Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD27290

A Major Recommendation System for University Admissions Based on Knowledge Graphs and Large Language Models

Yutong Leng

Institute of Software Engineering, Jilin University, Changchun, China
13756987653@163.com

Abstract. Most college entrance examination candidates in China face significant difficulties and confusion when filling out their university and major preferences. Existing systems that provide access to historical admission data can offer some assistance in decision-making. However, these systems typically require users to input information in a rigid, structured format, making it difficult to support natural and personalized interactions with students. On the other hand, large language models (LLMs) possess strong capabilities in natural language understanding and generation, enabling more user-friendly dialogue interfaces. Yet, due to the well-known issue of hallucination in LLMs, the reliability and factual correctness of their outputs cannot be fully guaranteed. To address these challenges, this paper constructs a structured knowledge graph of university-major admission data based on historical college entrance examination scores. This knowledge graph serves as a trustworthy information base to constrain and enhance the recommendation process. Building on this foundation, we propose a university and major recommendation framework that integrates extraction-enhanced techniques and large language models. By leveraging the knowledge graph to guide the generation process, our system improves both the accuracy and relevance of the recommendations. Experimental results demonstrate that the proposed method significantly enhances the precision of university and major recommendations, indicating its strong potential for real-world application and further development.

Keywords: Major Recommendation, Knowledge Graph, Large Language Model, College Admissions, Recommendation System

1. Introduction

College entrance examination candidates in China often experience confusion and anxiety when selecting universities and majors. According to Cheng et al. [1], more than 71% of senior high school students lack the ability to assess the alignment between their interests and potential majors, and approximately 66% are uncertain about their future career paths or how they relate to academic disciplines. Although artificial intelligence (AI)-based systems have been developed to support decision-making, in most families these tools are primarily operated by parents with limited student involvement. Existing systems tend to emphasize score-based matching while overlooking long-term career alignment and personal interest orientation [2]. Traditional recommenders tend to favor popular items, which may reinforce echo chambers and ignore users' less dominant interests or

DOI: 10.54254/2755-2721/2025.LD27290

emergent preferences [3]. Furthermore, studies reveal that various external influences—such as family expectations, school guidance, university promotion strategies, and online platforms—play a significant role in shaping students' decisions amid conflicting factors such as interest versus return on investment, idealism versus employment reality, and societal expectation versus personal uncertainty [4].

With the rapid advancement of artificial intelligence, large language models (LLMs) have emerged as powerful tools for natural language interaction and information processing. However, a critical limitation of LLMs lies in their susceptibility to hallucination—generating content that is factually incorrect or fabricated. Adel et al. [5] report that while generative AI tools can reduce task workload by 60–65% on average, their performance in tasks requiring explanation or precise reasoning remains extremely low, with accuracy dropping to as little as 4.6% and hallucination rates reaching up to 91%. Theoretical studies further argue that LLMs are fundamentally incapable of learning all computable functions, rendering hallucination a mathematically inevitable phenomenon [6]. Even among leading models, such as ChatGPT, the problem persists: GPT-3.5 exhibits a hallucination rate of 39.6%, GPT-4 at 28.6%, and Google Bard at a staggering 91.4% [7].

To address the aforementioned issues of interest mismatch and hallucination, this study proposes a novel recommendation system grounded in a structured knowledge graph constructed from historical university admission data. The knowledge graph integrates key entities such as universities, majors, regions, scores, and enrollment years, providing a reliable foundation for informed recommendations. On top of this graph, we develop a recommendation framework enhanced with an information extraction mechanism, which guides the large language model to generate more accurate and contextually relevant suggestions. Experimental results demonstrate that our approach significantly outperforms traditional score-based systems in terms of recommendation accuracy, indicating its practical value and potential for broader deployment. The overall system architecture is illustrated in Figure 1.

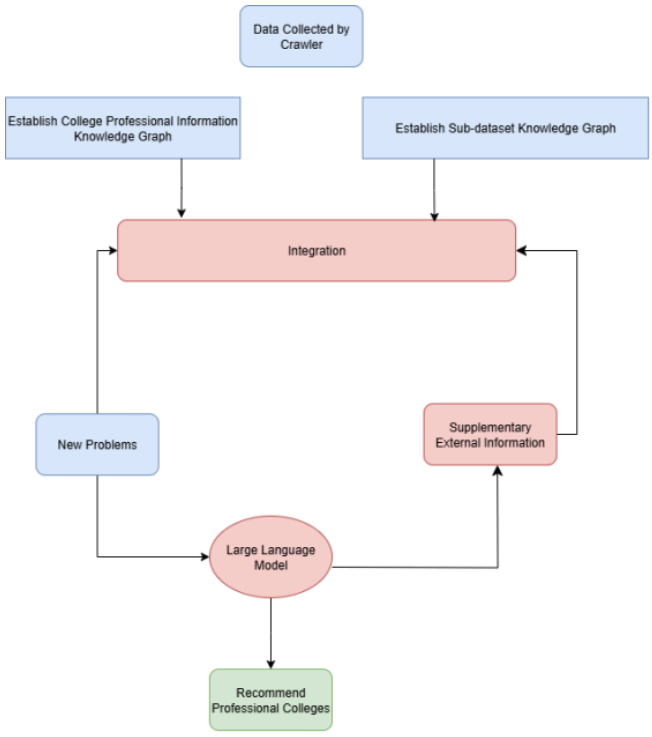

Figure 1. Overall system architecture

DOI: 10.54254/2755-2721/2025.LD27290

The remainder of this paper is organized as follows: Section 2 introduces the core technologies employed in our system; Section 3 details the implementation framework, including the construction of the knowledge graph and the enhanced extraction mechanism; Section 4 presents the experimental setup and evaluation results; Section 5 summarizes the key findings of the study; and Section 6 discusses potential directions for future research and system improvement.

2. Related work

2.1. Applications of Knowledge Graphs

The concept of the Knowledge Graph (KG) was officially introduced by Google on May 17, 2012, as an effort to enhance the capabilities of search engines by incorporating structured semantic information. However, its technological foundation can be traced back to Berners-Lee's 2006 proposition of the Linked Data paradigm, which envisioned the interconnection of distributed web data through standardized protocols. A knowledge graph typically serves as an auxiliary knowledge base that integrates entities and their relationships, thereby enabling semantic-level computation and reasoning.

The core technologies underpinning knowledge graph construction include knowledge extraction, knowledge fusion, and knowledge refinement. Knowledge extraction involves identifying entities and their relations from unstructured or semi-structured data sources; knowledge fusion resolves conflicts and redundancies across heterogeneous sources; and knowledge refinement ensures consistency and usability through schema alignment and quality control.

In the era of big data and the semantic web, knowledge graphs have emerged as foundational infrastructure across a wide array of applications. Their contributions to downstream tasks such as information visualization, question answering, intelligent summarization, and semantic search have been widely validated in both academic and industrial settings [8]. Contemporary AI systems, including prominent large-scale models like ChatGPT and DeepSeek, have increasingly incorporated knowledge graphs to mitigate hallucinations and enhance the factual correctness of generated outputs.

Empirical studies have further highlighted the pedagogical potential of knowledge graphs. For instance, the work by Yang et al. [9] demonstrates that integrating KGs into educational recommendation tools not only improves the precision of content suggestions but also fosters student engagement and enriches interactive learning experiences in online environments. These findings underscore the multifaceted utility of KGs in both cognitive computing and human-centric applications.

2.2. Applications of Large Language Models

Large Language Models (LLMs) represent a class of deep learning architectures designed to understand, generate, and manipulate human language at scale. They rely on a suite of enabling technologies, including scaling laws, data engineering, efficient pretraining strategies, capability elicitation, human alignment, and tool integration. The unprecedented depth of linguistic and conceptual understanding exhibited by LLMs distinguishes them from traditional natural language processing models and marks a significant leap in AI capabilities.

LLMs have demonstrated strong performance across a diverse set of domains. In the educational sector, they are widely applied to tasks such as exam question generation, automated grading, personalized feedback, and adaptive content recommendation, collectively encompassing nine

DOI: 10.54254/2755-2721/2025.LD27290

distinct categories of educational applications [10]. In the software engineering field, LLMs support code generation, bug fixing, code explanation, and natural language documentation, thereby accelerating the software development life cycle and lowering the technical barrier for non-expert programmers [11]. In the financial domain, LLMs can perform zero-shot and few-shot learning, fine-tuning, and the construction of customized models. A decision-making framework tailored for financial professionals has been proposed, highlighting the associated challenges and limitations [12].

The broader scientific community has also embraced LLMs as instruments for discovery. Zhang et al. conducted a comprehensive analysis of over 260 LLMs tailored for scientific use cases, documenting their integration into cross-modal tasks in chemistry, biology, astronomy, and climate science [13]. Their findings demonstrate the capacity of LLMs to facilitate hypothesis generation, data interpretation, and automated reasoning, thereby streamlining the scientific research pipeline.

Overall, the widespread deployment of LLMs across technical and non-technical domains has positioned them as transformative tools in artificial intelligence , with demonstrated capabilities in fields ranging from education and medicine to legal analysis and scientific discovery [14]. Moreover, their integration with structured knowledge sources—such as knowledge graphs—has been shown to significantly enhance factual reliability, interpretability, and performance in real-world applications, particularly in mitigating hallucinations and supporting semantic-level reasoning [15].

3. Methodology

3.1. Knowledge graph construction

This study constructs two types of domain-specific knowledge graphs to support university and major recommendation: (1) the "Major-University-Admission Score Graph" and (2) the "Major-Interest Graph." The former facilitates decision modeling for college entrance recommendation, while the latter supports interest-based major personalization.

3.1.1. Major-university-admission score graph

This graph comprises two substructures: (a) the university base information layer and (b) the 2024 admission score records for each major in each university in Liaoning Province, China. The university base information includes attributes such as university name, province, city, classification (e.g., 985, 211), and institution type (e.g., comprehensive, science & engineering). These data were collected from the official website via a web scraping pipeline, detailed below:

```
Input: "gx.csv" (list of university names)
Output: CSV file with structured fields [Name, Province, City,
Address, Introduction, 985, 211, Type, Attributes, Featured Majors]
1. Read "gx.csv" → school_list
2. For each school_name in school_list:
a. Search on https://gkcx.eol.cn using school_name
b. If detail page not found → log & continue
c. Parse detail page to extract:
- Name, Province, City, Address
```

DOI: 10.54254/2755-2721/2025.LD27290

```
- Introduction, 985/211, Type, Attributes, Featured Majors
d. Append extracted info to results
e. Sleep 1–2 seconds (randomized)
3. Write results to output CSV
```

The admission score data were sourced from the Liaoning Province official exam website. Specifically, the file titled "2024 Liaoning Province General Undergraduate Admission Scores for Additional Rounds" was extracted post-login. This dataset contains structured information on university, major, subject stream (Physics or History), batch, and minimum admission score, which are linked in the graph.

3.1.2. Major-interest graph

This graph captures the association between undergraduate major categories and student interests. First, we collected the full list of Ministry of Education (MoE) regulated undergraduate majors, categorized into 91 predefined major classes. Semi-automatic classification into these classes was performed using LLMs (e.g., DeepSeek, ChatGPT). Next, a hobby taxonomy was generated using DeepSeek, mapping each interest category to the most relevant major class. The algorithm below outlines the logic used to assign hobbies to major categories:

```
Input: CSV with [Major Name, Major Category]
Output: CSV with [Major Name, Major Category, Hobbies]
generate_hobby_templates():
Return dict: Major Category → list of 5–10 hobbies
read_majors(file_path):
Read CSV into list of dicts with "Major Name" and "Major Category"
assign_hobbies(majors, templates):
For each major:
major["Hobbies"] = templates.get(major["Major Category"], [])
save_csv(data, file_path):
Write CSV with columns: Major Name, Major Category, Hobbies (comma-separated)
main():
majors = read_majors("11.csv")
templates = generate_hobby_templates()
assign_hobbies(majors, templates)
save_csv(majors, "hobbies.csv")
Print "Done! X majors processed."
main()
```

The node and relationship types of the integrated knowledge graph is shown in Table 1 and Table 2.

DOI: 10.54254/2755-2721/2025.LD27290

Table 1. Node types in the knowledge graph

Node Type	Description
University	Higher education institution entity
Major	Specific academic discipline
MajorClass	Academic major category (e.g., Finance)
Hobby	Student interests and preferences
Province	Geographic location of university
Score	Admission score record
Stream	Exam stream: Physics or History
Batch	Admission batch (e.g., Regular)
Year	Admission year (e.g., 2024)

Table 2. Relationship types in the knowledge graph

Relationship	Description
HAS_HOBBY	Connects a major class to relevant hobbies
BELONGS_TO	Maps a specific major to a major class
ADMITTED_WITH_SCORE	Associates a major with its admission score
BELONGS_TO_STREAM	Links a score to its subject stream
BELONGS_TO_BATCH	Denotes admission batch of a score record
BELONGS_TO_YEAR	Identifies the year of a score record
OFFERS	University offers the given major
LOCATED_IN	Geographical relation between university and province
Relationship	Description
HAS_HOBBY	Connects a major class to relevant hobbies
BELONGS_TO	Maps a specific major to a major class
ADMITTED_WITH_SCORE	Associates a major with its admission score
BELONGS_TO_STREAM	Links a score to its subject stream
BELONGS_TO_BATCH	Denotes admission batch of a score record
BELONGS_TO_YEAR	Identifies the year of a score record
OFFERS	University offers the given major
LOCATED_IN	Geographical relation between university and province

The full ontology structure is illustrated in Figure 2.

DOI: 10.54254/2755-2721/2025.LD27290

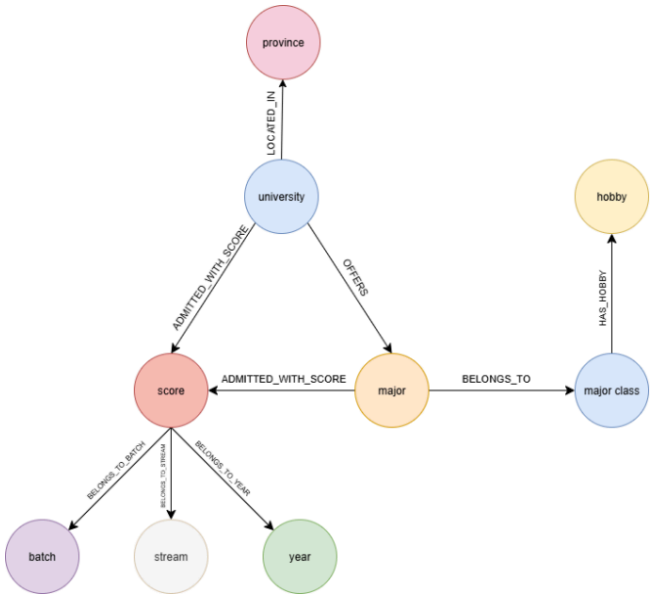

Figure 2. The ontology of the knowledge graph

3.2. Extraction-enhanced recommendation mechanism

To operationalize the constructed knowledge graph for real-time and user-centric recommendations, we propose an extraction-enhanced matching framework that systematically processes natural language user queries and provides ranked university-major recommendations. When a user submits a query such as"I am interested in quantitative trading, scored 546, and want to study in a university in Henan. I belong to the Physics stream." The system engages in the following comprehensive steps.

3.2.1. Information extraction

The framework employs natural language processing techniques to accurately parse the user's input, extracting a set of key attributes, including but not limited to personal interests, examination score, preferred geographic region, and subject stream (e.g., Physics or History). This process ensures that both explicit and implicit user intents are captured and mapped to corresponding entities in the knowledge graph.

3.2.2. Graph-based candidate generation and filtering

The system initiates the search by mapping the extracted user interests to the Hobby nodes within the graph. It then locates the associated MajorClass (e.g., Finance) that best aligns with these interests and retrieves all relevant Major nodes under this category. Next, the system filters for universities that offer these majors and are geographically situated in the user's specified province (e.g., Henan), by leveraging the LOCATED_IN relationship in the graph. For each university-major pair identified, the system further constrains the candidate set by examining historical admission records. Specifically, it selects Score nodes corresponding to the user's examination stream (e.g., Physics), designated batch (e.g., regular undergraduate), and the target year (e.g., 2024). Only those entries where the absolute difference between the user's score and the official admission score does not exceed a defined threshold (e.g., 50 points) are retained. The final set of eligible university-

DOI: 10.54254/2755-2721/2025.LD27290

major pairs is ranked according to the proximity of the admission score to the user's actual score. This ranking is performed in ascending order of |score_user - score_admission|, ensuring that the most competitive and achievable options are prioritized.

3.2.3. Recommendation output and interpretability

The system presents the user with a ranked list of recommended universities and majors, along with key information such as admission scores, score differentials, and institution location. This transparent presentation not only supports informed decision-making but also provides interpretability by elucidating the rationale behind each recommendation.

Through this end-to-end, extraction-enhanced recommendation process, the system effectively integrates user preferences, interests, and academic qualifications with historical admission data embedded in the knowledge graph. This approach delivers personalized, accurate, and interpretable recommendations, thereby supporting students in making optimal choices for their academic and professional futures.

4. Methodology

4.1. Knowledge graph construction

In this work, we leveraged the state-of-the-art GLM-4-plus large language model to systematically structure diverse entities—universities, majors, admission scores, and student interests—into a comprehensive knowledge graph. This knowledge representation supports entity pattern discovery, enabling the system to match candidate university-major combinations to individual student profiles. Furthermore, our system allows users to submit natural language queries and receive structured, data-driven recommendations in response.

To support these functionalities, we constructed five core data files: example.xlsx, gx.csv, hobbies.csv, lishi.csv, and wili.csv. University profile information was acquired via web scraping from the official platform, while the minimum admission scores for each major in 2024 were sourced from the official website of the Liaoning Provincial Education Examinations Authority.

The types and quantities of nodes and relationships in the constructed knowledge graph are summarized in Table 3.

Table 3. Node and relationship types and quantities in the university-major admission knowledge graph

Node Type	Quantity	Relationship Type	Quantity
University	2,867	HAS_HOBBY	453
Batch	14,298	BELONGS_TO	10,581
Hobby	438	ADMITTED_WITH_SCORE	28,596
Major	4,298	BELONGS_TO_STREAM	14,298
MajorClass	91	BELONGS_TO_BATCH	14,298
Province	33	BELONGS_TO_YEAR	14,298
Score	14,298	OFFERS	14,298
Stream	14,298	LOCATED_IN	2,822
Year	14,298		

DOI: 10.54254/2755-2721/2025.LD27290

This extensive schema ensures broad coverage and fine-grained granularity, providing the foundation for robust matching and recommendation.

4.2. Recommendation system performance evaluation

To rigorously evaluate the effectiveness of our proposed recommendation framework, we designed a set of performance tests using the example.xlsx dataset, which contains eight key fields (columns A–H):Interest (student's stated interest), Target Province (location of desired universities), Score (student's entrance exam score), Track (Physics or History stream), Final Major (major actually chosen), Final University (university actually attended), final_decision (recommendation generated by our model), baseline (recommendation from a baseline model).

A detailed description of each input and output field is provided in Table 4.

Table 4. Input and output variables for the college entrance recommendation system

Field	Description
Interest	User-specified interests
Target Province	Intended province for university selection
Score	User's entrance examination score
Track	Physics/History stream selection
Final Major	Major actually chosen by the user
Final University	University actually chosen by the user
final_decision	Recommendation provided by the proposed system
baseline	Recommendation provided by the baseline system

We conducted a comparative evaluation of our knowledge graph-enhanced system against a conventional baseline, which relies solely on keyword-based matching and search. The principal distinction is that our system applies extraction-enhanced reasoning grounded in a structured knowledge graph, while the baseline only performs surface-level language matching. As a result, the empirical data on actual final university and major choices demonstrate that our system significantly outperforms the baseline in terms of recommendation accuracy. Tables 5 and 6 present the comparative results for major and university recommendation accuracy, respectively.

Table 5. Major recommendation accuracy

	Top 1	Top 3	Top 5	Top 10	Top 15
Ours	0.230769	0.538462	0.923077	1	1
Baseline	0	0	0	0.230769	0.307692

Table 6. University recommendation accuracy

	Top 1	Top 3	Top 5
Ours	0.307692	0.769231	1
Baseline	0	0	0.692308

As indicated in the tables, our knowledge graph-driven recommendation system achieves substantial improvements across all evaluation metrics. The model not only enhances the accuracy

DOI: 10.54254/2755-2721/2025.LD27290

of top-ranked recommendations but also demonstrates robust performance as the candidate set expands, underscoring the practical utility and reliability of the proposed approach for supporting college entrance decision-making.

5. Conclusion

College entrance preference selection, as a pivotal process influencing students' future development, often presents considerable confusion and uncertainty for both candidates and their families. Although existing systems for querying historical admission data provide some reference value, they typically rely on structured input formats and thus lack support for natural language interaction, limiting their usability and user experience. Meanwhile, the advent of Large Language Models (LLMs) has created new opportunities for more natural and effective human-computer dialogue. However, the inherent "hallucination" problem in LLMs—that is, the potential to generate information that is inaccurate or not grounded in reality—raises concerns regarding the reliability of their outputs.

To address these challenges, this study proposes a knowledge graph construction methodology for university-major admission information, grounded in historical admission records. Building upon this knowledge graph, we further design a recommendation system framework that integrates extraction-enhanced strategies to facilitate more precise matching between candidates and academic programs. Experimental results demonstrate that the proposed approach yields substantial improvements in recommendation accuracy.

6. Discussion

There are two main limitations in this study. First, interest mapping is not fully accurate because the hobbies.csv file uses the national undergraduate major catalog published by the Ministry of Education, while the lishi.xlsx and wuli.xlsx files are based on 2024 admission data from the Liaoning Provincial Education Examination Authority. As a result, inconsistencies in major names across different sources led to incomplete matching. This could be addressed in future work by expanding and standardizing the major datasets for better coverage. Second, the data volume is limited, as the current system only incorporates admission records from Liaoning Province in 2024. To improve the system's robustness and practical value, future research should collect and integrate multi-year data from multiple provinces using the methods described in this paper.

References

[1] MCheng, Y., & Hamid, M. O. (2025). Social impact of Gaokao in China: A critical review of research. Language Testing in Asia, 15(1), 22.

[2] Chen, S., Xie, J., Wang, G., Wang, H., Cheng, H., & Huang, Y. (2024). From score-driven to value-sharing: Understanding Chinese family use of AI to support decision making of college applications. arXiv preprint arXiv: 2411.10280.

[3] Klimashevskaia, A., Jannach, D., Elahi, M., & Trattner, C. (2024). A survey on popularity bias in recommender systems. User Modeling and User-Adapted Interaction, 34(5), 1777-1834.

[4] Qiao, L., Tang, Z., & Zhang, W. (2024). Dilemmas and solutions in major selection for college entrance examination [高考志愿填报中专业选择的困境及其化解]. Journal of Education and Teaching Research, 38(3), 52–67.

[5] Adel, A., & Alani, N. (2025). Can generative AI reliably synthesise literature? Exploring hallucination issues in ChatGPT. AI & Society, 1–14.

[6] Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. arXiv preprint arXiv: 2401.11817.

DOI: 10.54254/2755-2721/2025.LD27290

[7] Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., ... & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. Journal of Medical Internet Research, 26(1), e53164.
[8] Wang, X., & Cheng, G. (2024). A survey on extractive knowledge graph summarization: Applications, approaches, evaluation, and future directions. arXiv preprint arXiv: 2402.12001.
[9] Yang, X., & Tan, L. (2022). The construction of accurate recommendation model of learning resources of knowledge graph under deep learning. Scientific Programming, 2022(1), 1010122.
[10] Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., ... & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology, 55(1), 90–112.
[11] Carver, J. C., Kendall, R. P., Squires, S. E., & Post, D. E. (2007, May). Software development environments for scientific and engineering software: A series of case studies. In 29th International Conference on Software Engineering (ICSE'07) (pp. 550–559). IEEE.
[12] Li, Y., Wang, S., Ding, H., & Chen, H. (2023, November). Large language models in finance: A survey. In Proceedings of the fourth ACM international conference on AI in finance (pp. 374-382).
[13] Zhang, Y., Chen, X., Jin, B., Wang, S., Ji, S., Wang, W., & Han, J. (2024). A comprehensive survey of scientific large language models and their applications in scientific discovery. arXiv preprint arXiv: 2406.10833.
[14] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv: 2108.07258.
[15] Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., & Wang, P. (2020, April). K-bert: Enabling language representation with knowledge graph. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 03, pp. 2901-2908).