Classifying the Research Level of Scientific Articles With Fine-Tuned LLMs

Xiaomei Wang [1 2], Xinyu Song [1 2], Zhengyin Hu [3]

[1] Institutes of Science and Development, Chinese Academy of Sciences

[2] University of Chinese Academy of Sciences.

[3] National Science Library(Chengdu), Chinese Academy of Sciences

2026.02.07

Xiaomei Wang

wangxm@casisd.cn

Professional Experience

- Researcher of Institutes of Science and Development, CAS
- Head of the Intelligence Analysis Technology and Data Platform team.
- Research primarily focuses on information analysis methods and technologies, scientific mapping analysis methods and tools, and Intelligent reasoning using scientific and technological big data.

Publications & Activities

- Mapping Science Strcture2022, China Science Press
- Mapping Technology Structure 2022, public release

Aims and contributions of our paper

01

In our paper, we aimed at:

- Constructing a multidisciplinary classification framework for research levels, overcoming insufficient theoretical foundations and limited semantic recognition of existing methods.
- Automatically classifying scientific papers to reveal disciplinary evolution and facilitate research planning & knowledge transfer.

02

Contributions of our study:

- Generalized framework – Established a generalized research-level classification framework for multi-disciplinary literature.
- Hybrid LLM tuning – Proposed a hybrid LLM tuning strategy (LoRA + prompt engineering) on LLaMA3-8B.
- Achieved 85.45% F1 and 85.50% accuracy, significantly outperforming logistic regression (62.71% / 62.00%).

CONTENT

PART 1 Introduction

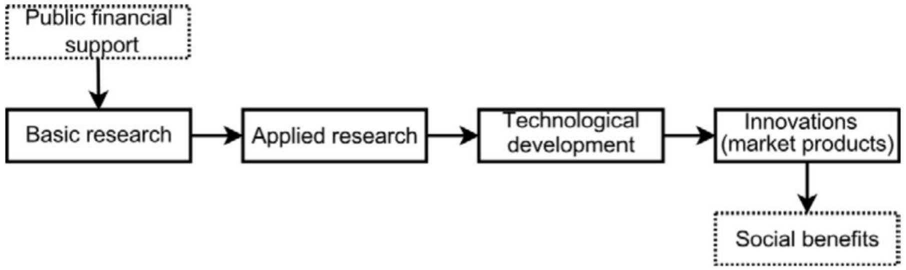
PART 2 Related Work

PART 3 Model

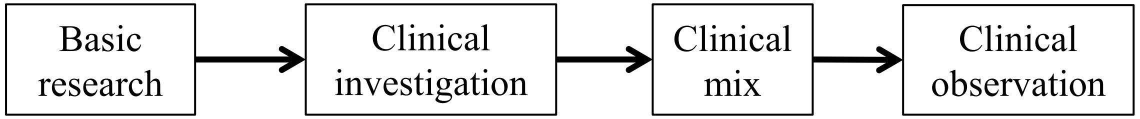
PART 3 Experimental Results

1.Introduction

Scientific research is often described as following a **linear model**.



In the 1970s, **Narin et al.** graded scientific research activities according to the "**Research Level**" in the medical journal system.



- Map the entire process of scientific research outcomes from theoretical discovery to practical application
- Support resource allocation and outcome evaluation
- Bridge basic research and applied practice to optimize the research ecosystem

Challenges and Limitations:

- Inconsistent definitions
- Blurred stage boundaries
- Limited cross-domain transferability
- Absence of unified interdisciplinary frameworks
- Lack of quantitative and reproducible criteria
- Coarse-grained classification at journal level
- ……

A strong need for a simple, unified, and cross-disciplinary way to classify research levels.

2.Related Work

Based on textual information

- Keyword statistics
- Triangle of Biomedicine
- Manual labeling
- ……

Relies on manual label and traditional machine learning, with **limited deep semantic understanding**

Shows **limited accuracy** when handling multi-domain scientific texts with high semantic complexity.

Based on citation network structure

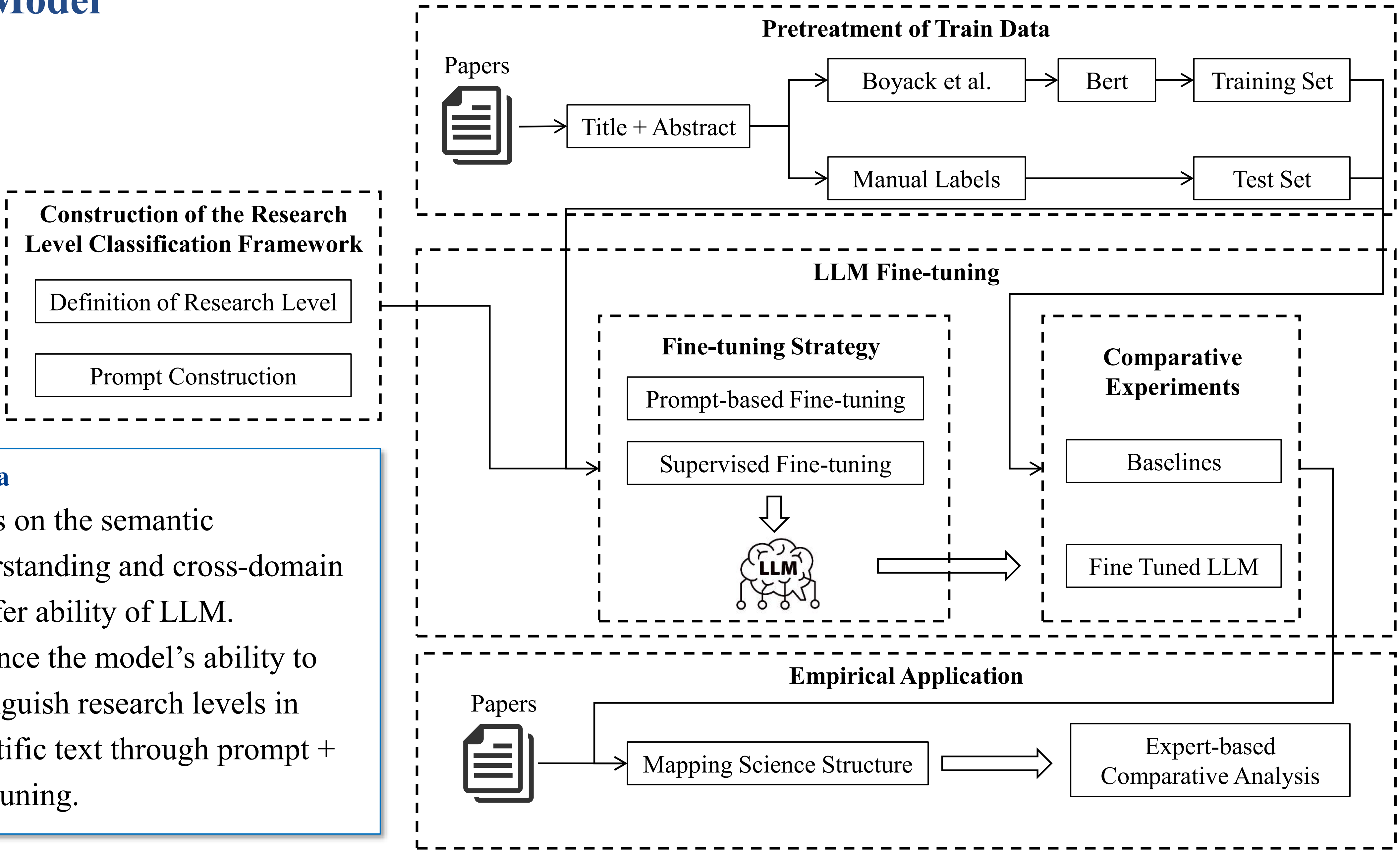
- Research Output Tracking Model
- Funding-Based Translational Analysis
- Relative Citation Ratio (RCR)
- ……

Large Language Models

- Performs well in tasks, such as sentiment analysis and topic classification
- Strong ability to model context and capture semantic differences
- Strong transferability and high automation
- ……

A research level classification method based on fine-tuned LLMs.

3.Model



Core Idea

- Focus on the semantic understanding and cross-domain transfer ability of LLM.
- Enhance the model's ability to distinguish research levels in scientific text through prompt + fine-tuning.

3.1 Research Level Definition

中国科学院科技战略咨询研究院
Institutes of Science and Development, Chinese Academy of Sciences

OECD × TRL: A Complementary Framework

OECD (Organisation for Economic Co-operation and Development) captures the drivers and goals of research.

2.9 The term R&D covers three types of activity: basic research, applied research and experimental development. **Basic research** is experimental or theoretical work undertaken primarily to acquire new knowledge of the underlying foundations of phenomena and observable facts, without any particular application or use in view. **Applied research** is original investigation undertaken in order to acquire new knowledge. It is, however, directed primarily towards a specific, practical aim or objective. **Experimental development** is systematic work, drawing on knowledge gained from research and practical experience and producing additional knowledge, which is directed to producing new products or processes or to improving existing products or processes. These three types of R&D are discussed further in Section 2.5.

×

TRL (Technology Readiness Levels) reflects technological realization pathways and maturity.

Implementation phase	Initiation						Maturity (Early Implementation)		
Domains of Readiness	Pre-readiness						Readiness		
TRL context	Conceptual Stage				Simulated & Relevant Environment		Real World Environment		
TRL functions	Prototyping				Piloting		Demonstration	Deployment	
TRL Levels	1	2	3	4	5	6	7	8	9
TRL description	Basic Principles	Knowledge base formalised	Proof of Concept	Prototype	Validation of Prototype	Test in a Relevant Environment	Demonstration in the Real-World Environment	Pre-release	Release

Research Level	TRL	OECD
Basic Scientific Research	TRL 1-2	Basic Research（2.25–2.28）
Applied Research	TRL 3	Applied Research（2.29–2.31）
Engineering–Technological Mix Reseach	TRL 4	Experimental Development （2.32， Early phase）
Technology Validation Research	TRL 5-6	Experimental Development （2.32–2.33， Later phase）

3.2 Prompt Design

Concise prompts

- Task instruction
- The numbers and names of the research levels.
- Relies on its prior knowledge and reasoning ability.

```
{
  "instruction": "You are a classifier of research levels for academic articles. Based on the article's title and abstract, categorize the article into one of the four levels:"
        "1. Basic Scientific Research "
        "2. Applied Research "
        "3. Engineering-Technological  Mix Research "
        "4. Technology Validation Research "
        "Please make a reasonable judgment on the article's research level according to the provided title and abstract.",
  "input": "title: {TI}, abstract: {AB}",
  "output": {}
}
```

Detailed prompts

Provide more explicit semantic guidance.

- Assigned role
- Clear task objective
- Introduction of discriminant criteria
- Provide classification rubric
- Context awareness
- Examples for support

Research level

Role

Task

Classification definition

Classification standard

Classification thinking framework

Classification rubric:

Step 1: Identify the Main Research Objective

If the goal is to understand fundamental principles → Level 1

If the goal is to apply known knowledge to a specific problem → Level 2

If the goal is to implement and test a working prototype → Level 3

If the goal is to validate the prototype in realistic or quasi-realistic environments → Level 4

Step 2: Identify the Nature of the Contribution

Generation of new knowledge, theories, or understanding → Level 1

Application of existing knowledge to new problems or contexts → Level 2

Implementation of full systems, prototypes, or platforms → Level 3

Feasibility assessments in quasi-real-world settings → Level 4

Context Sensitivity

Journal Context

Examples from different fields

3.3 Fine-tuning Strategy

Fine-tuning Strategy: Supervised + LoRA fine-tuning.

- **A text classification task**.
- Balance performance and resource efficiency.
- Directly update model parameters to make the outputs closer to the ground-truth labels.

Principles for Model and Platform Selection

- ❌ **Local deployment**
- ✅ **Cloud-based deployment**
- ❌ **Closed-source models**

Metrics

- **Classification metrics**:
  - Accuracy, Precision, Recall, and Weighted F1.
- **Generation quality metrics**:
  - ROUGE（Recall-Oriented Understudy for Gisting Evaluation）-1 / ROUGE-2
  - BLEU（Bilingual Evaluation Understudy）-4

Training platform：LLaMA Factory

- Supports full fine-tuning, LoRA, and Prefix Tuning.
- Flexible adaptation to mainstream language models.

LLaMA-Factory
Easy and Efficient LLM Fine-Tuning

Model: LLaMA3-8B-Chinese-chat

- Strong Chinese capability
- GPU-friendly

4. Experimental Results

High-confidence test set: Basic Capability Comparison

500 papers, each with a logistic regression probability exceeding 0.85, carefully balanced across all 22 ESI disciplines and further validated through manual spot-checking.

Model	F1	Accuracy	Precision	Recall
Bert+Attention	0.9549	0.9551	0.9549	0.9551
LlaMA3+Prompt (Detailed)	0.5090	0.5463	0.5083	0.5463
Llama3+Prompt (Concise) + Supervised	0.8519	0.8469	0.8657	0.8469
Llama3+Prompt (Detailed) + Supervised	0.9458	0.9459	0.9467	0.9459

Human-annotated test set: Classification Performance on Complex Tasks

200 papers, randomly sampled from data not used in training to avoid selection bias, with each paper independently labeled by two researchers following the predefined four-level classification scheme.

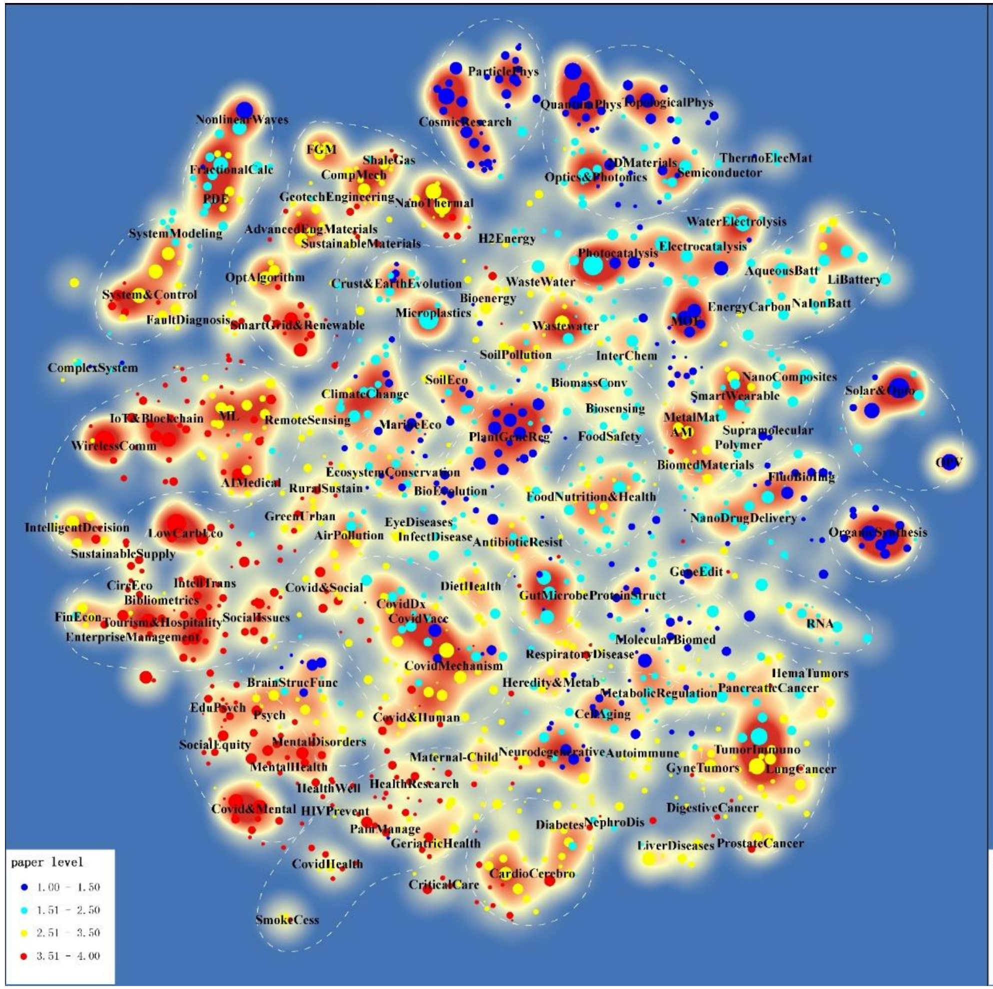
Model	F1	Accuracy	Precision	Recall
Logistic	0.6271	0.62	0.6481	0.62
Bert+Attention	0.6885	0.69	0.7173	0.69
LlaMA3+Prompt (Concise)	0.3543	0.34	0.4539	0.34
LlaMA3+Prompt (Detailed)	0.385	0.39	0.4332	0.39
Llama3+Prompt (Concise) + Supervised	0.7152	0.715	0.7292	0.715
Llama3+Prompt (Detailed) + Supervised	0.8545	0.855	0.8554	0.855
chatglm3-6b+Prompt (Detailed) + Supervised	0.7602	0.76	0.7888	0.76

4. Experimental Results

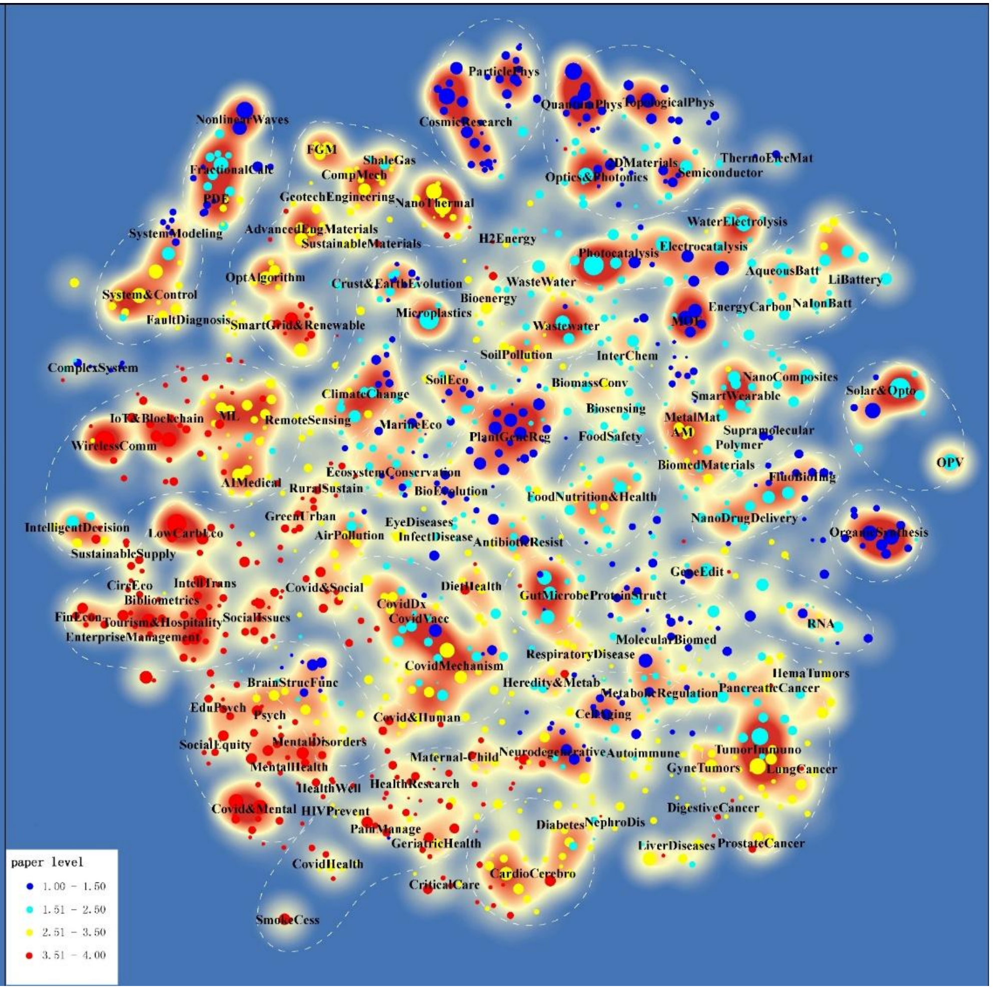
Data：Highly cited ESI papers （2016-2021）

中国科学院科技战略咨询研究院
Institutes of Science and Development, Chinese Academy of Sciences

Logistic

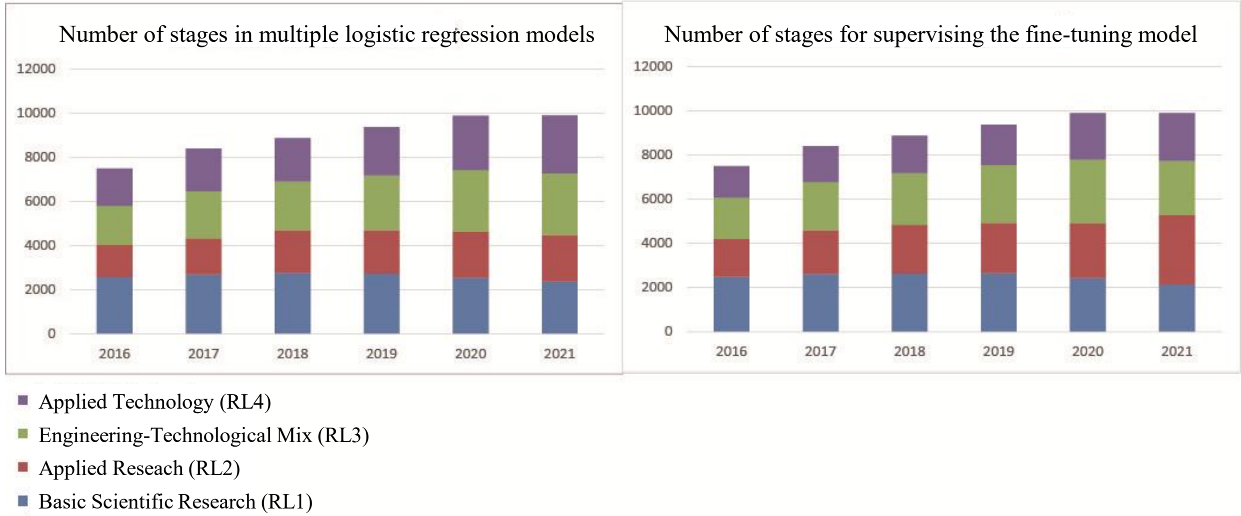


LLM

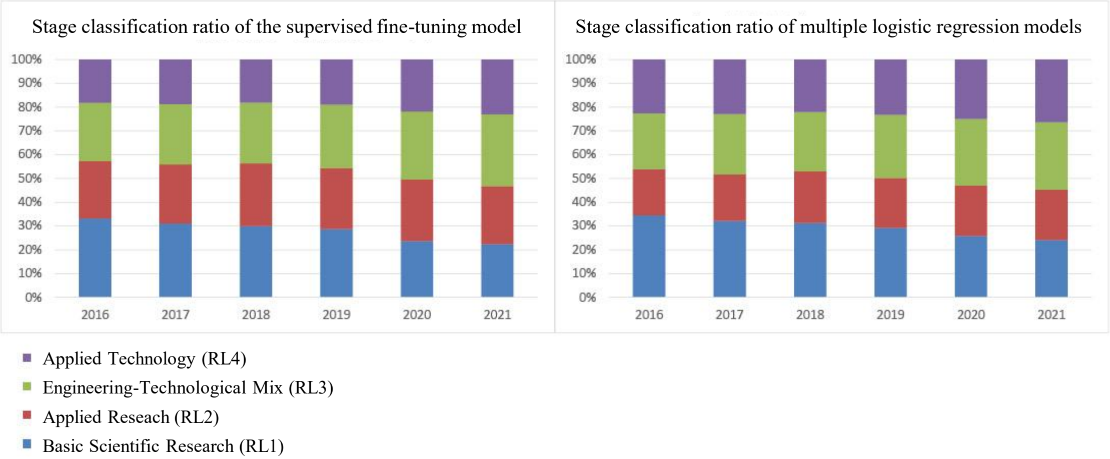


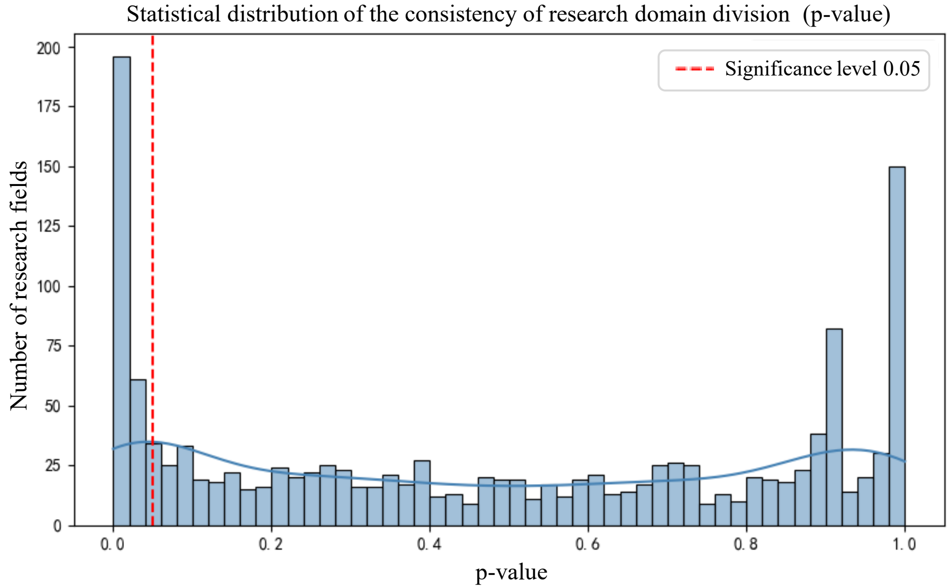
Each node represents a research area (RA) composed of highly cited papers. Node size reflects the number of papers, while node color is encoded by the weighted average of its research level (1–4). Lower scores indicate levels closer to basic research.

4. Experimental Results



Multinomial logistic regression and supervised fine-tuned models show consistent overall trends, but differ in detailed structural distributions.





The two methods match in about 80% of the fields. However, **272 fields (19.6%) show statistically significant differences ($p < 0.05$)**, with 197 fields having $p < 0.02$ — far beyond random expectation. This means that **in nearly one-fifth of all research fields, the LLM produces systematically different research level assignments from the logistic regression model,** revealing its finer semantic discrimination ability.

Conclusion

01

4-level classification framework

- Basic Research – Applied Research – Engineering-Technical Mix – Technology Validation
- based on OECD Frascati Manual (2015) and Technology Readiness Levels
- provides clearer definitions and criteria, better capturing scientific research progression and supporting LLM semantic understanding

02

First systematic application of LLMs for research level classification

- supervised fine-tuning + prompt engineering + LoRA (efficient tuning)

03

Comparative Experiments Validating Fine-Tuning Strategy

- compared multiple strategies, prompt styles, and model configurations.
- supervised fine-tuning achieves higher accuracy, stability, and generalization.

04

Validation via Science Structure Mapping

- visualized stage distributions in real-world research contexts (2016–2021).
- particularly effective in interdisciplinary and engineering-oriented fields, reducing keyword bias and improving cross-domain generalization.

Limitations & Future Work

Limitations: data bias from pseudo-labels; input limited to title and abstract; linear stage model may not capture non-linear research paths.

Data

- Combine main text paragraphs, citations, author affiliations, and field context.
- Explore signals from citations and institutions.
- ……

Prompt & Tuning

- AutoPrompt
- learnable prompts
- controllable parameter tuning
- ……

Applications

- Research and policy intelligence platforms
- Deploy in research management systems
- Research resource recommendation
- ……

Task Expansion

- Trend prediction
- Research value assessment
- Automatic review generation
- ……

Thanks!

2026.02.07