Research on Knowledge Extraction and Emerging Frontier Identification of Scientific Papers via Multi-LLM Integration: A Case Study of Artificial Intelligence

Hong Xu, Gang Li

2026-08-22

ChinaRxiv

Abstract

[Objective] To address the intelligence analysis challenges posed by the rapid growth in the number of scientific papers, this study explores the accuracy and reliability of multi-large language model (LLM) ensemble methods in knowledge extraction and frontier identification. [Methods] A five-stage research framework was constructed. Five LLMs, including GPT-4, were selected to perform entity and relation extraction on 2,000 AI papers. Two ensemble strategies, weighted voting and confidence fusion, were compared, and a multi-criteria evaluation framework (( )3/( )4/( )5 consensus) was used to construct reference standards —all F1 values should be understood as "agreement with model consensus" rather than absolute accuracy. [Results] GPT-4 achieved the best single-model entity F1 (0.855 ± 0.002). Under the ( )4 consensus standard, the ensemble entity F1 did not surpass the best single model; manual evaluation showed that this was partly because the recall of the reference standard was only 0.519. The ensemble achieved positive gains in relation extraction (F1 improvement of 15%–73%); weighted voting achieved 100% coverage of the reference standard, and among unique entities, 40% were true hallucinations while 60% were omissions by other models. A diagnostic stacking experiment obtained F1=0.944, but multi-criteria evaluation confirmed severe overfitting and thus it cannot serve as evidence of effectiveness. The knowledge graph contained 472 nodes and 15,144 edges; 29 emerging frontiers were identified according to the criteria of logarithmic growth rate ( )0.3 and recent frequency ( )5 (25–29 in five repeated experiments), and 6 of the Top 10 were directly or indirectly covered by the AI Index Report 2024. [Conclusion] The core contribution of this study is to reveal the decisive impact of evaluation benchmark construction on conclusions about ensemble evaluation and to confirm the unique value of ensembles in relation extraction and knowledge coverage.

Full Text

Research on Knowledge Extraction and Emerging Frontier Identification of Scientific Papers via Multi-LLM Integration: A Case Study of Artificial Intelligence

Hong Xu[1], Gang Li[2]

(1. Asset Management Office, Yunnan Open University, No. 318 Qixiu Street, Chenggong University Town, Kunming, Yunnan Province; 2. Digital Intelligence Center, Yunnan Open University, No. 318 Qixiu Street, Chenggong University Town, Kunming, Yunnan Province)

Abstract: **[Purpose]** In response to the challenges that the rapid growth in the number of scientific and technological papers poses for information anal-

ysis, this paper explores the accuracy and reliability of multi-large language model (LLM) integration methods in knowledge extraction and frontier identification. [**Method**] A five-stage research framework is constructed. Five LLMs, including GPT-4, are selected to extract entities and relations from 2,000 AI papers; two integration strategies, weighted voting and confidence fusion, are compared; and a multi-standard evaluation framework (( )3/( )4/( )5 consensus) is used to construct reference standards—all F1 values should be understood as "consistency with model consensus" rather than absolute accuracy. [**Result**] GPT-4 achieves the best single-model entity F1 (0.855 ± 0.002). Under the ( )4 consensus criterion, the integrated entity F1 does not exceed that of the best single model; manual evaluation shows that this is partly because the recall of the reference standard is only 0.519. Integration yields positive gains in relation extraction (F1 improvement of 15%–73%); weighted voting achieves 100% coverage of the reference standard, and among unique entities, 40% are hallucinations while 60% are omissions by other models. The diagnostic Stacking experiment obtains F1 = 0.944, but multi-standard evaluation confirms serious overfitting and cannot be used as evidence of validity. The knowledge graph contains 472 nodes and 15,144 edges; using criteria of log growth rate ≥ 0.3 and recent frequency ≥ 5, 29 emerging frontiers are identified (25–29 across five repeated experiments), of which 6 of the Top 10 are directly or indirectly covered by the AI Index Report 2024. [**Conclusion**] The core contribution of this paper is to reveal the decisive influence of evaluation benchmark construction on conclusions about integrated evaluation, and to confirm the unique value of integration in relation extraction and knowledge coverage.

Keywords: large language model; multi-model integration; knowledge extraction; knowledge graph; emerging frontier identification

CLC classification number: G350; TP18 **Document identification code:** A

Funding project: Research Fund Project of Yunnan Open University (Study on the Mechanism for Enhancing Multi-Brain Collaborative Reliability of Generative Large Models (25YNOU36))

Author profiles: Xu Hong (1997–), male, from Kunming, Yunnan Province, senior engineer; main research directions: AI application research and AI reliability research. E-mail: xuhong@ynou.edu.cn. Li Gang (1981–), male, from Kunming, Yunnan Province, doctoral student; main research directions: information management and information systems. E-mail: ligang@ynou.edu.cn.

Research on Knowledge Extraction and Emerging Frontier Identification of Scientific Papers via Multi-LLM Integration: A Case Study of Artificial Intelligence

Hong Xu[1], Gang Li[2]

1.Asset Management Office, Yunnan Open University, Kunming 650500, China

2.Digital Intelligence Center, Yunnan Open University, Kunming 650500, China

Abstract: [Purpose] This study explores multi-LLM integration to enhance knowledge extraction and frontier identification in scientific papers. [Method] A five-stage framework is applied to 2,000 AI papers using five LLMs and two integration strategies (weighted voting, confidence fusion); a multi-standard framework (( )3/( )4/( )5 consensus) constructs the reference standard, and all F1 values should be read as "consistency with model consensus" rather than absolute accuracy. [Result] GPT-4 achieves the best single-model entity F1

(0.855 ± 0.002). Under the ( )4 consensus standard, integration does not surpass the best single model on entity F1, partly because the reference standard' s recall is only 0.519; integration yields positive gains in relation extraction (F1 improvement of 15%–73%). Weighted voting covers 100% of the reference standard; among unique entities, 40% are true hallucinations and 60% are real entities missed by other models. A Stacking diagnostic experiment yields F1=0.944, but multi-standard evaluation confirms severe overfitting. The knowledge graph contains 472 nodes and 15,144 edges; under the criteria of log growth rate ( )0.3 and recent frequency ( )5, 29 emerging frontiers are identified (25–29 across five runs), and 6 of the Top 10 topics are directly or indirectly covered by the AI Index Report 2024. [Conclusion] The core contribution is revealing the decisive impact of evaluation benchmark construction on integration assessment and identifying integration' s unique value in relation extraction and knowledge coverage.

Key words: large language model; multi-model integration; knowledge extraction; knowledge graph; emerging frontier identification

1 Introduction

More than 5 million scientific and technological papers are published worldwide each year, and the number is increasing at an average annual rate of about 4% [1]. How to efficiently extract structured knowledge from massive literature and identify disciplinary frontiers has become a central issue in scientific and technological intelligence research.

Large language models (Large Language Model, LLM) have brought new opportunities for this. The new generation of LLMs represented by GPT-4 [2], Claude

[3], and Tongyi Qianwen [4] perform well in natural-language understanding, text generation, and knowledge reasoning, and have been used in information retrieval [5], intelligence analysis [6], and scientific and technological evaluation [7]. Compared with rule-based or shallow-learning approaches for named entity recognition (NER) and relation extraction (RE), LLMs have stronger semantic understanding and zero-shot/few-shot learning capabilities, enabling them to handle more complex extraction tasks from scientific and technological texts [8].

However, a single LLM has obvious limitations: first, the "hallucination" problem, namely generating extraction results that appear reasonable but are in fact incorrect [9]; second, model bias—different models have their own preferences and blind spots because of differences in training data, architecture, and optimization objectives, and extraction performance varies significantly across models and tasks [10]; third, insufficient domain adaptability, as general-purpose LLMs are not stable enough in extracting specialized terminology and complex relations in specific disciplines [11].

Multi-model integration (Ensemble) is one pathway for overcoming the above limitations. Ensemble-learning theory shows that reasonably combining the predictions of multiple base learners can reduce variance and bias and achieve more robust performance than a single model [12]. Existing studies have explored strategies such as voting, weighted fusion, and Stacking [13], but they have mostly focused on general NLP tasks. Systematic research aimed at scientific-paper knowledge extraction and intelligence-analysis scenarios remains lacking, and the following three issues have not yet been fully discussed: (1) patterns of performance differences and complementarity among different LLMs in extracting scientific and technological entities and relations; (2) comparison of the applicability of multiple integration strategies in scientific and technological text scenarios; (3) how to transform high-quality extraction results after integration into actionable signals for frontier identification.

Based on this, this paper proposes a multi-LLM integrated framework for knowledge extraction from scientific and technological papers and identification of emerging frontiers. Taking the field of artificial intelligence as an empirical case, it selects five representative models—GPT-4, Claude, Tongyi Qianwen, GLM-4, and DeepSeek—to conduct knowledge extraction on 2,000 papers, compares three integration strategies—weighted voting, confidence fusion, and Stacking—and constructs a knowledge graph based on the integrated results.

···spectrum, using community detection and logarithmic growth-rate methods to identify emerging fronts. The main contributions are threefold: (1) it reveals the decisive influence of how evaluation benchmarks are constructed on ensemble-evaluation conclusions—under the ≥ 4 consensus criterion, the ensemble' s entity F1 does not exceed that of the best single model, and the recall of some source reference criteria is only 0.519, indicating the need to guard against the methodological trap of "model consensus equals correctness" ; (2) it constructs a multidimensional evaluation framework covering multi-criterion

assessment ($\geq 3/ \geq 4/ \geq 5$), knowledge-coverage analysis, manual validation of unique entities, and manual evaluation of reference criteria, and finds the positive gains of ensembling in relation extraction (F1 improvements of 15%-73%); (3) it builds a knowledge graph based on the ensemble results and implements frontier identification, verifying the rationality of the identification results through structured comparison with external authoritative reports such as the AI Index Report.

It should be noted that while the five-model ensemble brings a marginal improvement in knowledge coverage (92.2%(→)93.3%), it is also accompanied by a fivefold computational cost. In practical applications, this trade-off must be weighed accordingly. If F1 is the objective, the single model GPT-4 already has an advantage; if maximizing coverage is the objective and the budget is sufficient, the ensemble has irreplaceable value.

2 Related Work

2.1 Applications of Large Language Models in Information Science

In literature analysis, the Transformer architecture proposed by Vaswani et al. [14] laid the technical foundation for modern LLMs. The derived BERT and GPT series of models have been widely used in automatic classification, abstract generation, and topic analysis of scientific and technological literature; GPT-3, proposed by Brown et al. [2], performs excellently in zero-shot text understanding and provides a new paradigm for automated processing of large-scale literature.

In information retrieval, the review by Zhu et al. [5] points out that LLMs show clear advantages in query understanding, document ranking, and conversational retrieval. In science and technology evaluation, Park et al. [7], based on quantitative analysis of tens of millions of papers and patents, revealed evolutionary patterns in the disruptiveness of scientific discoveries and also demonstrated the value of large-scale literature computational analysis.

Domestically, Li Yang and Sun Jianjun [6] analyzed the impact of large language models on the development of information science and possible response paths; Zhang Yiyi et al. [11] evaluated ChatGPT' s performance and usability in identifying entities in academic papers.

2.2 Knowledge Extraction Methods for Scientific and Technological Papers

Knowledge extraction from scientific and technological papers extracts structured knowledge units and their semantic relations from unstructured text. The core tasks are named entity recognition (NER) and relation extraction (RE).

Traditional methods rely on rule matching and statistical learning: the bidirectional LSTM-CRF of Lample et al. [15] performs well on sequence-labeling tasks;

DyGIE++ by Wadden et al. [16] uses a joint model to simultaneously extract entities, relations, and events; SciBERT [17] performs excellently on multiple scientific NER benchmarks and is often used as a reference baseline. However, such methods rely heavily on feature engineering and annotated data, and their generalization ability is limited in cross-domain and low-sample scenarios.

After the rise of pretrained language models, BERT-based NER/RE methods became mainstream: ChatIE by Wei et al. [8] verified the feasibility of using ChatGPT to complete zero-shot information extraction through conversational prompts; Wei et al. [18] found that instruction-tuned models generalize better in zero-shot information extraction. In recent years, LLMs such as GPT-4 have further promoted the transition of knowledge extraction toward a prompt-driven paradigm, reducing dependence on annotated data [19].

In terms of benchmark datasets, SciERC [20] annotated entities and relations in computer science literature, while SciREX [21] further covers coreference resolution, providing standardized references for method evaluation. Recently, works such as GraphRAG have combined entities and relations extracted by LLMs ···

organization as a layered community structure to support global analysis [22]. Domestic scholars have also carried out practice in large-model-assisted domain term extraction and graph construction [23].

2.3 Research on Multi-Model Ensemble Strategies

Ensemble learning improves overall performance by combining the predictions of multiple base learners. It is a classical method in machine learning [12] and has already been applied in NLP tasks such as text classification, sentiment analysis, and machine translation. For example, the self-consistency method proposed by Wang et al. [13] integrates multiple reasoning paths through majority voting, significantly improving generation quality.

In terms of multi-LLM ensembles, Zhou [24] and Zhang et al. [25] reviewed the theory and applications of ensemble learning; Chen et al. [26] systematically summarized the main paradigms of LLM ensembles, covering strategies such as output-level fusion, confidence-weighted fusion, and model selection; Jiang et al. [27] proposed LLM-Blender, which integrates the outputs of multiple LLMs through pairwise ranking and generative fusion; Pan et al. [28] provided a technical roadmap for integrating LLMs with knowledge graphs.

Domestically, Bao and Zhang Chengzhi [29] evaluated ChatGPT' s capabilities and boundaries in Chinese information extraction tasks such as named entity recognition and relation extraction. However, systematic comparative and ensemble research on multi-LLM systems for science and technology intelligence scenarios remains relatively lacking in China.

2.4 Research Review

Overall, existing research has three shortcomings: (1) applications of LLMs in intelligence studies mostly focus on a single model, lacking systematic comparison and ensemble research across multiple models; (2) research on knowledge extraction mostly focuses on extraction methods themselves, with relatively little discussion of the impact of multi-model fusion on extraction quality, and also lacks direct comparison with traditional baselines; (3) frontier identification based on knowledge graphs has not yet fully utilized the quality improvements brought by multi-LLM ensembles, and the identification results often lack comparative validation.

To address the above shortcomings, this paper takes the field of artificial intelligence as an empirical case and systematically studies the application of multi-LLM ensembles in knowledge extraction and frontier identification. It introduces traditional rule-based baselines as references and adopts comparative methods to validate frontier identification results. In terms of emerging-topic measurement, existing research has proposed a dynamic measurement framework for topic growth from knowledge emergence to diffusion—"novelty"–"growth" [30]. However, the impact of multi-model extraction quality on frontier identification results has not yet been discussed; this paper is precisely a response to this gap.

3 Research Framework and Methods

3.1 Overall Research Framework

This paper constructs a five-stage research framework, as shown in Figure 1.

Research framework

Figure 1. Research framework of multi-LLM integrated knowledge extraction and frontier identification for scientific papers

The five stages are: (1) data preprocessing: text cleaning, segmentation, and standardization; (2) multi-LLM knowledge extraction: each model separately extracts entities and relations; (3) ensemble fusion: extraction results are fused using three strategies—weighted voting, confidence fusion, and stacking; (4) knowledge graph construction: a graph is constructed based on the ensemble results and community analysis is performed; (5) frontier trend identification: emerging frontiers are identified through time slicing and growth-rate analysis, with the frequency-threshold method used as a comparison. The output of the preceding stage serves as the input of the following stage.

3.2 Multi-LLM Knowledge Extraction Module

This paper defines four types of scientific and technological entities and six types of semantic relations, as follows:

Entity types: (1) Method: algorithms, models, or techniques proposed in papers, such as "Transformer" and "BERT" ; (2) Task: the specific problems or application scenarios addressed by the research, such as "text classification" and "image recognition" ; (3) Dataset: data resources used in experiments, such as "ImageNet" and "COCO" ; (4) Metric: quantitative criteria for evaluating model performance, such as "Accuracy" and "F1 Score."

Relation types: (1) Applied-to: a method is applied to a certain task, such as "BERT is applied to text classification" ; (2) Based-on: a method is constructed or improved based on another method, such as "GPT-3 is based on Transformer"; (3) Evaluated-on: a method is evaluated on a specific dataset or metric, such as "BERT is evaluated on GLUE"; (4) Improves: one method improves performance or architecture on the basis of another method, such as "RoBERTa improves BERT" ; (5) Extends: one method extends the application scope of another method to a new scenario, such as "ViT extends Transformer to visual tasks" ; (6) Combines: two methods are fused and used to solve a specific problem, such as "CLIP combines contrastive learning with vision-language pretraining."

The prompt uses a structured template, uniformly including role setting, task description, definitions of entity and relation types, and output-format requirements, and is adapted according to model characteristics: GPT-4 and Claude use English prompts, Tongyi Qianwen, GLM-4, and DeepSeek use Chinese-English bilingual prompts.

3.3 Multi-Model Ensemble Strategies

3.3.1 Weighted Voting Strategy The weighted voting strategy assigns weights according to the performance of each model on the validation set, and determines the final extraction result by the majority-voting principle.

For entity extraction, suppose the voting weight of the k-th model for entity e is w_k. Then the aggregate score of entity e is:

$$S(e) = \sum_{k=1}^{K} w_k \cdot v_k(e)$$

where $v_k(e) \in \{0, 1\}$ indicates whether the k-th model extracts entity e, and w_k is obtained by normalizing the F1 value on the validation set. When $S(e)$ exceeds the threshold θ, the entity is included in the result. In this paper, $\theta = 0.5 \times \sum w_k$ (exceeding half of the total weighted votes). This threshold is set empirically: a higher θ improves precision but reduces recall, and vice versa; under this setting, the data in this paper achieve a good precision–recall balance.

3.3.2 Confidence Fusion Strategy The confidence fusion strategy uses the confidence scores output by each model for probability-weighted fusion. Suppose

the confidence of the k-th model for extraction result r is $c_k(r)$. Then the fused confidence is:

$$C(r) = \frac{\sum_{k=1}^{K} c_k(r) \cdot \delta_k(r)}{\sum_{k=1}^{K} \delta_k(r)}$$

where $\delta_k(r)$ is an indicator function, taking the value 1 when the k-th model extracts result r, and 0 otherwise. The final decision uses threshold filtering (fusion score ≥ 0.62 and support from at least 3 models). The threshold is determined by grid search on the validation set, and its sensitivity remains to be further explored in subsequent research.

How confidence is obtained. The confidence of each model is obtained by prompting the model to self-report in JSON format (e.g., "confidence: 0.85"), with values in the range $[0, 1]$. API logprobs are not used for two reasons: (1) different model APIs have inconsistent support for logprobs (some domestic models have not opened this interface), and self-reporting ensures comparability; (2) self-reported confidence reflects the model' s overall judgment of the extraction result, rather than token-level probabilities. It should be noted that LLM self-reported confidence has a known overconfidence bias [31], and the confidence scores of different models have not been systematically calibrated—GPT-4' s 0.8 and Tongyi Qianwen' s 0.8 are not strictly equivalent in a statistical sense. This may be one reason why confidence fusion (F1=0.779) is not significantly better than weighted voting (F1=0.771). In the future, fusion could be performed after calibration through temperature scaling or Platt scaling.

3.3.3 Stacking Ensemble Strategy Stacking adopts a two-layer architecture: the first layer uses the extraction results of 5 base LLMs as features, and the second layer makes an integrated decision using a meta-learner. To avoid data leakage, at the paper level, the 2000 papers are split 60/20/20 into a training set (1200 papers), validation set (400 papers), and test set (400 papers); the same paper does not appear simultaneously in both the training and test sets.

Using the F1 scores of each base model on the validation set as the basis, a leave-one-out method is used to compute the degree of matching in voting consensus between each model and the other four models, learn the optimal combination weights, and in the prediction stage perform a weighted summation of confidence scores, followed by threshold decision through the sigmoid function:

$$P(r|X) = \sigma\left(\sum_{k=1}^{K} w_k \cdot c_k(r)\right)$$

where w_k is the weight learned by the meta-learner, $c_k(r)$ is the confidence of model k, and σ is the sigmoid function. Stacking can adaptively learn the

strengths and weaknesses of each model across different entity and relation types.

3.4 Knowledge Graph Construction and Frontier Identification

Based on the integrated structured knowledge, a knowledge graph $G = (V, E)$ is constructed, where V is the set of entity nodes and E is the set of semantic-relation edges. During construction, synonymous entities are normalized, and entity occurrence frequency and relation strength are counted.

Community detection adopts the Louvain algorithm [32], identifying thematic community structures by optimizing modularity Q:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where A_{ij} is an element of the adjacency matrix, k_i is the degree of node i, m is the total number of edges, and c_i is the community to which node i belongs.

Frontier identification adopts logarithmic growth-rate analysis based on time slices: papers are divided into time windows by publication year, and the average occurrence frequencies of each entity in the early and recent windows are calculated. To avoid excessive sensitivity of traditional growth rates to low-frequency entities, the logarithmic growth rate is used:

$$GR = \ln(f_{\text{recent}} + 1) - \ln(f_{\text{early}} + 1)$$

where f_{early} and f_{recent} are respectively the average frequencies of an entity in the early and recent windows. Emerging frontier candidates must satisfy: (1) the logarithmic growth rate exceeds the threshold (set to 0.3 in this paper); (2) cumulative frequency in the recent window ≥ 5; (3) only method-type and task-type entities are considered, while metric-type entities are analyzed separately and are not treated as candidates.

To verify robustness, a frequency-threshold method is also introduced as a comparison: method/task entities with recent frequency ≥ 5 are taken as frontier candidates, ranked in descending order by frequency, and the reliability is evaluated through the consistency of the results of the two methods (intersection ratio).

4 Experimental Design

4.1 Dataset Construction

This paper constructs a dataset of 2,000 AI-domain papers through arXiv API retrieval, with a time span from 2019 to 2025, covering six subfields: natural language processing (about 340 papers), computer vision (about 340 papers),

multimodal learning (about 330 papers), graph neural networks (about 330 papers), reinforcement learning (about 330 papers), and large language models (about 330 papers). The selection strategy is: (1) search by subfield category; (2) prioritize top conferences and journal papers such as NeurIPS, ICML, ACL, EMNLP, and CVPR (identified via the arXiv comment field); (3) exclude review articles and Position Papers. The distributions across years and subfields are relatively balanced, and all metadata (title, abstract, arXiv ID, publication year, subfield) are stored in structured format, supporting result tracing and reproduction.

Annotation process and construction of evaluation benchmarks. In view of the lack of an existing gold standard, this paper adopts a systematic method for constructing reference standards: (1) two researchers with computer-science backgrounds independently compiled a dictionary of scientific and technological entities, covering four categories—methods, tasks, datasets, and metrics—with about 1,200 entries in total (approximately 450 methods, 300 tasks, 250 datasets, and 200 metrics), referring to the annotation specifications of ACL Anthology, Papers with Code, and SciERC; the overlap rate of the two annotators' terminology was about 82%, and disagreements were resolved through consultation; (2) five LLMs independently performed extraction under exactly the same prompts, without manual intervention or dictionary constraints; (3) the reference standard was constructed by supermajority voting: an entity/relation was included only if it was extracted by at least 4 models (80%) simultaneously. This standard is stricter than a simple majority (≥ 3), ensuring high confidence while preserving discriminative space—entities supported by only 3 models were counted as false positives, so that differences in the precision of different methods could be produced; (4) the relation reference standard required at least 4 models to agree on the same triple (source entity, target entity, relation type).

It should be noted that this construction method has the methodological limitation of circular validation: the reference standard is jointly constructed by 5 LLMs, while the evaluation objects are also these 5 LLMs and their ensemble. If the models share a certain systematic bias (for example, all identify "self-supervised learning" as a method rather than a task), this bias will be solidified by the consensus mechanism as the "correct answer." Therefore, the F1 value in this paper should be understood as "consistency with model consensus" rather than absolute accuracy. To mitigate this issue, this paper conducted multi-standard evaluation (see Section 5.2) and independent human evaluation of 100 papers; a larger-scale human annotation benchmark (such as complete entity and relation annotation for 200–300 papers) remains an urgent need for future research.

Reasons for choosing a ≥ 4 consensus threshold and its challenges. The main considerations for choosing ≥ 4 rather than ≥ 3 are: (1) higher confidence, reducing contamination of the benchmark by low-quality extractions; (2) under a 5-model configuration, ≥ 3 means "3 votes suffice to pass," which may include too many marginal consensus entities; (3) entities that reach ≥ 3 but not ≥ 4

are retained as discriminative space. However, the human evaluation in Table 8 poses a challenge to this: from ≥ 3 to ≥ 4, precision increases by only 0.9 percentage points (0.600 $\rightarrow$ 0.609), while recall decreases by 28.7 percentage points (0.806 $\rightarrow$ 0.519), and F1 drops from 0.688 to 0.560; that is, after the ≥ 4 threshold, a large recall loss is exchanged for a very small precision gain. In view of this, this paper reports F1 under both the ≥ 3 and ≥ 4 standards (see Table 5), allowing readers to judge the robustness of the conclusions for themselves.

Independent human evaluation. To verify the rationality of the reference standard, 100 papers (5% of the total sample) were randomly selected, and a third-party researcher who did not participate in the model experiments (with more than 3 years of AI research experience) evaluated the extraction results from two dimensions: entity boundaries and type judgments. The agreement between the reference-standard entity boundaries and human judgments was 84.3%, and the agreement for type judgments was 89.7%, indicating overall reliability; inconsistent cases were mainly concentrated in the handling of abbreviations (such as "ViT" and "Vision Transformer") and compound terms (such as marking "self-supervised pre-training" as a whole versus marking it separately).

4.2 Model Selection and Experimental Setup

This paper selects 5 representative LLMs: GPT-4 (gpt-4-0613), Claude (Claude 3 Opus), Tongyi Qianwen (Qwen-Max[4]), GLM-4 (GLM-4 Standard Edition[33]), and DeepSeek (DeepSeek-V2[34]), taking into account the diversity and availability of international commercial, domestic, and emerging open-source models. Version numbers were fixed to ensure reproducibility, and the temperature parameter was uniformly set to 0.1. All experiments were independently repeated 5 times (random seeds 42, 123, 456, 789, and 2024) for dataset splitting, and the reported results are means $\pm$ standard deviations. The experimental environment was Ubuntu 22.04 and Python 3.10; models were invoked through official APIs, with a maximum output of 4096 tokens; prompts, parameters, and random seeds were all managed with version control.

traditional baseline methods. To evaluate the improvement of LLMs over traditional methods, a rule-matching-based NER/RE baseline (RuleBaseline) was introduced, using dictionary matching and simple co-occurrence rules, without relying on any machine-learning model; it represents the baseline level achievable by traditional intelligence-analysis methods.

4.3 Evaluation Metrics

This paper uses precision, recall, and F1 value as the evaluation metrics for knowledge extraction performance, defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP is the number of correct extractions (the prediction result is consistent with the reference standard), FP is the number of erroneous extractions (the prediction result is not in the reference standard), and FN is the number of missed extractions (entities/relations in the reference standard that were not predicted).

Inter-model consistency analysis adopts Cohen's Kappa coefficient, with a value range of $[-1, 1]$. It is generally considered that $\kappa > 0.6$ indicates "high consistency," and $0.4 < \kappa \leq 0.6$ indicates "moderate consistency."

Statistical significance testing uses the paired t-test, with significance level $\alpha = 0.05$, to determine whether the difference between the ensemble strategy and the best single model is statistically significant.

5 Experimental Results and Analysis

5.1 Comparison of Single-Model Knowledge Extraction Performance

5.1.1 Analysis of Entity Extraction Results

The comparison of the performance of five LLMs and the traditional rule baseline on entity extraction is shown in Table 1. All results are the mean ± standard deviation of five independent repeated experiments.

Table 1 Comparison of entity extraction performance across models (mean±std)

Method	Precision	Recall	F1 Value	F1 Standard Deviation
RuleBaseline (Traditional baseline)	0.669 ± 0.003	0.398 ± 0.002	0.474 ± 0.002	0.242 ± 0.001
GPT-4	0.817 ± 0.003	0.920 ± 0.002	0.855 ± 0.002	0.124 ± 0.001
Claude	0.591 ± 0.002	0.940 ± 0.003	0.713 ± 0.002	0.132 ± 0.001
Tongyi Qianwen	0.610 ± 0.001	0.905 ± 0.002	0.716 ± 0.001	0.138 ± 0.001
GLM-4	0.883 ± 0.003	0.750 ± 0.002	0.795 ± 0.002	0.162 ± 0.001
DeepSeek	0.603 ± 0.002	0.917 ± 0.004	0.715 ± 0.002	0.137 ± 0.001

Overall, GPT-4 ranks first in entity F1 with 0.855, followed by GLM-4 (0.795), Tongyi Qianwen (0.716), Claude (0.713), and DeepSeek (0.715). All LLMs significantly outperform the rule baseline (F1=0.474), with improvements ranging from 50.4% to 80.4%.

The models show distinct precision–recall trade-offs: GPT-4 (precision 0.817, recall 0.920) achieves the best balance; GLM-4 has the highest precision (0.883) but lower recall (0.750), indicating a conservative strategy; Claude, Tongyi Qianwen, and DeepSeek all have relatively high recall.

with recall exceeding 0.90 but precision only 0.59–0.61, tending to output more candidates and thus yielding a higher false-positive rate. It can be seen that GPT-4 and GLM-4 favor "conservative" extraction, while the other three models favor "aggressive" extraction.

The F1 values of each model on different entity types are shown in Table 2.

Table 2 F1 scores across models and entity types (mean of 5 runs) Tab.2 F1 scores across models and entity types (mean of 5 runs)

Model	Method	Task	Dataset	Metric
RuleBaseline	0.512	0.461	0.398	0.312
GPT-4	0.892	0.851	0.823	0.756
Claude	0.745	0.698	0.671	0.589
Tongyi Qianwen	0.751	0.702	0.678	0.598
GLM-4	0.834	0.778	0.745	0.682
DeepSeek	0.748	0.699	0.674	0.592

Extraction performance is generally best for method-type entities (F1 of 0.745–0.892), because their naming patterns are fixed and their textual salience is high; metric-type entities are the most difficult (0.589–0.756), because their expressions are diverse (e.g., "Accuracy," "F1," "P@5") and are often mixed with numerical values.

5.1.2 Analysis of Relation Extraction Results

The performance comparison of each model on the relation extraction task is shown in Table 3.

Table 3 Comparison of relation extraction performance across models (mean±std) Tab.3 Comparison of relation extraction performance across models (mean±std)

Method	Precision	Recall	F1 value
RuleBaseline	0.145 ± 0.002	0.156 ± 0.003	0.132 ± 0.001
GPT-4	0.303 ± 0.005	0.642 ± 0.013	0.386 ± 0.007
Claude	0.282 ± 0.005	0.654 ± 0.008	0.369 ± 0.006
Tongyi Qianwen	0.357 ± 0.009	0.616 ± 0.011	0.425 ± 0.009
GLM-4	0.289 ± 0.004	0.268 ± 0.006	0.257 ± 0.005
DeepSeek	0.304 ± 0.005	0.641 ± 0.010	0.387 ± 0.007

Method	Precision	Recall	F1 value

Relation extraction is significantly more difficult than entity extraction: relation F1 ranges from 0.257 to 0.425, far lower than the entity F1 of 0.713–0.855. Tongyi Qianwen performs best (F1=0.425), benefiting from relatively high relation-type precision (0.357); GLM-4 performs worst (F1=0.257), with recall of only 0.268, indicating that it is overly conservative in relation identification.

All models have relatively low recall on "based on"-type relations, because such relations require understanding methodological inheritance and involve deeper semantic reasoning; surface relations such as "applied to" and "evaluated on" are easier to identify.

5.2 Analysis of Multi-Model Integration Effects

The performance comparison between the three integration strategies and single models is shown in Table 4. All results are the mean ± standard deviation over 5 independent repeated experiments.

Table 4 Comparison of integration strategies and single-model performance (mean±std) Tab.4 Comparison of integration strategies and single-model performance (mean±std)

Method	Entity Precision	Entity Recall	Entity F1	Relation Precision	Relation Recall	Relation F1
GPT-4 (optimal single model)	0.817 ± 0.003	0.920 ± 0.002	0.855 ± 0.002	0.303 ± 0.005	0.642 ± 0.013	0.386 ± 0.007
Weighted voting	0.645 ± 0.002	0.999 ± 0.001	0.771 ± 0.001	0.352 ± 0.006	0.685 ± 0.010	0.445 ± 0.007
Confidence fusion	0.691 ± 0.002	0.927 ± 0.003	0.779 ± 0.002	0.360 ± 0.005	0.656 ± 0.008	0.444 ± 0.005
Stacking (diagnostic experiment)	0.988 ± 0.001	0.913 ± 0.002	0.944 ± 0.001	0.678 ± 0.009	0.661 ± 0.009	0.668 ± 0.009

Note: The Stacking row has a serious risk of overfitting (see Section 5.2.4); its F1 value cannot be used as evidence of validity, and is used only for overfitting

diagnostic analysis. The main ensemble strategies in this paper are weighted voting and confidence fusion.

5.2.1 Precision–Recall Trade-off of Ensemble Strategies

The two main ensemble strategies exhibit different trade-off characteristics. Weighted voting achieves nearly perfect entity recall (0.999 ± 0.001), but its precision is relatively low (0.645 ± 0.002), and its entity F1 is 0.771 ± 0.001—majority voting recalls almost all reference-standard entities, but also introduces many low-consensus entities supported by only three models (treated as false positives in the reference standard). Confidence fusion has an entity F1 of 0.779 ± 0.002, not significantly different from weighted voting (t=-1.23, p=0.218). Stacking is included as a diagnostic experiment for analysis; under strict data partitioning, its entity F1 reaches 0.944 ± 0.001 and its precision is extremely high (0.988 ± 0.001), but multi-criterion evaluation (Section 5.2.4) confirms severe overfitting. Therefore, it is used only as a counterexample for methodological reflection and is not recommended as a strategy.

In contrast to the "negative gain" in entity extraction, the ensemble achieves positive gains in relation extraction. GPT-4' s relation F1 is 0.386; weighted voting (0.445) and confidence fusion (0.444) improve it by 15.3% and 15.1%, respectively (Stacking improves it by 73.1%, but cannot serve as evidence because of overfitting). A possible explanation is that a relation triple requires three elements to be simultaneously consistent in order to reach consensus; its consensus difficulty is much higher than that of a single entity. The number of relations in the reference standard is relatively small, so the probability that a high-recall strategy is penalized is correspondingly lower. Meanwhile, differences among models in judging relation types (e.g., GPT-4 favors the "based on" relation, while Claude favors the "applied to" relation) provide greater complementary space for the ensemble. This positive gain has a direct impact on downstream tasks: confidence fusion raises relation F1 from 0.386 to 0.444, meaning that approximately 15% more relation triples in the graph are correctly identified, which is beneficial for community discovery and frontier identification.

5.2.2 The Ensemble "Negative Gain" Phenomenon and Its Analysis

The key finding is that the entity F1 of weighted voting (F1=0.771) and confidence fusion (F1=0.779) is lower than that of the best single model, GPT-4 (0.855), meaning that "negative gain" occurs.

The root cause lies in the interaction between the construction of the reference standard and the ensemble strategy. The reference standard requires consensus among ( )4 models, whereas GPT-4 itself already has a recall of 0.920; some of the entities it identifies are supported by only three models and are therefore treated as false positives in the reference standard. When ensemble strategies (especially weighted voting) recall these low-consensus entities, they are instead penalized in terms of precision—high recall (0.999) leads to a decline in precision

under the strict standard.

The manual evaluation in Table 8 reveals a deeper cause: the ≥ 4 reference standard itself is incomplete, with a recall of only 0.519 against the manually constructed gold standard; that is, about 48% of true entities were missed because they did not receive consensus from 4 models. The F1 calculated on the basis of this standard is in fact obtained on an incomplete benchmark—even if a method perfectly identified all true entities, its upper bound on recall would be only 0.519; moreover, among the large number of "outside-the-reference-standard" entities recalled by high-recall strategies, a considerable portion truly exist (see Section 5.2.5), yet are counted as false positives. Therefore, the "negative gain" cannot be attributed solely to the ensemble recalling low-quality entities; it should also be partly attributed to the incompleteness of the reference standard, which systematically biases the evaluation framework toward conservative models (such as GPT-4).

This finding poses a challenge to the core argument: if an ensemble is inferior to a single model on F1, its value needs to be re-examined from other dimensions. The following subsections reposition the contribution of the ensemble from three perspectives: multi-standard evaluation, knowledge coverage analysis, and filtering of low-consensus entities.

5.2.3 Multi-standard evaluation: sensitivity analysis of the reference standard

To evaluate the impact of the strictness of the reference standard, this paper constructs reference standards under three consensus thresholds, ≥ 3, ≥ 4, and ≥ 5, and evaluates each method separately. The results are shown in Table 5.

Table 5 Multi-standard evaluation results (entity F1)

Method	≥ 3 consensus	≥ 4 consensus	≥ 5 consensus	F1 variation range
GPT-4	0.810	0.856	0.522	0.334
Claude	0.903	0.714	0.406	0.497
Tongyi Qianwen	0.887	0.714	0.429	0.458
GLM-4	0.681	0.794	0.637	0.157
DeepSeek	0.893	0.712	0.423	0.470
Weighted voting	1.000	0.772	0.414	0.586
Confidence fusion	0.924	0.780	0.422	0.502
Stacking	0.730	0.942	0.600	0.240

Note: The F1=1.000 of weighted voting under the ≥ 3 criterion is a tautology—the threshold setting of weighted voting ($\theta = 0.5 \times \sum w_k$, i.e., exceeding half of the

weights) is equivalent, under five approximately equal-weight models, to support from ≥ 3 *models, and completely coincides with the rule used to construct the* ≥ 3 *reference standard. Therefore, F1=1.000 has no independent evaluation significance. The focus of analysis should be placed on the differences in results under the* ≥ 4 *and* ≥ 5 *criteria.*

Multi-standard evaluation reveals the following important findings:

First, F1 values are highly sensitive to the strictness of the reference standard. From ≥ 3 to ≥ 5, the F1 of all methods fluctuates substantially: weighted voting decreases from 1.000 to 0.414, while GPT-4 decreases from 0.810 to 0.522. This demonstrates that the absolute F1 value should not be directly interpreted as "accuracy."

Second, the ranking of methods changes significantly with the standard. Under the ≥ 3 criterion, weighted voting (1.000) and confidence fusion (0.924) far exceed GPT-4 (0.810); under the ≥ 4 criterion, Stacking (0.942) leads, but weighted voting (0.772) is lower than GPT-4 (0.856); under the ≥ 5 criterion, GLM-4 (0.637) instead ranks first among single models. The instability of the rankings indicates that comparison under a single standard makes it difficult to draw robust conclusions.

5.2.4 Diagnosis of Stacking Overfitting

The abnormally high F1 (0.942) of Stacking under the ( )4 criterion requires dedicated analysis: its meta-learner is trained using the outputs of five models as features, while the reference standard is also constructed based on the consensus of these five models. The meta-learner may therefore be learning "how to predict model consensus" rather than "how to correctly extract entities."

Analysis of the meta-learner weights. The weights assigned by the meta-learner to the models are: Claude (0.221), DeepSeek (0.216), Tongyi Qianwen (0.215), GPT-4 (0.188), and GLM-4 (0.160)—models with high recall but low precision (all with recall above 0.90) receive high weights, whereas GLM-4 (0.883), which has the highest precision, receives the lowest weight. This distribution suggests that what the meta-learner has learned is to "maximize entities in the recall-model consensus" : outputs from high-recall models are more likely to overlap with the ( )4 consensus.

Cross-validation diagnosis. Table 5 shows that Stacking achieves an F1 of only 0.730 under the ( )3 criterion, which is 21.2 percentage points lower than its 0.942 under the ( )4 criterion. This strongly indicates that the meta-learner has learned the construction rule of the ( )4 consensus criterion rather than a general entity-recognition capability—if what it learned were genuine recognition ability, F1 should not fluctuate drastically with the criterion. In comparison, although GPT-4 has a relatively large F1 variation range (0.334), its pattern is smooth and does not show "abnormally high performance under a specific criterion."

The precision of Stacking also changes synchronously with the consensus thresh-

old: 0.998 under ( )3, 0.989 under ( )4, and then plunges to 0.466 under ( )5. This shows that its high precision under the ( )4 criterion is not an improvement in recognition ability, but rather that it has learned the decision rule of "retaining only high-consensus entities."

5.2.5 Multidimensional Evaluation of Ensemble Value

Given that F1 is substantially affected by the reference criterion and exhibits "negative gain," this section re-evaluates the value of ensembles from two dimensions—knowledge coverage and low-consensus entity filtering—and repositions them in conjunction with the manual validation in Section 5.2.6.

(1) Knowledge coverage analysis Coverage is defined as the proportion of reference-standard entities covered by each method' s predictions (i.e., recall). The results are shown in Table 6.

Table 6 Knowledge coverage of each method (( )4 reference standard)
Tab.6 Knowledge coverage of each method (( )4 reference standard)

Method	Total predicted entities	Covered reference-standard entities	Total reference-standard entities	Coverage
GPT-4	14,678	12,077	13,097	92.2%
Claude	20,586	12,340	13,097	94.2%
Tongyi Qianwen	19,176	11,850	13,097	90.5%
GLM-4	11,039	9,820	13,097	75.0%
DeepSeek	19,612	11,958	13,097	91.3%
Weighted voting	20,136	13,097	13,097	**100.0%**
Confidence fusion	17,589	12,223	13,097	93.3%
Stacking	12,085	11,985	13,097	91.5%

Weighted voting achieves 100% coverage of the reference standard, whereas the best single model, GPT-4, reaches 92.2%, still missing about 1,020 reference-standard entities. For science and technology intelligence analysis, whose primary goal is to "identify knowledge as comprehensively as possible,"improvement in coverage has greater practical significance than improvement in F1.

However, the high-coverage strategy carries a risk of noise propagation. The weighted-voting precision is only 0.645, and about 35% of predicted entities are erroneous outputs. If used directly for graph construction, it would generate a large number of spurious nodes and edges, interfering with community detection

and frontier identification. Therefore, this paper ultimately selects confidence fusion (precision 0.691, recall 0.927) as the graph data source, achieving a balance between coverage and noise control (see Section 5.5).

(2) Analysis of filtering low-consensus entities

To evaluate the filtering effect of integration on low-consensus entities, this paper defines "unique entities" as entities recognized by only one model and not recognized by the other four models, and analyzes their actual correctness under manual evaluation. The results are shown in Table 7.

Table 7 Human evaluation of unique entities across models

Model	Number of unique entities	Manually judged correct	Manual accuracy	Number of hallucinations	Hallucination rate
GPT-4	107	64	59.8%	43	40.2%
Claude	153	93	60.8%	60	39.2%
Tongyi Qianwen	142	87	61.3%	55	38.7%
GLM-4	73	42	57.5%	31	42.5%
DeepSeek	146	90	61.6%	56	38.4%

Note: Unique entities were extracted from a subset of 100 manually evaluated papers (i.e., the same 100 papers used for the manual evaluation in Table 8) and judged one by one by one evaluator. "Manually judged correct" means that the entity truly exists in the original text (the boundary is correct) and that the entity type is correctly judged (i.e., both boundary and type are correct). "Hallucination" means that the entity does not exist in the original text or that the type judgment is wrong. The distribution of unique entities across the 100 papers is sufficient for each model (73–153 entities), enough to support statistical conclusions.

It should be noted that in Section 5.2.3, the "hallucination rate" of single-model unique entities under the ( )4-reference criterion is 100%, but this is a tautological consequence of the reference-standard definition: the ( )4 criterion requires at least four models to recognize an entity jointly, whereas a unique entity is recognized by only one model. The two are logically incompatible. This "100% hallucination rate" is not an empirical finding.

To distinguish "hallucinations" (generated nonexistent entities) from "missed detections" (true entities missed by other models), this paper uses the 100 manually evaluated data items to independently verify unique entities. The results show that their manual accuracy is 57.5%–61.6%; that is, about 60% of unique entities are correct entities that truly exist, while hallucinations account for only 38.4%–42.5%.

This finding has methodological implications: (1) the ( )4-reference criterion indeed involves missed detection—about 48% of correct entities are omitted because they do not obtain consensus (recall only 0.519; see Table 8); (2) while the consensus mechanism filters out about 40% of hallucinations, it also incorrectly excludes about 60% of true entities; (3) the quality differences among the unique entities of the various models are not large (57.5%–61.6%), indicating that unique entities arise more from differences in extraction strategies among models than from any particular model being especially prone to hallucination.

(3) Repositioning the value of integration

Synthesizing the analyses in Sections 5.2.5(1)–(2) and 5.2.6, this paper repositions the core value of multi-model integration. The value of integration should not be measured only by the F1 metric, but should be understood from the following three dimensions:

(1) Maximization of knowledge coverage: weighted voting achieves 100% coverage for the ≥ 4 reference standard, outperforming any single model (maximum 94.2%); however, its precision is only 0.645, with about 35% of outputs being erroneous. Therefore, this paper selects confidence fusion (precision 0.691, recall 0.927) to construct the graph spectrum, balancing coverage and precision.

(2) Filtering of low-consensus entities: the consensus mechanism filtered out about 40% of hallucinations, but also mistakenly excluded about 60% of true entities. This "double-edged sword" effect suggests that consensus thresholds should be adjusted according to the scenario—use ≥ 4 in high-precision scenarios and ≥ 3 in high-recall scenarios.

(3) Tunability of precision–recall: the three strategies provide different trade-off options (weighted voting favors recall, confidence fusion seeks balance, and Stacking favors precision), allowing intelligence analysts to choose as needed.

The high F1 (0.944) of Stacking under the ≥ 4 criterion should be interpreted cautiously after diagnosis. Its advantage may partly stem from overfitting to the rules used to construct the reference standard; in the future, its generalization ability should be verified on an independent manually annotated benchmark.

Indirect comparison with public benchmarks. The evaluation setting in this paper differs from public benchmarks such as SciERC, but can serve as a reference. On SciERC, the best joint extraction models (such as DyGIE++[16]) achieve an entity F1 of about 0.68 and a relation F1 of about 0.48[20]. The entity F1 of the LLM single models in this paper (0.713–0.855) is overall higher than that level; however, because SciERC has finer annotation granularity, more types, and uses a manually curated gold standard, numerical comparisons should be made with caution. SciBERT-CRF typically reaches 0.65–0.75[17] on scientific NER tasks; the rule-based baseline in this paper (0.474) represents only the lower bound of "no machine learning."

5.2.6 Manual Evaluation Validation of the Reference Standard

To verify the validity of the ≥ 4 reference standard itself, 100 manually evaluated data items were used as an independent gold standard. The precision, recall, and F1 of the reference standards were calculated (Table 8), which also constitutes an independent test of the reliability of the entire evaluation framework.

Table 8 Comparison of reference standards against human gold standard

Reference standard	Total entities	Correct by human judgment	Precision	Covered human standard	Recall	F1
≥ 3 consensus	145	87	0.600	87	0.806	0.688
≥ 4 consensus	92	56	0.609	56	0.519	0.560

The manual evaluation of the reference standards reveals the following key findings:

First, the precision of the ≥ 4 reference standard is about 0.609. That is, about 60.9% of entities were judged completely correct by humans (both boundaries and types correct), with 93.2% boundary correctness and 61.4% type correctness. Although majority voting did not eliminate errors, it filtered out a large number of low-quality extractions.

Second, the recall of the ≥ 4 reference standard is only 0.519. That is, about 48.1% of correct entities in the human standard were missed because they did not obtain ≥ 4 consensus, confirming the finding in Section 5.2.5(2). The recall of the ≥ 3 standard increases to 0.806, while precision decreases slightly to 0.600.

Third, the strictness of the reference standard faces a precision–recall trade-off. From ≥ 3 to ≥ 4, precision increases by only 0.9 percentage points ($0.600 \rightarrow 0.609$), recall decreases by 28.7 percentage points ($0.806 \rightarrow 0.519$), and F1 drops from 0.688 to 0.560. Under the 5-model configuration, the ≥ 4 threshold may be overly strict; ≥ 3 performs better in F1 but contains more noise.

This manual evaluation involved only one evaluator; the dimensions were limited to entity boundaries and types and did not cover relations. Statistical power is limited (100 articles, 1 person, 95% confidence interval about ($\pm$)7%). In the future, more evaluators should be added to calculate inter-rater consistency, and the evaluation should be extended to relation extraction.

Model inter-consistency analysis (original string-matching criterion): Cohen' s Kappa consistency coefficient and Jaccard similarity heatmaps.

Figure 1: Model inter-consistency analysis (original string-matching criterion): Cohen' s Kappa consistency coefficient and Jaccard similarity heatmaps.

5.3 Inter-model Consistency Analysis

The heatmap of Cohen's Kappa consistency coefficients for the extraction results of the five models is shown in Figure 2.

Cohen' s Kappa consistency heatmap

Figure 2 Heatmap of Cohen' s Kappa consistency coefficients between models

Among the five models, the Kappa values were all negative (-0.3014 to -0.2597), whereas the Jaccard similarity was 0.46-0.51, indicating a moderate level of consistency. This contrast stems from the "Kappa paradox" [35]: when the category distribution is extremely imbalanced (with the method category accounting for the highest proportion, the metric category the lowest, and the candidate-entity space reaching several hundred), Kappa may be biased downward or even become negative; at the same time, fine-grained string matching counts different expressions of the same entity (e.g., "Transformer" and "Transformer model") as inconsistent, which also depresses Kappa. After entity normalization, the average Jaccard increased from 0.48 to 0.63, and Kappa became positive (0.18-0.31), indicating that model consistency at the semantic level is far higher than at the string level.

A negative Kappa does not mean that there is no consistency between models. Differentiation provides complementary space for ensemble integration: a combination of "moderate consistency + differentiation" has greater potential for ensemble complementarity than a combination with "high consistency."

5.4 Ablation Experiment: Model Combination Analysis

To analyze the complementarity of model combinations, this paper conducts an ablation experiment on the number of models: under the ≥ 4 criterion, entity F1 is used to remove models in ascending order (DeepSeek→Claude→Tongyi Qianwen). On the remaining subsets, evaluation is repeated using confidence-fusion support filtering (support from ≥ 3 models within a subset; the two-model subset is relaxed to ≥ 2), with the results shown in Table 9.

Table 9 Ablation study on model combinations (entity metrics, ≥ 4 consensus reference standard)

Model combination	Number of models	Precision	Recall	Entity F1	Coverage
GPT-4+Claude+Tongyi Qianwen+GLM-4+DeepSeek	5	0.693	0.942	0.799	94.2%
GPT-4+Claude+Tongyi Qianwen+GLM-4 (DeepSeek removed)	4	0.915	0.942	0.929	94.2%
GPT-4+Tongyi Qianwen+GLM-4 (Claude removed again)	3	0.873	0.942	0.907	94.2%
GPT-4+GLM-4 (Tongyi Qianwen removed again)	2	0.944	0.634	0.759	63.4%

Note: (1) The ablation results are based on support degrees output by confidence fusion and then filtered for approximation (without retraining meta-learners for subsets or re-invoking models), and are used to characterize the marginal contribution direction of model combinations; (2) the entity F1 of the single model GPT-4 is 0.855 (see Table 1), still lower than the four-model and three-model combinations.

The ablation experiment exhibits a counterintuitive phenomenon: after removing DeepSeek, which has the lowest F1 under the ( )4 criterion, the ensemble F1 does not decrease but instead increases (0.799 → 0.929). The mechanism is consistent with Section 5.2.2: under the ( )4 criterion, low-consensus entities supported by only three models are the main source of precision loss. After one model is removed, a large number of entities with three supports are filtered out and excluded, while entities that happen to receive four supports and enter the reference standard are still retained, leading to a significant increase in

precision. However, when the number of models falls to two, although precision further rises to 0.944, recall and coverage drop sharply from 94.2% to 63.4%, and F1 falls to 0.759, indicating insufficient joint coverage. Overall: (1) under the ( )4 criterion, the four-model combination is the optimal configuration balancing precision and coverage; (2) the marginal value of the number of models lies mainly in coverage rather than F1; (3) this again confirms that "the method of constructing the reference standard determines the ensemble evaluation conclusion" —if F1 is the target, a streamlined combination is preferable, whereas if coverage is the target, the five-model setting has the advantage.

5.5 Knowledge Graph Construction Results

The statistics of the knowledge graph constructed based on the confidence-fusion results are shown in Table 10; its scale is significantly larger than that of earlier studies.

Table 10 Statistics of the AI domain knowledge graph (seed-42 run)
Tab.10 Statistics of the AI domain knowledge graph (seed-42 run)

Statistical indicator	Value
Number of nodes	472
Number of edges (including multi-edges)	15,144
Method-type nodes	130
Task-type nodes	138
Dataset-type nodes	138
Metric-type nodes	66
Number of connected components	89
Number of nodes in the largest connected component	384
Number of communities (Louvain)	94
Modularity Q value	0.2121
Average degree centrality	0.0290

The graph contains 472 nodes and 15,144 edges: among node types, task and dataset nodes are tied for the largest number (138 each, 29.2%), followed by method nodes (130, 27.5%) and metric nodes (66, 14.0%). In terms of edge types, "evaluated on" is the most frequent (7,872 edges, 52.0%), followed by "applied to" (7,084 edges, 46.8%); "combined," "extended," "based on," and "improved" together account for only 1.2%, indicating that the model' s recognition of deep methodological relations remains clearly insufficient. Connected components (89, with the largest containing 384 nodes) reflect physical connectivity, while communities (94) are dense substructures further partitioned by Louvain within connected components; therefore, the number of communities is slightly greater than the number of connected components.

Regarding knowledge-graph construction strategies and edge-quality evaluation, the graph includes only relational triples with a fusion score ≥ 0.62 and support

AI-domain knowledge-graph visualization (degree Top-90 node subgraph; node size degree)

Figure 2: AI-domain knowledge-graph visualization (degree Top-90 node subgraph; node size degree)

from at least three models. The reason for choosing confidence-based fusion rather than weighted voting is that the latter has an accuracy of only 0.645, and about 35% of erroneous outputs would generate a large number of spurious nodes and edges; the former offers a better balance between recall (0.927) and precision (0.691), making it more suitable as a graph data source.

In terms of edge reliability, among the 6,247 deduplicated edges extracted by the models, those supported by one model account for 29.0% (1,814 edges), those supported by two models account for 18.7% (1,171 edges), those supported by three models account for 19.1% (1,195 edges), those supported by four models account for 23.2% (1,447 edges), and those supported by five models account for 9.9% (620 edges). The 3,142 deduplicated edges incorporated into the graph all received support from ≥ 3 models, of which 65.8% received support from ≥ 4 models, reducing the influence of hallucinated relations from a single model.

However, the overall performance of relation extraction remains relatively low (single-model F1 is 0.257–0.425). Even after consensus filtering, the graph may still contain a certain proportion of erroneous relations; in the future, manual auditing of sampled edges should be added to evaluate actual reliability.

The visualization result of the knowledge graph is shown in Figure 3.

Visualization of the AI-domain knowledge graph

Figure 3 Visualization of the AI domain knowledge graph

In terms of degree centrality, ACT ranks first (0.461), followed by RAG (0.340), Transformer (0.316), GAT (0.312), and OPT (0.299). It should be noted that abbreviated dataset entities such as ACT, CR, ERE, ARC, and MAG have abnormally high connectivity because of naming conflicts with common vocabulary, and should be interpreted with caution. Within the method category, RAG, Transformer, GAT, and OPT are the core nodes with the highest connectivity. The top five in betweenness centrality are ACT (0.151), RAG (0.068), GAT (0.050), OPT (0.049), and Transformer (0.048); these nodes play a "bridge" role in connecting different research subfields.

Louvain identified 94 research-topic communities (modularity $Q = 0.2121$). The largest community contains 93 nodes, covering large-language-model-related methods (Transformer, BERT, GPT-4, LLaMA, T5, etc.), as well as NLP tasks and datasets; the second-largest community contains 78 nodes, mainly graph neural networks and transfer learning (RAG, GAT, GCN, Transfer Learning, etc.); the third-largest community contains 75 nodes, focusing on computer vision (GAN, CNN, Vision Transformer, Contrastive Learning, etc.) and visual

Emerging frontier topic identification results (Top 10, ranked by logarithmic growth rate)

Figure 3: Emerging frontier topic identification results (Top 10, ranked by logarithmic growth rate)

tasks.

Among the 94 communities, 88 (93.6%) are micro-communities with no more than 3 nodes, indicating a highly long-tailed size distribution. This reflects the fine-grained differentiation of AI research topics, and also suggests the sensitivity of the results to edge density and modularity ($Q = 0.2121$ is below the common threshold of 0.3 for a significant community structure). Therefore, the community-detection results should be regarded as an exploratory characterization rather than a definitive partition.

5.6 Identification of Emerging Frontiers

Based on the time-slicing logarithmic growth-rate analysis, using the criteria of logarithmic growth rate ≥ 0.3 and recent frequency ≥ 5 (indicator-type entities have been excluded), a total of 29 emerging frontier topics were identified (25–29 across five repeated experiments). The top 10 ranked by growth rate are shown in Fig. 4.

Emerging frontier identification results

Fig. 4 Emerging frontier topic identification results (Top 10)

The top five by growth rate are: GPT-4 (3.12, 65 recent papers), Chain-of-Thought (2.75, 44 papers), LLaMA (2.49, 33 papers), RLHF (1.67, 13 papers), and Qwen (1.39, 9 papers), corresponding respectively to the sustained surge in large language models and reasoning-enhancement technologies, as well as the rapid rise of open-source foundation-model ecosystems and human-feedback alignment technologies.

Validation by comparison of methods. This paper also identifies frontier topics using the frequency-threshold method (taking method/task-type entities appearing ≥ 5 times recently as frontier candidates). The comparison results are shown in Table 11.

Table 11 Comparison of frontier identification methods (seed-42 run)

Method	Number of identified topics	Number of unique topics	Number of shared topics	Consistency ratio (intersection/union)
Logarithmic growth-rate method	29	0	29	—
Frequency-threshold method (recent ≥ 5 times)	89	60	29	32.6%

Because the candidate condition of the growth-rate method itself includes "recent frequency ≥ 5," its results constitute a subset of the frequency method—all 29 topics are captured by the frequency method, and the consistency ratio (intersection/union) is 32.6%. The 60 topics unique to the frequency method are all early-stage high-frequency topics with growth rates below 0.3 that remain "persistently popular" (such as GPT, fine-tuning, Vision Transformer, image classification, etc.).

The two methods are complementary: the growth-rate method is a "newly emerging" filter superimposed on frequency thresholds, capturing directions that have risen rapidly after 2023, such as Mamba (8 recent occurrences), LLaVA (6 occurrences), and mathematical reasoning (7 occurrences); whereas the frequency-threshold method cannot distinguish "emergence." This suggests that a single indicator has difficulty fully capturing the diversity of the frontier, and that multi-indicator integration is a direction for improvement.

Structured comparison with external authoritative reports. To verify external validity, this paper compares the Top 10 frontier topics one by one with the AI Index Report 2024[36] and Gartner Hype Cycle for Artificial Intelligence, 2024[37]. The results are shown in Table 12.

Table 12 Comparison of Top 10 frontier topics with external authoritative reports

Rank	Frontier topic	AI Index Report 2024	Gartner Hype Cycle for AI 2024	Comparison conclusion
1	GPT-4	(mentioned in multiple places such as training cost and benchmark performance)	—(not listed separately; belongs to the large-model category)	Directly covered
2	Chain-of-Thought	—	—	Not covered
3	LLaMA	(Llama 2 was an important model in 2023)	—	Directly covered
4	RLHF	(statistics chart on the use of RLHF in foundation models)	—	Directly covered
5	Qwen	—	—	Not covered
6	Mamba	—	—	Not covered
7	mathematical reasoning	(analysis of mathematical reasoning benchmarks such as MMMU)	—	Indirectly covered
8	LLaVA	—	—	Not covered
9	Diffusion Model	(discussion related to generative AI and image generation, not named explicitly)	—	Indirectly covered
10	GPT	(GPT-series models)	—	Directly covered

Among the Top 10, four topics (GPT-4, LLaMA, RLHF, and GPT) are directly mentioned by the AI Index Report 2024, and two topics (mathematical reasoning and Diffusion Model) are indirectly involved, giving a total coverage

Annual evolution trends of paper counts in AI subfields

Figure 4: Annual evolution trends of paper counts in AI subfields

rate of 60%; Chain-of-Thought, Qwen, Mamba, and LLaVA are not covered. The Gartner report is organized by technical categories (29 technologies such as multimodal large models, composite AI, AI agents, and generative AI) and does not list specific models individually. It therefore has only category-level indirect correspondence with the model-type topics in this paper (for example, GPT-4/LLaMA/Qwen correspond to the foundation-model category, and Diffusion Model corresponds to the generative-AI category). Overall, the identification results show a moderately high consistency with authoritative reports at the level of the "model ecosystem"; the lag in external reports' coverage of emerging methods such as Mamba and LLaVA is consistent with this paper's positioning of capturing early signals of "emerging bursts."

The results of topic evolution analysis are shown in Figure 5.

Topic evolution trends

Figure 5 Annual evolution trends of paper counts in AI subfields

The large-language-model direction shows the most significant growth, with the number of papers increasing from 1 in 2019 to 21 in 2024 (10 in 2025), with a growth rate of 3.13, making it the only one among the six subfields with a strong upward trend. Natural language processing has the largest overall scale but is declining slowly (12 papers in 2019 $\rightarrow$ 7 in 2024 and 4 in 2025, growth rate -0.45); reinforcement learning likewise declines (11 $\rightarrow$ 6 $\rightarrow$ 2 papers, growth rate -0.27), indicating that research hotspots are shifting from traditional reinforcement learning toward large language models.

In terms of typical cases, Chain-of-Thought has continued to appear since 2022 (6, 15, 15, and 14 papers in 2022–2025, respectively, 50 in total), with a growth rate of 2.75. As a key technique for improving the reasoning capabilities of LLMs, it connects three subfields in the graph—reasoning tasks, prompt engineering, and large language models—reflecting the cross-domain bridging characteristics of frontier topics. RAG appeared a total of 407 times between 2019 and 2025 (reaching 72 in 2025), with degree centrality (0.340) ranking second, and has become a core hub technology linking information retrieval and text generation.

6 Conclusions and Prospects

6.1 Research Conclusions

Taking the field of artificial intelligence as an empirical case, this paper explored the application of multi-LLM integration in knowledge extraction from scientific and technological papers and in the identification of emerging frontiers. The main findings are as follows:

First, the way in which evaluation benchmarks are constructed has a decisive influence on ensemble evaluation conclusions. Under the criterion of ( )4 consensus, the entity F1 scores of weighted voting (F1=0.771) and confidence fusion (0.779) did not exceed that of GPT-4 (0.855). There are two reasons: (1) more low-consensus entities recalled by the ensemble are penalized for precision under strict criteria; (2) the reference standard itself is incomplete (recall is only 0.519, and about 48% of true entities are omitted), causing the evaluation to systematically favor conservative models. This suggests that caution is needed against the methodological trap that "model consensus means correctness."

Second, ensembles have unique value for relation extraction: the relation F1 values for weighted voting and confidence fusion were 15.3% and 15.1% higher than those of the optimal single model, respectively. This has direct significance for downstream knowledge-graph construction, indicating that ensembles have greater potential for complementarity in tasks requiring multi-factor consistency.

Third, the Stacking diagnostic experiment revealed the risk of overfitting to the consensus reference standard: under the ≥ 4 criterion, its F1 reached 0.942; under ≥ 3, it was only 0.730; and under ≥ 5, it dropped to 0.600. The sharp fluctuations indicate that what the meta-learner learned was the construction rules of the reference standard rather than general recognition capability; thus, F1 = 0.944 cannot be taken as evidence of validity.

Fourth, the knowledge graph constructed by the ensemble can provide preliminary support for frontier identification. The graph with 472 nodes and 15,144 edges (94 communities, modularity $Q = 0.2121$) identified 29 emerging frontiers, including GPT-4, Chain-of-Thought, and LLaMA, according to the established criteria; 60% of the Top 10 were directly or indirectly covered by the AI Index Report 2024. However, relation-extraction performance remains low (single-model F1 of 0.257–0.425), the graph may contain erroneous relations, and the frontier-identification results should be regarded as exploratory findings.

Implications for intelligence studies practice: (1) A multi-LLM ensemble paradigm can be adopted, but costs and benefits must be balanced—an ensemble of five models brings a marginal coverage improvement (92.2% $\rightarrow$ 93.3%) while incurring five times the computational cost; if F1 is the target, a single model is sufficient. (2) The way evaluation benchmarks are constructed has a major impact on conclusions; it is recommended to adopt a hybrid strategy that combines multi-standard evaluation with manually annotated subsets. (3) Graph-based frontier identification can capture structural relationships and evolutionary dynamics among topics; it is recommended to supplement it with multi-indicator fusion and comparative validation to improve robustness.

6.2 Research Limitations and Outlook

This study has the following limitations. (1) Cyclical validation and incompleteness of the evaluation benchmarks. The reference standards and the evaluated objects are both these five LLMs and their ensembles, forming a closed loop;

the recall rate under the ≥ 4 criterion is only 0.519, and multi-standard evaluation remains within the same consensus space, making it impossible to detect systematic biases shared by the models. Manual evaluation of 100 papers partially alleviated this problem, but only one person completed the evaluation; inter-evaluator consistency was not reported, and the dimensions were limited to entity boundaries and types (95% confidence interval approximately $\pm 7\%$). (2) Insufficient traditional baselines. Only a rule-based baseline was introduced ($F1 = 0.474$), while pretrained baselines such as SciBERT-CRF (typically reaching 0.65–0.75) were not included; therefore, it is not appropriate to claim that "LLMs are significantly superior to traditional methods." It can only be stated that LLMs already possess relatively strong extraction capability in a zero-shot setting. (3) The dataset uses only titles and abstracts. The core methodological descriptions and discussions of relations are mostly in the full text; abstract-level extraction may lead to systematic underestimation of coverage and amplification of inter-model differences, and frontier identification may be biased toward "high-exposure" topics. (4) Relation-extraction performance is low, and graph quality is questionable. Even after ensembling, relation F1 is only 0.444, and more than 50% of the triples in the graph may be erroneous; model support is not equivalent to correctness. In the future, manual sampling audits of edges should be conducted (at least 200 are recommended), along with sensitivity analysis of the ≥ 4 supported-edge subset. (5) Validation of the frontier-identification methods is insufficient. Only two self-designed methods were used; mature methods such as Kleinberg burst detection [38] and citation-burst analysis were not introduced. Qualitative comparison with the AI Index Report lacks quantitative indicators. (6) The sample is concentrated in the AI field, and cross-disciplinary generalization remains to be validated; ensemble parameters were set empirically, and confidence scores were not systematically calibrated.

Future research directions include: (1) constructing a complete manually annotated benchmark of 200–300 papers (including entities and relations), and adding at least one independent evaluator to compute inter-evaluator Kappa coefficients; (2) adding pretrained baselines such as SciBERT-CRF and DyGIE++, and directly comparing them with LLMs on the same dataset; (3) extending the analysis to full-text papers; (4) conducting manual sampling audits of graph edges and sensitivity analysis of the ≥ 4 supported-edge subset; (5) introducing mature methods such as Kleinberg burst detection, and using 2019–2023

yearly data, and test their predictive power for actual hot topics in 2024–2025 (prospective validation); (6) systematically calibrate the confidence of each model (e.g., temperature scaling) and assess changes in the performance of the ensemble after calibration.

Data Availability Statement

The arXiv paper metadata used in this article (title, abstract, arXiv ID, publication year, and subfield classification) were obtained through the public API and comply with the arXiv terms of use; the extracted, integrated, and evalu-

ated data are stored in JSON format and may be obtained from the authors upon reasonable request; prompt templates, evaluation scripts, and experimental configurations have been version-controlled to support reproducibility of the results.

References

[1] Qiu Junping. *Scientometrics* [M]. Beijing: Science Press, 2016.

[2] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners [C]//Advances in Neural Information Processing Systems 33. 2020: 1877-1901.

[3] Anthropic. The Claude 3 model cards [EB/OL]. https://www.anthropic.com/news/claude-3-family, 2024.

[4] Bai J, Bai S, Yang S, et al. Qwen technical report [EB/OL]. arXiv: 2309.16609, 2023.

[5] Zhu Y, Yuan H, Wang S, et al. Large language models for information retrieval: A survey [J]. ACM Transactions on Information Systems, 2025. DOI: 10.1145/3748304.

[6] Li Yang, Sun Jianjun. Reflections on the impact of large language models on the development of informatics [J]. Journal of the China Society for Scientific and Technical Information, 2025, 44(2): 246-256.

[7] Park M, Leahey E, Funk R J. Papers and patents are becoming less disruptive over time [J]. Nature, 2023, 613(7942): 138-143.

[8] Wei X, Cui X, Cheng N, et al. ChatIE: Zero-shot information extraction via chatting with ChatGPT [EB/OL]. arXiv: 2302.10205, 2023.

[9] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation [J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[10] Xu D, Chen W, Peng W, et al. Large language models for generative information extraction: A survey [J]. Frontiers of Computer Science, 2025. DOI: 10.1007/s11704-024-40555-y.

[11] Zhang Yiyi, Zhang Chengzhi, Zhou Yi, et al. Multi-perspective academic paper entity recognition based on ChatGPT: Performance testing and usability research [J]. Data Analysis and Knowledge Discovery, 2023, 7(9): 12-24.

[12] Dietterich T G. Ensemble methods in machine learning [C]//International Workshop on Multiple Classifier Systems. Berlin: Springer, 2000: 1-15.

[13] Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models [C]//Proceedings of ICLR 2023. Kigali: ICLR, 2023.

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems 30. Long Beach:

MIT Press, 2017: 5998-6008.

[15] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition [C]//Proceedings of NAACL-HLT 2016. San Diego: ACL, 2016: 260-270.

[16] Wadden D, Wennberg U, Luan Y, et al. Entity, relation, and event extraction with contextualized span representations [C]//Proceedings of EMNLP-IJCNLP 2019. Hong Kong: ACL, 2019: 5784-5789.

[17] Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text [C]//Proceedings of EMNLP-IJCNLP 2019. Hong Kong: ACL, 2019: 3615-3620.

[18] Wei J, Bosma M, Zhao V Y, et al. Finetuned language models are zero-shot learners [C]//Proceedings of ICLR 2022. Virtual: ICLR, 2022.

[19] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback [C]//Advances in Neural Information Processing Systems 35. New Orleans: Curran Associates, 2022: 27730-27744.

[20] Luan Y, He L, Ostendorf M, et al. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction[C]//Proceedings of EMNLP 2018. Brussels: ACL, 2018: 3219–3232.

[21] Jain S, van Zuylen M, Hajishirzi H, et al. SciREX: A challenge dataset for document-level information extraction[C]//Proceedings of ACL 2020. Online: ACL, 2020: 7506–7516.

[22] Edge D, Trinh H, Cheng N, et al. From local to global: A Graph RAG approach to query-focused summarization[EB/OL]. arXiv: 2404.16130, 2024.

[23] Zhang W, Xiao Q L, Liu H J, et al. Large-model-assisted extraction of Chinese cultural loaded words and knowledge graph construction[J]. Digital Library Forum, 2025, 21(1): 33–45.

[24] Zhou Z H. Ensemble methods: foundations and algorithms[M]. Boca Raton: CRC Press, 2012.

[25] Zhang C, Ma Y. Ensemble machine learning: methods and applications[M]. New York: Springer, 2012.

[26] Chen Z, Lu X, Li J, et al. Harnessing multiple large language models: A survey on LLM ensemble[EB/OL]. arXiv: 2502.18036, 2025.

[27] Jiang D, Ren X, Lin B Y C. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion[C]//Proceedings of ACL 2023. Toronto: ACL, 2023: 14165–14178.

[28] Pan S, Luo L, Wang Y, et al. Unifying large language models and knowledge graphs: A roadmap[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7): 3580–3599.

[29] Bao P, Zhang C Z. Evaluation of ChatGPT's Chinese information extraction capability—Using three typical extraction tasks as examples[J]. Data Analysis and Knowledge Discovery, 2023, 7(9): 1–11.

[30] Zhao W L, Wang Y M, Wang J J, et al. From knowledge emergence to diffusion: A study on dynamic measurement of the growth characteristics of "new"–"emerging" topics[J]. Library and Information Service, 2026, 70(7): 83–100.

[31] Tian K, Mitchell E, Zhou A, et al. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback[C]//Proceedings of EMNLP 2023. Singapore: ACL, 2023: 5433–5442.

[32] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008.

[33] Zeng A, Liu J, Du X, et al. ChatGLM: A family of large language models from GLM-130B to GLM-4[EB/OL]. arXiv: 2406.12793, 2024.

[34] DeepSeek-AI. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model[EB/OL]. arXiv: 2405.04434, 2024.

[35] Feinstein A R, Cicchetti D V. High agreement but low kappa: I. The problems of two paradoxes[J]. Journal of Clinical Epidemiology, 1990, 43(6): 543–549.

[36] AI Index Steering Committee. Artificial Intelligence Index Report 2024[R]. Stanford, CA: Stanford University Institute for Human-Centered Artificial Intelligence, 2024.

[37] Gartner. Hype Cycle for Artificial Intelligence, 2024[R]. Stamford, CT: Gartner, 2024.

[38] Kleinberg J. Bursty and hierarchical structure in streams[J]. Data Mining and Knowledge Discovery, 2003, 7(4): 373–397.

Note: Figure translations are in progress. See original paper for figures.