspective (p_{proc}):** Framing the extraction task within the context of domain-specific workflows or events. It prompts the model to identify entities involved in dynamic processes, such as actions, state changes, or procedural operations. This captures terms that function as integral components of domain activities rather than static objects.

•**Linguistic structure perspective (p_{struct}):** Focusing on the morphological and syntactic construction of terms. It explicitly targets complex noun phrases, specific compound formations (e.g., modifier-head structures), abbreviations, or other lexical patterns characteristic of technical jargon in the target field.

•**Semantic role perspective (p_{role}):** Deconstructing the sentence into logical semantic arguments. It asks the model to identify terms playing specific functional roles, such as the Locus (spatial location), Instrument (tool used), or Source (origin), thereby capturing terms based on their contextual utility and logical relations.

For each input $s \in \mathcal{D}$, we query the model independently with each prompt $p_k \in \mathcal{P}$. The candidate set for s, denoted as $\mathcal{C}_s^{raw}$, is constructed by aggregating the unique outputs from all perspectives:

$$\mathcal{C}_s^{raw} = \cup_{k=1}^{K} \mathrm{LLM}(p_k(s, d, E); \tau). \tag{2}$$

Finally, the global raw candidate pool $\mathcal{C}_{global}$ is formed by the union of all extractions across the corpus:

$$\mathcal{C}_{global} = \cup_{s \in \mathcal{D}} \mathcal{C}_s^{raw}. \tag{3}$$

This phase prioritizes recall. By aggregating these different perspectives, $\mathcal{C}_{global}$ successfully captures a wide spectrum of terminology, which introduces false positives that require subsequent verification.

Phase 2: Coarse-to-fine verification

Subjecting every candidate in $\mathcal{C}_{global}$ to LLM verification is inefficient and increases the risk of over-correction due to model stochasticity. To address this, we introduce a statistics-guided coarse-to-fine verification mechanism. This approach utilizes corpus-level statistical patterns for candidate stratification,

reserving LLM-based semantic verification strictly for ambiguous cases.

Global statistic construction

For each unique candidate term $t \in \mathcal{C}_{global}$, we traverse the corpus $\mathcal{D}$ to construct two key context sets.

Occurrence set ($\mathcal{S}_{occ}$): The set of all sentences in $\mathcal{D}$ where the term t appears. Let $N_{occ}(t) = |\mathcal{S}_{occ}(t)|$.

Extraction set ($\mathcal{S}_{ext}$): The subset of sentences where t was explicitly extracted by any of the prompts in Phase 1. Let $N_{ext}(t) = |\mathcal{S}_{ext}(t)|$.

Based on these sets, we define two objective metrics to quantify the term-hood of each candidate.

(1) Global extraction rate (R_{ext}): This metric measures the consistency of the term's recognition across its occurrences. A high R_{ext} indicates that the phrase is consistently recognized as a term in the majority of contexts where it appears.

$$R_{ext}(t) = \frac{N_{ext}(t)}{N_{occ}(t)}. \tag{4}$$

(2) Prompt consensus rate (R_{prompt}): This metric measures the agreement between different semantic perspectives. We calculate the ratio of unique prompt types that successfully extracted t, denoted as $n_p(t)$, relative to the total number of prompts K.

$$R_{prompt}(t) = \frac{n_p(t)}{K}. \tag{5}$$

Statistical stratification mechanism

Using the metrics defined above, alongside the absolute extraction frequency $N_{ext}(t)$, we classify candidates into three distinct categories based on predefined thresholds: δ_{ext} (extraction rate), δ_{prompt} (prompt consensus), and δ_{freq} (minimum frequency). Notably, δ_{freq} serves as a stability constraint to prevent low-frequency noise (e.g., terms appearing and extracted only once) from being prematurely accepted.

•**High-confidence acceptance ($\mathcal{C}_{accept}$):** Candidates that exhibit both sufficient frequency and strong consistency. To be accepted without further verification, a term must strictly satisfy the frequency constraint and exceed the thresholds for both extraction rate and prompt consensus:

$$t \in \mathcal{C}_{accept} \Leftrightarrow \left(N_{ext} \geqslant \delta_{freq}\right) \wedge \left(R_{ext} \geqslant \delta_{ext}^{accept}\right) \wedge \left(R_{prompt} \geqslant \delta_{prompt}^{accept}\right). \tag{6}$$

These are high-quality terms with unanimous statistical support and are directly added to the final set $\mathcal{T}$.

•**Noise rejection ($\mathcal{C}_{reject}$):** Candidates that exhibit consistently weak evidence. Terms are discarded if they fail to meet the thresholds for both extraction rate and prompt consensus:

$$t \in \mathcal{C}_{reject} \Leftrightarrow \left(R_{ext} \leqslant \delta_{ext}^{reject}\right) \wedge \left(R_{prompt} \leqslant \delta_{prompt}^{reject}\right). \tag{7}$$

These candidates typically represent hallucinated noise or irrelevant common words.

•**Ambiguity verification ($\mathcal{C}_{verify}$):** Candidates falling into the intermediate range (e.g., frequently extracted but with low consistency). These boundary cases contain valid but ambiguous terms and are passed to the next stage for semantic verification.

LLM-based semantic verification

For candidates falling into the $\mathcal{C}_{verify}$ category, we perform fine-grained verification supported by global contexts. Unlike local verification which examines a single sentence, we aggregate a support set $\mathcal{S}_{support}$ consisting of up to M sentences sampled from the occurrence set $\mathcal{S}_{occ}$.

We construct a verification prompt that presents the candidate term alongside its global contexts. To ensure optimal instruction following, we employ prompts in the native language of the dataset (i.e., English prompts for the ACTER dataset and Chinese prompts for the CMeEE dataset). For clarity, the English template of our prompt is presented below:

Verify whether the following candidate terms are standard professional terms in the field of {domain}.

Contexts: $\{s_1, s_2, ..., s_M\}$

Candidate: *term*

Judgment Rules:

• *Mark as **1** if it is a valid proprietary term in this field.*

• *Mark as **0** if it is a general word, verb, adjective, generic noun, or conjunction.*

Requirement: *For each candidate, **analyze** whether it fits the domain definition (or why not) first, and then provide the judgment.*

Output Format: *[Candidate]: [Analysis] → [1 or 0]*

Algorithm 1 MP-CoVe framework for training-free term extraction

Input: Corpus $\mathcal{D}$, prompt set $\mathcal{P}$, thresholds $\left\{\delta_{freq}, \delta_{ext}^{accept}, \delta_{prompt}^{accept}, \delta_{ext}^{reject}, \delta_{prompt}^{reject}\right\}$

Output: Final term set $\mathcal{T}$

// **Phase 1: Multi-Perspective Generation**

Initialize candidate pool $\mathcal{C}_{global} \leftarrow \varnothing$;

for $s_i \in \mathcal{D}$ **do**

  for $p_k \in \mathcal{P}$ **do**

    $\mathcal{C}_{local} \leftarrow \text{LLM}\left(p_k(s_i, d); \tau\right); \mathcal{C}_{global} \leftarrow \mathcal{C}_{global} \cup \mathcal{C}_{local}$

  end

end

// **Phase 2: Coarse-to-Fine Verification**

Initialize $\mathcal{T} \leftarrow \varnothing$;

for *each candidate term* $t \in \mathcal{C}_{global}$ **do**

  // Step 2.1: Statistic Construction

  Calculate absolute frequency $N_{ext}(t)$ and prompt count $n_p(t)$; Calculate occurrences $N_{occ}(t)$ and occurrence set $\mathcal{S}_{occ}(t)$ in $\mathcal{D}$; Compute metrics: $R_{ext} \leftarrow N_{ext} / N_{occ}, R_{prompt} \leftarrow n_p / K$

  // Step 2.2: Stratification & Verification

  if $N_{ext} \geqslant \delta_{freq} \wedge R_{ext} \geqslant \delta_{ext}^{accept} \wedge R_{prompt} \geqslant \delta_{prompt}^{accept}$ **then** // High-Confidence: Accept

    $\mathcal{T} \leftarrow \mathcal{T} \cup \{t\}$

  else if $R_{ext} \leqslant \delta_{ext}^{reject} \wedge R_{prompt} \leqslant \delta_{prompt}^{reject}$ **then** // Noise: Reject

    Discard t

  else // Ambiguous: Verify

    $\mathcal{S}_{support} \leftarrow SampleContexts\ (\mathcal{S}_{occ}(t))$

    $Verdict \leftarrow \text{LLM}_{Verify}(t, \mathcal{S}_{support})$

    if $Verdict = 1$ **then**

      $\mathcal{T} \leftarrow \mathcal{T} \cup \{t\}$

(*Continued*)

Algorithm 1 MP-CoVe framework for training-free term extraction

```
            end
        end
    end
    return 𝒯
```

Upon completion of the verification process, we parse the model's judgment. Candidates receiving a positive label are incorporated into the final terminology set $\mathcal{T}$, while those marked as **0** are discarded.

The complete execution procedure of our framework is summarized in Algorithm 1. The process begins with the generation of a raw candidate pool via multi-perspective prompting (Lines 1–6). Subsequently, statistical metrics are computed for each candidate (Line 9) to categorize them into three categories. High-confidence terms are directly accepted, low-confidence noise is discarded, and ambiguous candidates undergo LLM-based semantic verification using their global context support sets (Lines 10–20).

EXPERIMENTS

Datasets

To assess both the effectiveness and generalizability of MP-CoVe, experiments are conducted on two widely used benchmark datasets encompassing English and Chinese.

ACTER (English). The annotated corpora for term extraction research (ACTER) dataset, version 1.5 [1], is a standard benchmark for ATE[1)]. In this study, we focus on the HTFL domain subset, which comprises unstructured clinical texts characterized by dense and specialized terminology. The dataset provides four annotation categories: specific term, common term, out-of-domain (OOD) term, and named entity. In accordance with the conventional scope of terminology extraction, instances labeled as named entity are excluded, while the remaining three categories are retained as valid ground truth.

CMeEE (Chinese). The CMeEE dataset is derived from the Chinese biomedical language understanding evaluation (CBLUE) benchmark [24] and consists of medical texts annotated with nine entity types. To adapt this named entity recognition (NER) dataset for the ATE task, annotated entities are treated as domain-specific terms. The ground truth is further refined by heuristic filtering to better align with the linguistic definition of terminology. Specifically, candidates that violate length constraints (restricted to 2–25 characters) or exhibit non-nominal patterns characteristic of sentence fragments, such as verbs, adverbs, conjunctions, numeric expressions, and punctuation, are excluded. As the test set annotations are not publicly available, evaluation is conducted on the validation split.

To ensure a rigorous experimental protocol and to prevent potential data leakage, a consistent partitioning strategy is applied across both datasets. For each dataset, the first five documents of the target corpus are reserved to construct a candidate pool for few-shot demonstrations. All quantitative evaluations are performed exclusively on the remaining documents. In the primary experimental setting, MP-CoVe operates under a strictly zero-shot setting, processing only the raw input texts without access to ground-truth annotations. Detailed dataset statistics are summarized in Table 1.

1) https://github.com/AylaRT/ACTER.

Table 1 Statistical overview of benchmark datasets. Summary of the English (ACTER) and CMeEE datasets used for evaluation, detailing the language, domain scope, and corpus size[a)].

Dataset	Lang	Domain	# Docs	# Terms
ACTER-HTFL	EN	Heart failure	190	2361
CMeEE	ZH	Medical	5000	11,776

a) # Terms denotes the total number of unique terms. # Docs denotes the number of test instances: for ACTER-HTFL, each instance corresponds to one document; for CMeEE, each instance corresponds to one sentence-level test sample.

Baselines

We compare MP-CoVe against comprehensive LLM-based extraction paradigms. All baselines utilize the same backbone model as our method to ensure a fair comparison.

•**Direct generation (zero-shot):** The standard extraction approach. We query the LLM with a single, generic definition-based prompt to extract terms in a one-pass manner. To minimize generation randomness, we use a deterministic sampling strategy with a temperature of $\tau = 0.1$.

•**Few-shot:** An enhanced baseline that prepends 5 demonstrations (selected from the reserved pool described in the Datasets subsection) to the prompt. Inference is performed with $\tau = 0.1$.

•**Self-consistency:** A state-of-the-art ensemble strategy [5] designed to mitigate generation instability. We sample the zero-shot prompt 5 times using varying temperatures ($\tau \in \{0.3, 0.5, 0.6, 0.7, 0.8\}$) to encourage diversity. A candidate is retained only if it appears in at least 3 out of 5 generations (majority voting).

•**Generate-Verify:** A two-stage pipeline baseline [6,20]. Candidates are first generated using the zero-shot prompt and then self-verified by the model. Both stages employ a low temperature ($\tau = 0.1$). Crucially, unlike our method, this standard verification assesses terms solely within the local sentence context from which they are extracted.

Evaluation metrics

We evaluate ATE performance using precision (P), recall (R), and $F1$-score ($F1$). To accurately reflect the overall extraction quality across the dataset, we adopt a corpus-level calculation strategy (micro-average). Specifically, the counts of true positives (TP), false positives (FP), and false negatives (FN) are accumulated across all test documents to compute the final global metrics, defined as:

$$P = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} (TP_i + FP_i)}, R = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} (TP_i + FN_i)}, F1 = \frac{2 \cdot P \cdot R}{P + R}, \tag{8}$$

where N denotes the total number of documents in the corpus.

We enforce strict exact match criteria for evaluation. An extracted candidate is considered a true positive only if it is identical to the ground-truth label (case-insensitive for English).

Implementation details

Backbone model. We employ Qwen3-14B[2)] [25] as the backbone LLM for all experiments to ensure a fair comparison. The model is deployed using the vLLM framework to enable high-throughput inference. All

2) https://huggingface.co/Qwen/Qwen3-14B.

experiments are conducted using 2× NVIDIA A100 GPUs, utilizing tensor parallelism.

MP-CoVe settings

Our framework operates strictly in a zero-shot setting and utilizes a consistently low temperature ($\tau = 0.1$) throughout all stages to ensure reproducibility and stability.

•**Phase 1: Multi-perspective generation.** We set the number of perspectives $K = 5$ and use all the 5 prompts introduced before.

•**Phase 2: Stratification & verification.** Candidates are stratified into accept, reject, and verify sets based on frequency and consensus statistics. To ensure the robustness of our results and avoid overfitting, the specific threshold configurations for extraction rate and prompt consensus are analyzed in detail in the parameter sensitivity analysis. Finally, for candidates requiring verification, the LLM acts as a judge with $\tau = 0.1$.

EXPERIMENTAL RESULTS AND ANALYSIS

Main results

Table 2 presents a comparative evaluation of MP-CoVe against four representative baseline paradigms on both the English ACTER dataset and the CMeEE dataset. The results indicate that MP-CoVe achieves the highest *F*1-scores across both benchmarks, thereby substantiating the effectiveness of the proposed multi-perspective framework. Several key observations emerge from this comparison.

Substantial gains in recall. A primary strength of MP-CoVe lies in its enhanced ability to identify a broader range of domain-specific terms. On the ACTER dataset, the proposed method attains a recall of 70.54%, exceeding the strongest baseline in this regard (zero-shot prompting) by 6.73 percentage points. A similar trend is observed on the CMeEE dataset, where MP-CoVe improves recall by 11.48 percentage points over the Generate-Verify baseline (72.62% versus 61.14%). Conventional zero-shot and few-shot approaches tend to rely on generalized definitional prompts, which bias the model toward prototypical term instances and constrain recall (typically within the 61%–63% range). In contrast, MP-CoVe employs multiple semantic perspectives to expand the search space, thereby enabling the extraction of diverse term variants, including complex compound expressions and abbreviations that are often overlooked by single-prompt strategies.

Overall performance (*F*1-score). MP-CoVe achieves the highest *F*1-scores on both datasets, reaching 51.05% on ACTER and 63.21% on CMeEE. While certain baselines, such as Generate-Verify, attain

Table 2 Main comparison results on ACTER (English) and CMeEE (Chinese) datasets (%). The best results are highlighted in bold. All methods utilize the same Qwen3-14B backbone.

Method	ACTER (Heart failure)			CMeEE (Chinese medical)		
	P	*R*	*F*1	*P*	*R*	*F*1
Zero-shot	37.50	63.81	47.24	61.63	62.29	61.96
Few-shot	43.02	61.21	50.53	63.54	61.57	62.54
Self-consistency	38.60	62.76	47.80	62.55	61.38	61.96
Generate-Verify	**44.66**	59.22	50.92	**64.72**	61.14	62.88
MP-CoVe (Ours)	40.00	**70.54**	**51.05**	55.97	**72.62**	**63.21**

competitive *F*1-scores due to relatively high precision, their overall performance remains limited by insufficient recall. The results suggest that prioritizing recall improvements, even at the expense of moderate reductions in precision, can yield a more favorable balance between the two metrics, ultimately leading to superior aggregate performance. This finding highlights the suitability of MP-CoVe for applications where comprehensive term coverage is essential.

Precision–recall trade-off. MP-CoVe exhibits comparatively lower precision than methods such as few-shot prompting and Generate-Verify, reflecting a fundamental divergence in extraction strategies. On the one hand, precision-oriented baselines adopt a conservative extraction paradigm, achieving higher precision (e.g., Generate-Verify reaches 44.66% on ACTER) but failing to surpass a recall ceiling of approximately 64%. This suggests that such approaches predominantly capture high-confidence, canonical terms. On the other hand, MP-CoVe follows a recall-driven strategy characterized by broader exploration. By aggregating candidates from multiple perspectives, the framework substantially enlarges the candidate space. Although this expansion introduces additional noise, resulting in a precision of 40.00%, the corresponding gain in recall (+11.32% relative to Generate-Verify) demonstrates the effectiveness of this approach. Given the inherent difficulty of identifying long-tail terminology in ATE, this trade-off is well-justified, as high recall is often a prerequisite for achieving strong performance in downstream applications.

Ablation study

To systematically assess the contribution of each component within MP-CoVe and to elucidate the role of multi-perspective generation, we conduct a comprehensive ablation study on the CMeEE dataset. The full MP-CoVe framework is compared against four variants: (1) w/o multi-perspective, which employs a single definition-based prompt for candidate generation followed by the proposed coarse-to-fine verification module; (2) w/o stratification & verification, which directly aggregates candidates from all five perspectives without any filtering; (3) w/o verification, which retains the stratification step but accepts both high-confidence and ambiguous candidates without LLM-based verification; and (4) w/o stratification, which removes statistical stratification and subjects all generated candidates to uniform LLM-based verification. The results are summarized in Table 3.

Impact of multi-perspective generation. The findings underscore the critical importance of diversity in candidate generation. As shown in Table 3, removing the multi-perspective component (w/o multi-perspective) leads to a substantial decrease in recall, dropping by 12.07 percentage points (from 72.62% to 60.55%). This result indicates that reliance on a single, generic definitional prompt imposes a significant coverage constraint, limiting the model's ability to retrieve long-tail and domain-specific terminology. In

Table 3 Ablation study on the CMeEE dataset. We systematically evaluate the contribution of the multi-perspective generation module and the coarse-to-fine filtering mechanism (stratification & verification). Results are reported in percentage (%). The best results are highlighted in bold.

Variant	*P*	*R*	*F*1
MP-CoVe (Full)	55.97	72.62	**63.21**
w/o Multi-perspective	**64.67**	60.55	62.54
w/o Stratification & verification	53.30	**73.49**	61.78
w/o Verification	53.76	73.36	62.05
w/o Stratification	55.57	72.63	62.97

contrast, the full MP-CoVe framework leverages complementary semantic perspectives to expand the candidate space, thereby overcoming this bottleneck. These observations suggest that semantic diversity is a key prerequisite for achieving high recall in term extraction.

Effectiveness of coarse-to-fine filtering. With respect to the filtering mechanism, the results reveal that naively aggregating all candidates (w/o stratification & verification) yields the highest recall but at the expense of substantially reduced precision (53.30%) and overall $F1$-score. This finding highlights that while extensive exploration enhances coverage, it must be coupled with an effective selection strategy to ensure output quality. Notably, the full MP-CoVe framework outperforms the variant that applies LLM-based verification uniformly to all candidates (w/o stratification). This comparison suggests that, for high-confidence candidates, corpus-level statistical consensus provides a sufficiently reliable indicator of correctness. Subjecting such candidates to redundant verification not only fails to improve performance but may also introduce instability due to model stochasticity, as reflected in the slight decline in $F1$-score (from 63.21% to 62.97%). Overall, the proposed stratification mechanism effectively balances computational efficiency and predictive robustness by reserving semantic verification for genuinely ambiguous cases while preserving high-confidence candidates through statistical evidence.

Analysis of perspective complementarity

To investigate why multi-perspective prompting leads to superior recall, we conduct a fine-grained analysis of the extracted terms from three dimensions: quantitative overlap, semantic specialization, and performance on long-tail entities.

Quantitative contribution of perspectives

We first analyze the number of unique terms extracted by each perspective on the CMeEE dataset, as illustrated in Figure 2. The results indicate that 10,172 terms are extracted by all five prompts, representing a robust consensus set. In addition to this intersection, individual perspectives provide significant unique contributions that are statistically non-overlapping. Specifically, the process-centric perspective identifies the largest number of unique terms (2535), followed by the linguistic structure perspective (1940). This confirms that while different prompts share a common core of high-frequency terms, they effectively retrieve distinct subsets of knowledge, validating the additive value of our prompt design.

Semantic specialization

Beyond quantitative overlap, we verify whether the multi-perspective prompts function as designed. Table 4 presents representative unique terms extracted exclusively by each perspective, revealing distinct semantic tendencies that align with our method definitions:

Ontology-driven perspective (p_{ont}): By guiding the model to scan for taxonomic categories, this prompt excels at identifying canonical disease names and hierarchical concepts (e.g., organic acid metabolic diseases). Such terms represent the static taxonomic knowledge of the domain but are often missed by generic queries that lack explicit categorical constraints.

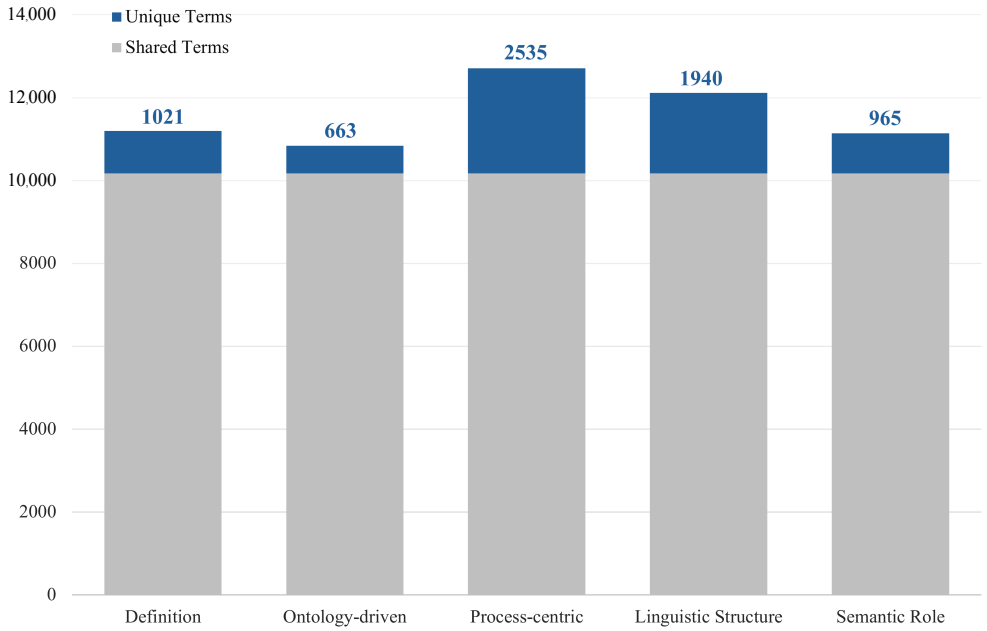

Figure 2 Analysis of perspective complementarity on CMeEE. The gray bar represents the number of consensus terms extracted by all five prompts. The blue bars indicate the count of unique terms extracted exclusively by each specific perspective.

Table 4 Qualitative analysis of unique terms extracted by different perspectives. We select representative terms that were exclusively retrieved by specific prompts. Note that the definition prompt serves as a general baseline, while the four perspectives below demonstrate clear semantic specializations.

Perspective	Semantic focus	Representative unique terms (Chinese/English)
Ontology-driven (p_{ont})	Standard diseases & hierarchical concepts	巨细胞动脉炎 (Giant cell arteritis) 原发性糖尿病 (Primary diabetes) 有机酸代谢性疾病 (Organic acid metabolic diseases)
Process-centric (p_{proc})	Procedures, symptoms & dynamic states	动脉导管结扎 (Arterial duct ligation) 轻度脱水 (Mild dehydration) 颅骨脑回压迹增多 (Increased convolutional markings)
Linguistic structure (p_{struct})	Complex compounds, codes & abbreviations	抗双链DNA抗体 (Anti-dsDNA antibody) VDJ基因 (VDJ gene) FMR-1前突变 (FMR-1 premutation)
Semantic role (p_{role})	Anatomical sites (locus) & spatial regions	肺动脉环 (Pulmonary artery ring) 后尿道瓣膜处 (Posterior urethral valve site) 室壁心肌 (Ventricular wall myocardium)

Process-centric perspective (p_{proc}): Framing the extraction task within domain workflows enables the retrieval of entities involved in clinical events and state changes. For instance, terms such as arterial duct ligation (procedure) and mild dehydration (state) are successfully captured. This confirms that p_{proc} effectively targets dynamic components of domain activities rather than isolated static objects.

Linguistic structure perspective (p_{struct}): Focusing on morphological construction proves critical for recovering long-tail technical jargon. This perspective uniquely identifies complex compound terms and alphanumeric codes, such as anti-dsDNA antibody and VDJ gene. These entities, characterized by non-standard syntactic structures, are typically filtered out by models relying on standard grammatical templates.

Semantic role perspective (p_{role}): By deconstructing sentences into functional arguments (specifically the Locus), this prompt ensures the coverage of spatial and anatomical entities. Terms such as pulmonary artery

ring and posterior urethral valve site are extracted exclusively by this perspective, thus highlighting the importance of explicitly querying for spatial relations.

Collectively, these distinct extraction patterns demonstrate that MP-CoVe aggregates complementary semantic evidence, thereby maximizing the overall terminological coverage.

Performance on long-tail terms

Conventional single-prompt baselines often struggle with long-tail terms, defined as rare terms that appear infrequently in the corpus. To evaluate MP-CoVe's capability in this aspect, we categorize the ground-truth terms into three groups based on their corpus-level frequency: rare (1–2 occurrences), medium (3–9 occurrences), and frequent (⩾ 10 occurrences). We focus on recall for this analysis, as assigning false positives to specific frequency groups is inherently infeasible.

As shown in Table 5, the performance divergence between the Generate-Verify baseline and our method amplifies as term frequency decreases. For frequent and medium terms, both methods maintain high recall levels (exceeding 90%), with MP-CoVe yielding marginal to moderate gains (+0.41% and +2.08%, respectively). This suggests that the definition-based prompt is generally sufficient for retrieving repetitive, prototypical concepts. However, the most significant contrast appears in the rare group. While the single-prompt baseline suffers a marked decline to 57.45%, MP-CoVe exhibits superior robustness by maintaining a 70.61% recall, delivering a substantial improvement of +13.16%. These results indicate that the multi-perspective mechanism introduced by MP-CoVe is essential for uncovering low-frequency, obscure terms that are typically missed by baseline approaches.

Error analysis and case study

To provide a comprehensive understanding of the framework, we performed an error analysis on the CMeEE dataset. We first examine the filtering effectiveness of the coarse-to-fine pipeline by categorizing the candidates rejected at each stage, then present a qualitative analysis of the remaining errors categorized by false negatives, false positives, and verification errors against the ground-truth annotations.

Filtering effectiveness analysis

To evaluate whether both filtering stages are necessary, we cross-compare the full pipeline against single-stage ablations. Table 6 summarizes the results, revealing a strong complementarity between the two stages.

Table 5 Recall performance across different term frequency groups. We categorize ground-truth terms into rare, medium, and frequent groups based on their occurrence counts in the CMeEE dataset. Δ denotes the absolute recall improvement of MP-CoVe over Generate-Verify. Bold values indicate the better performance within each group.

Group	Frequency	Recall (%)		Δ (%)
		Generate-Verify	MP-CoVe	
Rare	1–2	57.45	**70.61**	**+13.16**
Medium	3–9	90.34	**92.42**	+2.08
Frequent	⩾ 10	93.88	**94.29**	+0.41

Table 6 Complementarity analysis of the two filtering stages on the CMeEE dataset. Counts and representative examples are derived from cross-comparing the full pipeline against single-stage ablations. English translations are provided in parentheses where applicable.

Scenario	Count	Representative examples
Noise that only stratification catches (verification would miss)	727	萎缩 (atrophy), 400 mg/(kg·d), 蛛网膜下腔穿刺(subarachnoid puncture), gestational age
Noise that only verification catches (stratification would miss)	1074	水果 (fruit), 温度 (temperature), Taussiglm, 18岁 (age 18)
Valid terms protected by stratification (verification would over-correct)	26	全身血管 (systemic vasculature), 催化酶 (catalytic enzyme), 代谢率改变 (metabolic rate alteration)

Of the 845 correctly rejected noise terms in the stratification stage, 727 (86%) would have been incorrectly accepted by the LLM verifier as valid terminology if sent to verification directly, as their surface forms appear sufficiently domain-relevant (e.g., 萎缩 (atrophy), 400 mg/(kg d)) despite their weak corpus-level statistical signals. Conversely, all 1074 correctly rejected candidates in the verification stage would have passed stratification, as they exhibit plausible statistical profiles despite being non-terminological (e.g., 水果 (fruit), Taussiglm). These two patterns confirm that the statistical and semantic filtering mechanisms address complementary types of noise.

In addition to noise filtering, the stratification stage also plays a protective role. The comparison of the full framework against the w/o stratification variant reveals 26 valid ground-truth terms that are correctly preserved by the stratification module as high-confidence direct accepts but would have been incorrectly rejected under LLM verification. For instance, 全身血管 (systemic vasculature) and 催化酶 (catalytic enzyme) possess strong statistical evidence yet were rejected by the LLM verifier in the ablated setting, largely due to their specialized nature exceeding the model's domain knowledge. Such over-correction phenomenon empirically validates the design rationale of our stratification mechanism. For high-confidence candidates supported by strong corpus-level evidence, statistical consensus provides more reliable judgment than LLM-based verification, which is susceptible to stochastic errors on highly specialized terminology.

Qualitative error case analysis

We further analyze the errors made by MP-CoVe by examining false negatives, false positives, and verification errors against the ground-truth annotations. Table 7 presents a systematic categorization with representative examples.

False negative analysis. Among the 3303 missed terms, the dominant source (94.5%) is terms that were never extracted by any of the five perspectives, revealing systematic blind spots in LLM-based extraction. These are concentrated in four categories. (1) Short polysemous words that also function as common vocabulary components. The LLM tends to extract longer, more specific compound terms while overlooking their constituent morphemes as independent terms. (2) English abbreviations that appear without expanded forms and thus lack contextual cues for recognition. (3) Mixed Chinese-English terms that pose tokenization challenges at code-switched boundaries. (4) Chinese compound phrases, the largest category, consisting of multi-character expressions that describe specific clinical findings or pathological conditions. These terms are rarely reproduced verbatim by the model, resulting in surface-form mismatches with the ground truth. A smaller portion (5.5%) consists of 183 terms that were correctly extracted by at least one perspective but subsequently removed by the coarse-to-fine pipeline (41 by stratification and 142 by verification).

Table 7 Representative categories of error cases on the CMeEE dataset. English translations are provided in parentheses where applicable. In the representative examples, gold denotes the ground-truth answer in the annotation.

Error	Category	Count	Representative examples
False negative (3303)	Short polysemous terms	119	受损 (Damaged), 癌症 (Cancer), 腭弓 (Palatal arch)
	English abbreviations	231	DKA, DNA, HLA, ICU, MTX
	Mixed Chinese-English terms	439	DQ位点 (DQ locus), 多巴胺D2受体 (Dopamine D2 receptor)
	Chinese compound phrases	2271	反复下呼吸道感染 (Recurrent lower respiratory tract infection), 肾素和醛固酮分泌不足 (Renin and aldosterone insufficiency)
	Extracted but filtered	183	皮质 (Cortex), 狼疮 (Lupus), 循环 (Circulation)
False positive (7414)	Boundary mismatch (subset of gold)	2477	支原体肺炎 (Mycoplasma pneumonia) → gold: 肺炎支原体肺炎 (Mycoplasma pneumoniae pneumonia); 感染灶 (Infection focus) → gold: 隐性感染灶 (Occult infection focus)
	Boundary mismatch (superset of gold)	2992	脑膜炎球菌 (Meningococcus) → gold: 脑膜炎 (Meningitis); 肝功能随访 (Liver function follow-up)→ gold: 肝功能 (Liver function)
	General medical vocabulary	23	患儿 (Pediatric patient), 临床表现 (Clinical manifestation)
	Dosages and formulas	337	500 μg/吸, 250 mg/kg, 5年无病生存率
Verification error (142)	Polysemous short terms	40	皮质 (Cortex), 糖尿 (Diabetes), 脂肪 (Fat)
	Rare specialized terms	84	风湿性感染(Rheumatic infection), 预防感染 (Infection prevention)

False positive analysis. Of the 7414 false positive terms, the most notable finding is that 73.8% (5469 cases) arise from boundary mismatches with gold-standard terms rather than from entirely spurious extractions. In 2477 cases, the model extracts a shorter span that is a substring of the gold term (e.g., 支原体肺炎 (mycoplasmal pneumonia) vs. gold 肺炎支原体肺炎 (mycoplasma pneumoniae pneumonia)). In 2992 cases, the model extracts a longer span that contains the gold term (e.g., 肝功能随访 (liver function follow-up) and gold 肝功能 (liver function)). This pattern suggests that a substantial fraction of measured false positives reflects annotation boundary inconsistency rather than genuine model errors. The remaining false positives include general medical vocabulary that exhibits high corpus frequency yet is not annotated as domain-specific, and dosages or formulas whose alphanumeric structure superficially resembles technical codes.

Verification error analysis. Among the 1216 candidates rejected by LLM verification, 142 (11.7%) are valid ground-truth terms, constituting the verification-stage subset of the false negatives discussed above. These wrongly rejected terms fall into two categories. (1) Polysemous short words such as 皮质 (cortex), 糖尿 (diabetes), and 脂肪 (fat), where the LLM interpreted them as general vocabulary rather than domain-specific terminology despite their annotated status in the CMeEE ground truth. (2) Rare specialized expressions such as 风湿性感染 (rheumatic infection) and 预防感染 (infection prevention), where the model lacked sufficient domain knowledge for accurate judgment. Together with the 26 over-correction cases identified in the filtering effectiveness analysis, these errors illustrate a fundamental limitation of LLM-based verification. Its judgment is reliable for prototypical domain terms but degrades at both ends of the specificity spectrum.

Parameter sensitivity analysis

To assess the robustness of MP-CoVe, we investigate the impact of two key hyperparameters: the number of perspectives (K) and the thresholds for stratification.

Impact of perspective count (K)

To evaluate the contribution of perspective diversity, we measure the Micro-Recall of the candidate

generation phrase as we incrementally increase K. We report the average recall across all possible combinations of K perspectives, along with the standard deviation to indicate performance stability. As shown in Figure 3, our results demonstrate a robust positive correlation between perspective diversity and coverage:

Recall improvement. The average recall rises steadily from 58.50% at $K = 1$ to 73.49% at $K = 5$, yielding an absolute improvement of +14.99%. This confirms that different perspectives retrieve complementary subsets of the terminology, effectively bridging the semantic gaps inherent in single-prompt extractions.

Robustness to prompt selection. Notably, the standard deviation drops sharply from ±6.35% ($K = 1$) to ±0.78% ($K = 4$). The high variance at $K = 1$ indicates that single-prompt baselines are highly sensitive to specific prompt designs. In contrast, the rapid convergence of performance variance at $K \geqslant 2$ demonstrates that aggregating perspectives significantly mitigates this sensitivity, ensuring consistent and robust performance regardless of the specific combination of prompts used.

Impact of stratification thresholds

We investigate the sensitivity of MP-CoVe to the stratification hyperparameters defined in Eqs. (7) and (8). Our analysis focuses on how these thresholds balance the trade-off between extraction quality ($F1$-score) and system efficiency (verification cost).

For the high-confidence acceptance ($\mathcal{C}_{accept}$), we fix the prompt consensus threshold at $\delta_{prompt}^{accept} = 0.8$. This setting requires a majority agreement (at least 4 out of 5 perspectives), which serves as a fundamental baseline for reliability. Consequently, our analysis examines the interaction between the frequency constraint (δ_{freq}) and the extraction rate (δ_{ext}^{accept}).

Analysis of acceptance criteria (δ_{freq} and δ_{ext}^{accept}). To isolate the impact of the acceptance thresholds on extraction quality, we conduct a grid search analysis while keeping the rejection parameters constant. Specifically, we fix the rejection thresholds at $\delta_{ext}^{reject} = 0.1$ and $\delta_{prompt}^{reject} = 0.4$. We then evaluate the $F1$-score

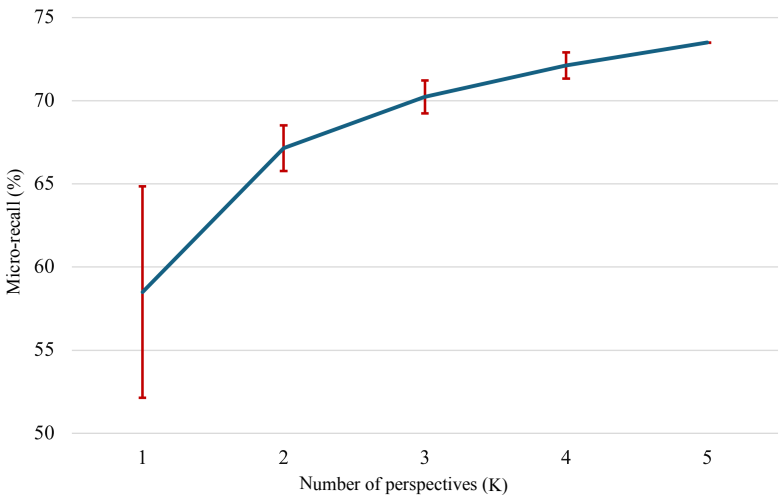

Figure 3 Micro-recall of phase 1 (candidate generation) across different numbers of perspectives (K). The curve illustrates the mean recall, with error bars representing the standard deviation across all possible prompt combinations.

by varying the acceptance extraction rate (δ_{ext}^{accept}) from 0.3 to 0.8 across three different frequency constraints (δ_{freq} from 1 to 4). The resulting performance for precision, recall, and $F1$-score are visualized in Figure 4. The analysis reveals two critical characteristics of our framework.

Insensitivity to hyperparameter variations. As illustrated in the $F1$-score heatmap in Figure 4, the framework exhibits high stability across the parameter grid. The $F1$-scores fluctuate within a narrow margin (63.13%–63.23%), resulting in a uniform color distribution. This indicates that MP-CoVe is robust to variations in acceptance thresholds.

We attribute the overall stability of the $F1$-score to the balanced interaction between precision and recall. Specifically, increasing the frequency constraint (δ_{freq}) from 1 to 2 yields a clear improvement in precision (from ~55.7% to ~56.0%) by effectively filtering single-occurrence candidates. This gain is counterbalanced by a marginal decrease in recall (~0.3%). Furthermore, the framework remains robust to variations in the extraction rate (δ_{ext}^{accept}). As shown in the vertical axis of the heatmaps, varying this threshold results in negligible performance fluctuations. This indicates that the subsequent verification stage effectively processes candidates that fall below the high-confidence acceptance criteria, preventing valid terms with lower extraction rates from being lost.

Analysis of noise rejection criteria (δ_{ext}^{reject} and δ_{prompt}^{reject}). The rejection mechanism ($\mathcal{C}_{reject}$) aims to discard obvious noise to reduce the computational burden on the verification module. We fix $\delta_{freq} = 2$ and $\delta_{ext}^{accept} = 0.5$, and investigate the impact of the noise rejection thresholds. Specifically, we evaluate the performance by varying the rejection extraction rate (δ_{ext}^{reject}) from 0.1 to 0.4 and the prompt consensus threshold (δ_{prompt}^{reject}) across {0.2, 0.4, 0.6}. The experimental results, summarized in Figure 5, reveal a distinct trade-off compared to the acceptance phase.

Sensitivity of recall to rejection thresholds. Unlike the stability observed in the acceptance phase, the framework demonstrates significant sensitivity to rejection criteria. As the prompt consensus threshold (δ_{prompt}^{reject}) increases from 0.2 to 0.6, recall decreases notably from 72.69% to 68.86%. This trend suggests that valid long-tail terminology often exhibits weak prompt consensus. Consequently, applying high rejection

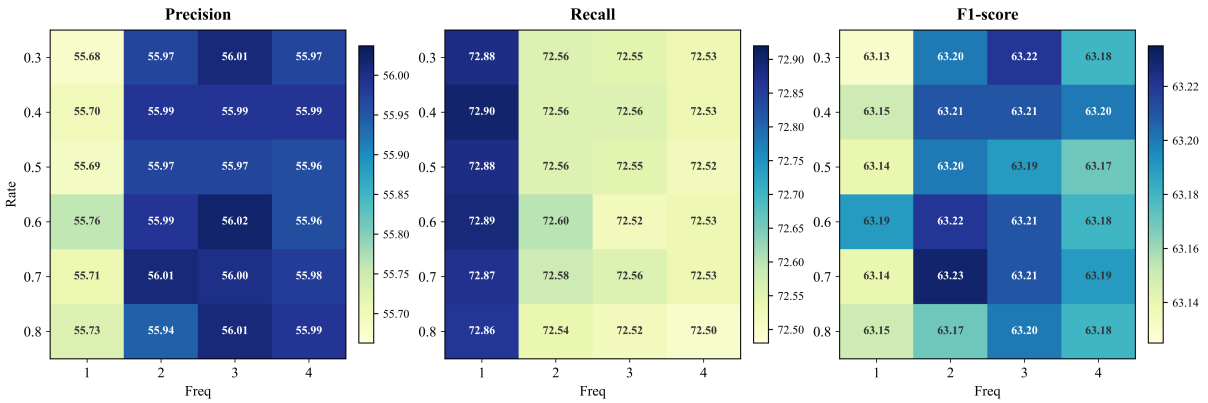

Figure 4 Sensitivity analysis of δ_{freq} and δ_{ext}^{accept}. The X-axis represents the frequency constraint (δ_{freq}), and the Y-axis denotes the acceptance extraction rate (δ_{ext}^{accept}). The three heatmaps illustrate precision, recall, and $F1$-score from left to right, respectively.

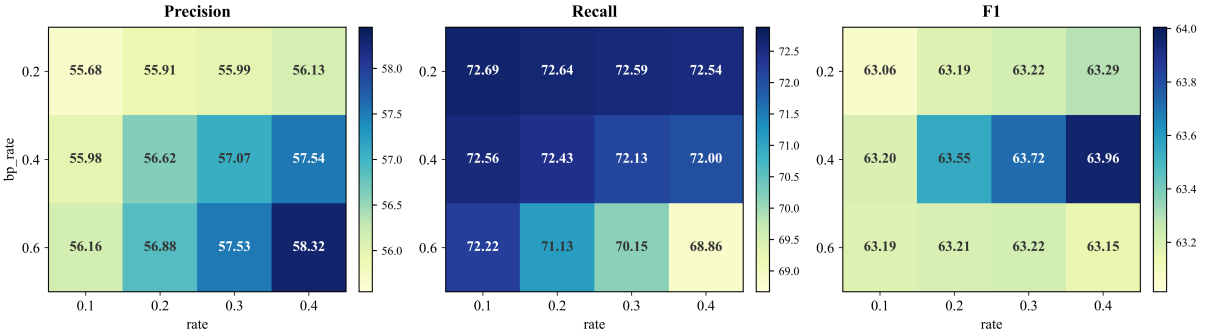

Figure 5 Parameter sensitivity analysis of δ_{ext}^{reject} and δ_{prompt}^{reject}. The *X*-axis represents the rejection extraction rate (δ_{ext}^{reject}), and the *Y*-axis denotes the prompt consensus threshold (δ_{prompt}^{reject}).

thresholds ($\delta_{prompt}^{reject} \geqslant 0.6$) filters out these candidates before the verification stage, resulting in irreversible information loss.

Precision improvement via strict filtering. Conversely, stricter rejection criteria effectively reduce noise in the candidate pool. Setting $\delta_{prompt}^{reject} = 0.6$ and $\delta_{ext}^{reject} = 0.4$ yields the highest precision (58.32%) by eliminating low-confidence candidates. However, this improvement comes at the direct cost of the recall degradation described above.

Optimal threshold configuration. The *F*1-score heatmap indicates that optimal performance is achieved at intermediate thresholds. The configuration of $\delta_{prompt}^{reject} = 0.4$ and $\delta_{ext}^{reject} = 0.4$ results in the peak *F*1-score of 63.96%. This setting effectively removes obvious noise (increasing precision to 57.54%) while retaining a sufficient number of candidates for verification (maintaining recall at 72.00%).

CONCLUSION

This study introduces MP-CoVe, a training-free framework for domain-specific terminology extraction that integrates multi-perspective candidate generation with a coarse-to-fine verification mechanism. In the generation stage, diverse prompting strategies are employed to elicit candidate terms from multiple semantic viewpoints, thereby enhancing coverage of domain-relevant terminology. To refine these candidates, the framework leverages corpus-level statistical signals to partition them into high-confidence, ambiguous, and noisy subsets, enabling a principled balance between recall and precision. Experimental results on domain-specific benchmarks demonstrate the effectiveness of the proposed approach. Further analysis indicates that MP-CoVe exhibits strong robustness with respect to acceptance thresholds, as the subsequent verification stage can effectively handle candidates that fall below strict statistical criteria. In contrast, the framework is more sensitive to rejection thresholds, where conservative filtering is necessary to avoid discarding valid long-tail terms.

The training-free nature of MP-CoVe, coupled with its strong recall performance on rare and specialized terminology, makes it well-suited for practical applications such as terminology extraction from electronic health records, where complex vocabularies, long-tail distributions, and limited annotations pose persistent

challenges. Its high-recall characteristics are also advantageous for knowledge graph construction and multilingual terminology management, as evidenced by consistent performance across both English and Chinese datasets.

Nevertheless, error analysis reveals that LLM-based verification remains vulnerable to misclassification at both ends of the specificity spectrum, occasionally rejecting short polysemous terms as well as infrequent but valid domain expressions. Future work will focus on enhancing the robustness of the verification process, for example, by incorporating external terminological resources to support boundary-case decisions and by developing ensemble verification strategies that combine multiple LLMs to mitigate stochastic errors. Additionally, extending the framework to handle nested term structures, as well as introducing structural pre-filtering mechanisms to exclude dosage expressions and sentence fragments, may further improve overall extraction quality.

Acknowledgements

We sincerely thank the reviewers for their insightful comments and constructive suggestions.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62506359).

References

1 Rigouts Terryn A, Hoste V, Drouin P, *et al*. TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset. In: *Proceedings of the 6th International Workshop on Computational Terminology*. Marseille, 2020, 85–94.

2 Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: The C-value/NC-value method. *Int J Digit Libr* 2000; **3**: 115–130.

3 Tran HTH, Martinc M, Pelicon A, *et al*. Ensembling transformers for cross-domain automatic term extraction. In: *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries*. Cham: Springer, 2022.

4 Tran HTH, Martinc M, Caporusso J, *et al*. The recent advances in automatic term extraction: A survey. arXiv: 2301.06767.

5 Wang XZ, Wei J, Schuurmans D, *et al*. Self-consistency improves chain of thought reasoning in language models. In: *Proceedings of the Eleventh International Conference on Learning Representations*. 2023, 2550–2575.

6 Weng YX, Zhu MJ, Xia F, *et al*. Large language models are better reasoners with self-verification. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, 2550–2575.

7 Rigouts Terryn A, Hoste V, Lefever E. In no uncertain terms: A dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Lang Resour Eval* 2020; **54**: 385–418.

8 Justeson JS, Katz SM. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Nat Lang Eng* 1995; **1**: 9–27.

9 Xu K, Feng YF, Li QD, *et al*. Survey on terminology extraction from texts. *J Big Data* 2025; **12**: 29.

10 Bourigault D, Jacquemin C. Term extraction + term clustering: An integrated platform for computer-aided terminology. In: *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*. Bergen, 1999, 15–22.

11 Zhang ZQ, Iria J, Brewster C, *et al*. A comparative evaluation of term recognition algorithms. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, 2008.

12 Conrado M, Pardo T, Rezende S. A machine learning approach to automatic term extraction using a rich feature set. In:

Proceedings of the 2013 NAACL HLT Student Research Workshop. Atlanta, 2013, 16–23.

13 Kucza M, Niehues J, Zenkel T, *et al*. Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks. In: *Proceedings of the 19th Annual Conference of the International Speech Communication*, *INTERSPEECH 2018*. 2018, 2072–2076.

14 Tran HTH, Martinc M, Repar A, *et al*. A transformer-based sequence-labeling approach to the slovenian cross-domain automatic term extraction. In: *Slovenian Conference on Language Technologies and Digital Humanities*. 2022, 196–204.

15 Lang C, Wachowiak L, Heinisch B, *et al*. Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021, 3607–3620.

16 Yuan SY, Yang DQ, Liu JX, *et al*. Causality-aware concept extraction based on knowledge-guided prompting. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, 2023, 9255–9272.

17 Aggarwal P, Madaan A, Yang YM, *et al*. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore, 2023, 12375–12396.

18 Raj H, Rosati D, Majumdar S. Measuring reliability of large language models through semantic consistency. arXiv: 2211.05853v2.

19 Chen B, Zhang F, Nguyen A, *et al*. CodeT: Code generation with generated tests. arXiv: 2207.10397.

20 Gero Z, Singh C, Cheng H, *et al*. Self-verification improves few-shot clinical information extraction. In: *Proceedings of the ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*. 2023.

21 Huang J, Chen XY, Mishra S, *et al*. Large language models cannot self-correct reasoning yet. arXiv: 2310.01798.

22 Stechly K, Valmeekam K, Kambhampati S. On the self-verification limitations of large language models on reasoning and planning tasks. In: *Proceedings of the Thirteenth International Conference on Learning Representations*. 2025, 98190–98243.

23 Shinn N, Cassano F, Gopinath A, *et al*. Reflexion: Language agents with verbal reinforcement learning. In: *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*. 2023, 8634–8652.

24 Zhang NY, Chen MS, Bi Z, *et al*. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, 2022, 7888–7915.

25 Yang A, Li AF, Yang BS, *et al*. Qwen3 technical report. arXiv: 2505.09388.