The Societal Benefits of AI Interventions and Intelligent Technology

The Societal Benefits of AI Interventions and Intelligent Technology

Edited by

Joan Lu and Qiang Xu

Cambridge
Scholars
Publishing

The Societal Benefits of AI Interventions and Intelligent Technology

Edited by Joan Lu and Qiang Xu

This book first published 2026

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-0364-7006-7
ISBN (Ebook): 978-1-0364-7007-4

TABLE OF CONTENTS

CHAPTER ONE

CHALLENGES AND DEVELOPMENTS OF LARGE LANGUAGE MODELS FOR MATHEMATICAL COMPUTATION, REASONING AND SPECIALIZATION

JINWEN MA[1,2], JO-KU CHENG[2] AND JINGYANG DENG[2]
[1]INSTITUTE OF SYSTEMS SCIENCE, BEIJING WUZI UNIVERSITY, BEIJING, 101149, CHINA
[2]SCHOOL OF MATHEMATICAL SCIENCES, PEKING UNIVERSITY, BEIJING, 100871, CHINA

Abstract. Ever since the release of ChatGPT, the transformer-based Large Language Models (LLMs) have shown tremendous performance improvements and achieved great success in Natural Language Processing (NLP) as well as in Artificial Intelligence Generated Content (AIGC). However, the key challenges of an LLM in its further developments and applications are the low abilities of mathematical computing and reasoning, respectively. Moreover, a big knowledge gap exists between the general-purpose and domain-specific applications. In this chapter, we investigate these three challenges and try to improve the large language model to achieve the higher performance in mathematical computing and reasoning by adding certain inference mechanisms, adopting the multi-modal architectures, and eliminating the knowledge gap. Specifically, we analyze the Probability Inference mechanism of LLMs and discuss how to make an LLM achieve mathematical computation. To improve its mathematical reasoning, we propose a Diagram Formalization Enhanced Multi-Modal LLM for geometry problem solver. Finally, on the specialization of an LLM, we present a specialized LLM in the State-owned Assets and Enterprises Domain.

Keywords: Large Language Model (LLM); Artificial Intelligence Generated Content (AIGC); Mathematical Computation; Reasoning; Specialization.

1. Introduction

With the development of computer technology, Natural Language Processing (NLP) has attracted great attention from the field of Artificial Intelligence. It aims to utilize computational linguistic methods to enable computers to perform tasks such as automatic translation, summarization, classification, and other forms of language processing involving text and speech data. However, due to the complexity of natural language, it is rather difficult for computer programs to complete these tasks effectively and efficiently. A feasible way of NLP is to understand the language from a mathematical perspective. Mathematically, a language model (LM) typically represents the probability distribution of a natural language. The significant statistical LMs were introduced and developed in the two decades spanning the 1980s and 1990s, being helpful for many tasks like machine translation (MT), natural language generation (NLG), optical character recognition (OCR), and handwriting recognition [1]. They also played an important role in grammar induction, information retrieval (IR), and speech recognition.

From 2006 to 2017, the introduction and rapid development of deep learning techniques led to their integration into NLP for processing large-scale data. As a result, deep learning-based language models were proposed and widely applied to tackle complex NLP tasks, achieving state-of-the-art performance. However, these deep learning-based LMs also encounter certain challenges, including processing rare or previously unseen words, overfitting during training, and difficulty in capturing complex linguistic phenomena [2].

To overcome these challenges, starting from 2017, the Generative Pre-Trained Transformer (GPT) LMs as well as the other large language models were developed, such as ChatGPT, GPT-4, and GPT-4o, leveraging large transformer architecture with a huge number of parameters. These large language models (LLMs) are distinguished by their capacity to perform various NLP tasks, including text generation and classification [3]. Actually, they gain their capabilities through a rigorous training process that involves learning statistical relationships from extensive text corpora through both self-supervised and semi-supervised approaches. In this way, they develop quickly and promote artificial

intelligence generated content (AIGC) so that many required texts and documents can be generated through LLMs.

As LLMs have obtained the top performances on the tasks of NLP, their developments are quickly beyond NLP, especially being extended to image as well as video processing through Vision Transformer (ViT) architecture [4]. Specifically, it still utilizes transformer architecture but processes the images through patch embedding, position embedding, transformer encoders and MLP Heads. In such a way, we can establish a large model to process the images in a similar way as that of LLMs. On the other hand, we can also establish a Multi-modal LLM that is still based on the transformer architecture but use both texts and visual inputs (e.g., images or videos) as well as certain formalization languages inputs [5]. As two or more kinds of information are merged in the large model, its learning performance can be improved remarkably. In fact, the multi-modal LLMs have been developed to solve geometry problems and obtain better results.

In China, LLMs have developed very quickly and there are many foundation models which are widely used in practical applications. In the following, we list the main large language models released in China:

1. Wenxin Large Models (Baidu, Inc.);
2. Tongyi Qianwen (Qwen) Large Models (Alibaba (China) Co., Ltd);
3. Hunyuan Large Models (Tencent Holdings, Ltd);
4. Pangu Large Models (Huawei Technologies Co., Ltd);
5. Doubao Large Models (Douyin Co., Ltd);
6. Xinghuo Large Models (iFLYTEK Co., Ltd);
7. MathGPT (Jiuzhang Large Model) (TAL Education Group);
8. DeepSeek-R1 (Hangzhou DeepSeek Artificial Intelligence Co., Ltd).

All of these large language models are free and open to society. Everyone in the world can load and use it. It should be noted that even though DeepSeek-R1 was just released in January 2025, it has already attracted great attention for its strong reasoning capabilities and low computation cost. Currently, DeepSeek-R1 has been widely applied and further developed in China.

Although LLMs have quickly developed and are widely applied to various AI problems, there are three major challenges. The first is mathematical computation. In fact, although the current LLMs are able to solve many

complicated problems like language translation and text generation, they cannot make a mathematical computation as a computer or even a cell phone does. The second is mathematical reasoning. Certain Large Reasoning Models (LRMs) like DeepSeek-R1 have made certain progress on mathematical reasoning like solving geometry problems, but the results need to be greatly improved on both answer accuracy and the reasoning process. The third challenge is domain specialization. It is still difficult to learn the domain-specific knowledge without forgetting the general knowledge.

In this chapter, we try to investigate and analyse the critical problems of these challenges and propose new strategies and methods to address them. In Section 2, we begin to analyse the reasons for inaccurate mathematical computation of LLMs and suggest certain strategies to make correct or accurate computation. We then improve the reasoning ability by incorporating a formalization language description of a geometric diagram into a multi-modal LLM for solving geometry problems in Section 3. We further investigate the domain specialization problem of the China State-owned Assets and Enterprises and present the new methods and models on this specific domain in Section 4. Finally, we make a brief conclusion and certain future remarks in Section 5.

2. Probability Inference and Mathematical Computation

It is clear that LLMs as well as other machine learning models are trained with a set of random samples, and therefore use a strategy of probability inference to get the result according to the statistical learning theory [6]. In this way, they do not use mathematical logic or rules to perform mathematical computations for mathematical problems. Hence, the computing result of an LLM is able to be wrong with a positive risk. That is, it cannot always be correct or accurate. Actually, we can observe that as the computational problem is complex to a certain degree, any LLM cannot get the correct or accurate result at all.

According to the above analysis, we can find out that the current LLMs cannot make the correct or accurate mathematical computation as a computer or even a cell phone does in our daily life. In order to enable an LLM to get the true result of mathematical computation, the underlying inference mechanism should be modified. The most possible strategy is to adopt Python or any other programming language to make the mathematical computation for a requested computational problem. That is,

when a computational problem is questioned into the LLM system, it should understand that this computational problem is a kind of computation task that can be written into a Python program, and then run the Python program to get the correct or accurate result. This strategy seems reasonable and can be easily realized in the LLM system since the LLM system also runs on the Python computing platform. However, there still exist some critical problems. (1). It is still difficult for an LLM system to decide which problem is a computational problem with a computation task that can be computed step by step (2). An LLM system can write an effective Python program for some simple computational tasks, but as the computational task becomes complex, it is rather difficult for an LLM system to write an effective Python program. (3). There may be many computational problems questioned synchronously at a certain time, which can delay or destroy the system because of the momentary tremendous computation load with those computational tasks. With the continuous progress and improvement of LLMs, these problems may be improved step by step. However, there is still a long way to the correct or accurate mathematical computation along this direction.

The other possible strategy is to combine a system of mathematical computation logic rules with the LLM system and perform reasoning based on logical inference and output the correct or accurate result. That is, the LLM system should have a hybrid inference framework such that it can get the logic result in the determinate situations, but the possible prediction in the random situations through probability or Bayesian inference. It is clear that our wise brains of human being actually use this kind of hybrid inference rule. However, it is rather difficult to realize the hybrid inference rule in the LLM system since the difference between the determinate and indeterminate problems is not so clear from the database. As long as the LLMs can truly understand the meaning of the problems, they can use the hybrid inference rule to solve the determinate and indeterminate problems differently and select the correct result, and then the LLM system can make the correct or accurate mathematical computation.

3. Mathematical Reasoning and Geometry Problem Solver

We further consider the challenge of mathematical reasoning. Actually, it is rather difficult for an LLM to make mathematical reasoning during solving a mathematical problem such as a geometry one because the current LLMs is effective to learn the answer for a question (i.e., a pair of

input and output data), but poor to learn the reasoning process on solving the mathematical problem. As the LLMs use a probability inference rule in each step of the reasoning process, they may suffer severe hallucination that greatly decreases both answer accuracy and inferring process correctness. Recently, great efforts have been made to improve the mathematical reasoning performance of LLMs. Like DeepSeek-R1-Zero, it utilizes large-scale Reinforcement Learning (RL) without Supervised Fine-Tuning (SFT). We can also strengthen the reasoning ability with the mechanisms of long Chain of Thoughts (CoTs), self-verification, etc. Here, we try to use the Multi-Modal LLMs (MLLMs) to improve the performance of reasoning, especially using a certain formalization language input of the diagram as a kind of modality for the geometry problem solver in the following subsections.

3.1 Geometry Problem Solving

As a typical reasoning task, geometry problem solving (GPS) is very challenging in AI research [7,8]. Some GPS methods typically rely on symbolic reasoning or textual descriptions, but these approaches often struggle to incorporate the visual aspects inherent in geometry, such as geometric diagrams and figures [9]. In order to overcome this weakness, the multi-modal models have been developed to make the reasoning over both text and diagrams simultaneously, effectively solving geometry problems involving angles, shapes, and relationships, which provides new opportunities for personalized education. Recent advancements in the LLMs and vision-based models have led to the rise of MLLMs, combining the strengths of NLP and computer vision [10]. However, these models have shown certain limitations when applied to GPS, particularly when tasked with understanding and reasoning about complex geometric diagrams [11]. The key reason is that vision encoders are often pretrained on natural images, which differ significantly from the abstract and formalized visual structures commonly found in geometric diagrams.

Our research addresses the challenges in solving geometry problems with MLLMs by introducing the Diagram Formalization Enhanced Geometry Problem Solver (DFE-GPS) [12], which incorporates diagram formalization to enhance the model's ability to understand geometric relationships. This model introduces new methods for interpreting geometric diagrams and generating step-by-step solutions.

3.2 MLLMs

MLLMs have shown remarkable progress in integrating visual and linguistic modalities to perform a wide range of tasks, such as image captioning, visual question answering (VQA), and multimodal dialogue. Representative models like GPT-4, Flamingo, and Kosmos-1 have demonstrated that combining visual encoders with large-scale pretrained language models can yield powerful cross-modal reasoning capabilities. For instance, GPT-4 supports image and text inputs in a unified interface, achieving human-level performance on many academic benchmarks. Flamingo connects a frozen vision encoder to a pretrained language model via learnable adapters, enabling strong few-shot learning in vision-language tasks.

Despite these advances, existing MLLMs struggle with tasks requiring precise geometric reasoning and symbolic manipulation. Most current models rely heavily on natural image datasets and free-form language corpora, which lack structured mathematical or geometric knowledge. Moreover, while MLLMs can describe diagrams or answer basic visual questions, they often fail to capture formal geometric constraints such as congruence, symmetry, or deductive steps found in textbook problems. Recent efforts like PaLM-E begin to explore embodied visual-language reasoning, but their focus remains on perception and control rather than structured mathematical reasoning.

These limitations motivate the research that goes beyond conventional MLLMs by introducing structured representations, formal languages, and domain-specific supervision to enhance models' reasoning over geometry problems.

3.3 MLLM-based GPS Models

In the era of LLMs, more researchers are exploring the reasoning abilities of the models, examining the mathematical and logical thinking capabilities of large models to solve mathematical problems. Geometry problem solving requires handling both textual and visual information, making MLLMs a key testing criterion for model reasoning ability. The current models using multimodal large models for geometry problem-solving can be divided into two categories. One category relies on geometry solvers [13], while the other category directly outputs human-readable solutions to geometry problems without using additional solving tools.

The first category includes works like GeoX [14], which outputs formal language reasoning steps via an MLLM pretrained with formalized geometric language, followed by a geometric solver to produce the final result. However, this solution process is unreadable, making its application challenging. The second category includes works like G-LLaVA [15], which directly output natural language reasoning steps. G-LLaVA utilizes GPT-3.5 for data expansion based on geometric problem characteristics, using a dataset of over 170K geometric diagram-caption and question-answer pairs. The dataset includes cross-modal geometry data, where the LLM generates textual descriptions (captions) for images based on problem questions and answers. It also includes prompts for questions about relationships between geometric elements and uses previously generated text to have the LLM explain the diagram details. Instruction tuning data is used for equation solving, numerical scaling, condition reconstruction, sentence rewriting, etc., to expand the existing question-answer dataset. The training phase uses standard LLaVA methods for geometric visual-language alignment, training the visual encoder and language model projection layers, and further fine-tuning them with LLM via geometric instruction tuning to enhance the model's problem-solving ability. However, there are issues with the diagrams in the dataset. On the text side, the correctness of results generated by GPT-3.5 based on text modification related to numerical data is hard to guarantee. Modifications to the text are not reflected in the original geometry images, which can cause confusion during the learning process.

3.4 SynthGeo228K

The key limitation of the existing geometry models is the lack of a large-scale, high-quality dataset specifically designed for multi-modal geometric problem solving. To address this issue, we introduced SynthGeo228K [12], a synthetic dataset that combines geometric diagrams with both formal and natural language descriptions. Inspired by FormalGeo [16], we adopted the Conditional Declaration Language (CDL) as the formal representation framework. In total, SynthGeo228K comprises over 228,000 diagram-text pairs, offering a rich and diverse resource to support training and evaluation of multi-modal models in complex geometry reasoning tasks.

Based on the Geometry Model Building Language (GMBL), we designed a wide variety of geometric composition templates, enabling the large-scale generation of geometric diagrams. Unlike randomly assembling geometric elements, our approach generated the images with clear

geometric meaning, better assisting models in understanding geometric relationships. We started with basic geometric elements such as points, line segments, and fundamental triangles (e.g., isosceles, equilateral, right angle triangles), as well as other shapes (e.g., rectangles, trapezoids, circles, regular pentagons), gradually increasing the complexity of the compositions. Ultimately, three levels of geometric elements were incorporated into the construction sequences.

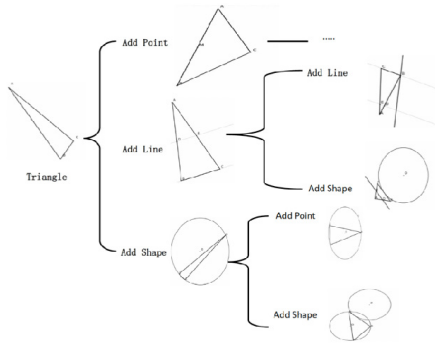

Fig. 1. SynthGeo228K Show Case.

As shown in Figure 1, we use triangles as an example to demonstrate part of the evolutionary process starting from a triangle. We attempted to add new elements such as points, lines, and shapes to see whether they could form new geometric figures. For instance, we can add point elements to a triangle such as the midpoints of sides, the excentre, centroid, circumcentre, and incentre. When adding lines, based on their positional relationships with the triangle, they may be disjoint or intersecting. In the intersecting case, the new line could be parallel to a side, a perpendicular bisector, or an angle bisector, among others. When adding new shapes such as triangles, quadrilaterals, or circles various geometric relationships arise.

On this basis, we can further enrich the constructions. For example, starting with a triangle with a line parallel to one of its sides, we can add a new circle, and so on. Following this logic, we can evolve many more templates from the existing ones. During the template construction process, we continuously introduced new elements and systematically planned the overall design. Additionally, we translated all 462 templates into both a formal language (ConstructionCDL and ImageCDL) and detailed natural

language descriptions (captions). Each generated image incorporates some randomness, as the initial points are randomly selected during the generation process.

3.5 DFE-GPS

With the above dataset SynthGeo228K, we further established a novel framework called the Diagram Formalization Enhanced Geometry Problem Solver (DFE-GPS) [12]. The overview flowchart of DFE-GPS is given in Fig.2.

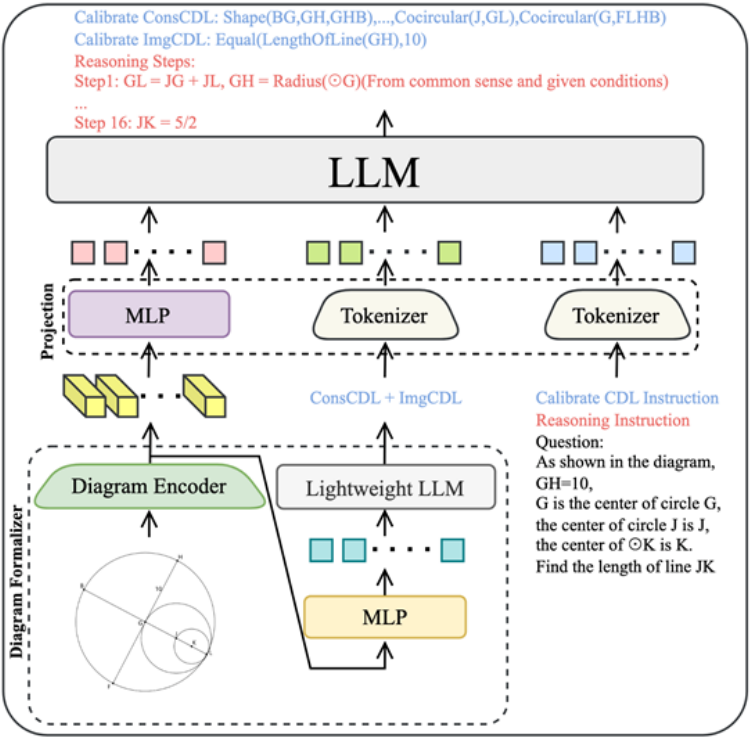

Fig. 2. Overview Flowchart of DFE-GPS [12].

Besides synthetic data generation methods shown as above, the DFE-GPS model hinges on the following key innovations:

(a) Multi-Modal Input

The system processes three types of input: image-based features from diagrams, formal symbolic representations of geometric relationships, and natural language problem descriptions. These diverse inputs are projected into a shared semantic space, allowing the model to use them cohesively for reasoning tasks.

(b) Diagram Formalization Component

A specialized module is trained to translate visual diagrams into structured formal language, encoding geometric relationships (like angles, lengths, and shape properties). This representation supplements the raw visual data and helps the model reason symbolically.

(c) Three-Stage Training Regimen

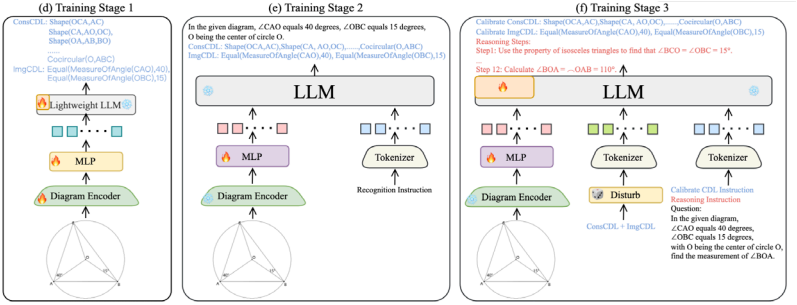

Fig. 3. Three-Stage Training Regimen [12].

The training process is divided into three focused stages:

Stage 1: Train the diagram formalization model to generate formal language representations from diagrams.

Stage 2: Learn how to align visual and linguistic features through a projection module, enabling meaningful multi-modal interactions.

Stage 3: Fine-tune the model to perform reasoning tasks, including correcting formal representations and generating problem-solving steps in natural language.

(d) Evaluation Metrics

Beyond the traditional accuracy, we further design a new metric, the Process Evaluation Score (PES), which evaluates the generated reasoning steps based on correctness, logical structure, and clarity. This is performed using an auxiliary LLM.

The core evaluation of the DFE-GPS model focused on its ability to solve geometry problems, both in multiple-choice and open-ended formats. In the multiple-choice mode, where the model generates a solution and selects the correct answer from a set of provided options, the DFE-GPS-34B model achieved an accuracy of 82.38%. In contrast, the open-ended mode, which requires the model to generate a solution independently without predefined options, yielded an accuracy of 75.33%. These results highlight the model's capability to handle both simple and more complex problem-solving scenarios.

To provide a comprehensive understanding of the model's performance, we compared DFE-GPS with other state-of-the-art models in the domain of geometry problem solving, including both single-modal and multi-modal models. In multiple-choice tasks, DFE-GPS-34B outperformed most other models, including DeepSeek-Prover-V1.5 and Yi-VL series. When handling open-ended problems, DFE-GPS exhibited a notable advantage over other models, which often struggled with the more complex tasks.

3.6 Discussions

Comprehension of geometry involves not only the ability to solve problems but also the skills to visualize geometric relationships and the ability to create new problems. Building on this, it can foster an engaging educational environment.

Personalized education requires not only detailed solutions to problems but also a diverse set of questions tailored to meet the varying learning needs of individual students. With the advancement of unified methods, it is foreseeable that they can be applied in education to deliver customized learning content. In the future, we can explore generating new problems with accompanying diagrams, aligned with specific knowledge points to better support differentiated instruction.

4. Specialization of LLMs in the State-owned Assets and Enterprises Domain

The domain specialization of LLMs is very important and valuable for practical applications in the era of AI. We investigate this challenge through our research of specialization of LLMs in the state-owned assets and enterprises domain of China in the following subsections.

4.1 Domain Specialization

Recently, LLMs are quickly developed and widely used across industries, with their natural language processing capabilities driving innovation in healthcare [17,18], finance [19], law [20], and beyond. However, adapting these models to a typical ecosystem of the State-Owned Assets and Enterprises (SOAEs) of China leads to certain distinct challenges rooted in the sector's complexity and specificity. In fact, the SOAEs operate across the critical sectors such as energy, transportation, and manufacturing, where work-flows are shaped by specific regulatory frameworks, industry terminology, and compliance requirements. Unlike the general domains, SOAEs face inherent data limitations: sparse labeled datasets, uneven data quality, and a lack of task diversity that hinder the effective training of LLMs. Moreover, the need to balance broad language proficiency with deep domain expertise creates a dual challenge: off-the-shelf LLMs often struggle to interpret policy nuances or align with organizational values, leaving a gap in AI-driven solutions tailored to SOAE needs.

Against this backdrop, our research initiatively seeks to develop specialized LLMs that address these bottlenecks through a three-pronged approach: robust data infrastructure, innovative training methodologies, and end-to-end model optimization. By focusing on the unique constraints and opportunities of the SOAE domain, our goal is to enable these enterprises to leverage AI for digital transformation and operational efficiency. In the following, we present our main strategies and methods on how to specialize an LLM into the domain of SOAEs.

4.2 Specialization Strategies

(a) Data Construction and Domain-Specific Ecosystem Development

High-quality, contextually relevant data are essential for training the LLMs that understand SOAE workflows. We have created a multi-

dimensional data ecosystem that reflects the sector's unique characteristics. This includes a pretraining corpus compiled from millions of tokens of SOAE-related documents, such as policy manuals and operational reports, structured to capture industry-specific terminology and regulatory frameworks. Complementing this are fine-tuning datasets annotated for SOAE-specific tasks like policy compliance analysis and strategic suggestion extraction. This ecosystem addresses data scarcity and heterogeneity through rigorous curation, removing noise, standardizing for-mats, and enriching datasets with case studies. As detailed in [21], this ecosystem is further anchored by a benchmarking framework that evaluates models against the metrics like policy interpretation accuracy and cross-departmental communication coherence, ensuring the relevance to real-world SOAE challenges.

(b) Adaptive Training Strategies for Domain Alignment

General LLMs often fail to meet the SOAE requirements due to mismatches in domain knowledge and organizational values. To address this, we developed a hybrid training strategy that merges domain adaptation with preference alignment. It begins with domain-adaptive pre-training, where models are exposed to SOAE-specific knowledge graphs and policy documents to build sector-specific expertise while retaining general language capabilities via a replay mechanism that mitigates catastrophic forgetting. Selective supervised fine-tuning (SFT) then refines models on task-specific datasets, while innovative techniques like Kahneman-Tversky Optimization enable the strategic use of low-quality data by accounting for decision-making biases inherent in organizational contexts. This ensures that the model outputs align with the SOAE regulatory standards and operational priorities, transforming general LLMs into tools capable of generating contextually appropriate, policy-compliant responses. The experimental results in [22] highlight the strategy's effectiveness, showing improved accuracy in tasks like policy analysis compared to traditional training methods.

(c) Holistic Model Optimization for Performance and Efficiency

Developing the SOAE-specific LLMs requires overcoming technical trade-offs between specialization and generalization, data efficiency, and inference speed. Our proposed strategy in [22,23] introduces a comprehensive optimization framework designed to enhance both the performance and efficiency of LLMs in the SOAE domain. This framework emphasizes the integration of domain knowledge throughout

the model development process while preserving the models' general language capabilities. By adopting a progressive learning strategy during the fine-tuning phase, the models are gradually exposed to tasks of increasing complexity, which helps in building their adaptability across different types of tasks. Furthermore, the framework incorporates an advanced decoding method to accelerate the inference process for documents with long contexts, ensuring that the models can provide timely responses without sacrificing the quality of their outputs.

4.3 Main Achievements

(a) SOAEs-DataSuite

The foundation of our research lies in the careful preparation and preprocessing of data tailored to the SOAE domain. In [21], we compiled an extensive pre-training corpus called SOAEs-DataSuite by drawing on a diverse array of domain-specific data sources, including but not limited to policy documents, operational reports, and news articles. This corpus, which contains approximately 19.57 billion tokens, serves as a rich foundation for the model pretraining. To ensure the quality and usability of the data, we implemented a multi-step preprocessing pipeline. This pipeline includes content parsing to extract text from various formats, deduplication to remove redundant information, and cleaning to eliminate noise and irrelevant content. For PDF files, which often contain embedded images, we used the advanced OCR technology to extract text while preserving the original document structure. For the Word files, we utilized specialized libraries to convert the content into a uniform text format. These processes collectively ensure that the data is well-structured and suitable for training high-quality LLMs.

In fact, SOAEs-DataSuite is a cornerstone in the development of specialized LLMs for the SOAE domain. It provides a comprehensive collection of data resources meticulously designed to address the unique challenges of SOAEs. The suite includes a diverse range of data types, each carefully selected and processed to meet the specific needs of SOAE applications. The pre-training corpus within the suite offers extensive exposure to the terminologies and contexts unique to SOAEs, ensuring that models are well-versed in the linguistic nuances and information density characteristic of SOAE communications. This exposure is crucial for developing models that can understand and generate contextually appropriate responses within the SOAE environment.

The fine-tuning datasets included in SOAEs-DataSuite further enhance the models' capabilities by providing annotated data for various tasks such as extracting critical information from policy documents and classifying operational texts. These datasets are specifically designed to improve the models' precision and contextual understanding, which are essential for real-world SOAE applications. The benchmark component of the suite establishes a standardized evaluation protocol tailored to the SOAE domain. It employs a mix of evaluation metrics, including accuracy, BLEU, and ROUGE scores, to provide a comprehensive assessment of model performance across different task types. This enables reliable measurement and comparison of various LLMs in addressing SOAE-specific challenges, thereby facilitating the advancement of specialized LLM development.

(b) Hybrid Training Strategy for Enhancing Domain Adaptation

The key challenge in applying a general LLM to the SOAE domain is the need to bridge the gap between general language capabilities and domain-specific requirements. Our proposed approach in [22] involves a combination of domain-adaptive pretraining and innovative training strategies, which is a kind of hybrid training strategy for enhancing the SOAE domain adaption. During the domain-adaptive pretraining phase, we expose the model to a large volume of SOAE-specific data, enabling it to acquire familiarity with domain-specific terminology and contexts. To address the issue of catastrophic forgetting, where the model may lose the previously learned knowledge, we incorporate a replay mechanism that interleaves general domain data with SOAE-specific data. This ensures that the model retains its general language abilities while adapting to the SOAE domain. Then, we employ a selective supervised fine-tuning strategy, where a portion of high-quality domain-specific data is used to further refine the model's performance on specific tasks. Additionally, we introduce a novel data scheduling approach that leverages both high-quality and low-quality data. By integrating low-quality data through the Kahneman-Tversky Optimization technique, we make effective use of data that might otherwise be discarded, thereby enhancing the model's robustness and adaptability to real-world conditions.

This hybrid training strategy effectively bridges the gap between general LLMs and the specialized requirements of SOAEs by addressing two critical problems: the knowledge gap and value disagreement. Through domain-adaptive pretraining, the model is immersed in a rich array of SOAE-specific data, allowing it to build a comprehensive understanding

of the domain's unique knowledge base. This phase is crucial for equipping the model with the necessary domain knowledge to perform specialized tasks effectively.

The integration of a replay mechanism during the domain-adaptive pretraining phase ensures that the model retains its general language capabilities while acquiring new domain-specific knowledge. This mechanism prevents catastrophic forgetting, preserving the model's versatility and enabling it to handle a broader range of tasks. The subsequent selective supervised fine-tuning sharpens the model's performance on specific SOAE tasks by focusing on high-quality, task-relevant data. This targeted approach enables the development of specialized skills while maintaining efficiency, ensuring that the model can deliver accurate and reliable responses to domain-specific queries.

Additionally, our hybrid strategy incorporates low-quality data through the Kahneman-Tversky Optimization technique. This innovative approach expands the effective training data pool and helps the model learn to navigate the complexities of real-world data. By utilizing data that might otherwise be discarded, we enhance the model's robustness and adaptability, making it better equipped to handle the varied and sometimes messy nature of SOAE data. This holistic approach ensures that the resulting LLMs are accurate, reliable, and contextually appropriate, making them powerful tools for SOAE applications.

(c) SOAEsV2 Model Series: Balancing Specialization and Versatility

To enhance both the performance and efficiency of LLMs in the SOAE domain, we established the SOAEsV2 models through a comprehensive model optimization mechanism in [22, 23]. Actually, this mechanism emphasizes the gradual integration of domain knowledge into the model while preserving its general language capabilities. The model optimization process begins with continual pretraining, where the model is progressively exposed to SOAE-specific data, allowing it to build a comprehensive understanding of the domain. Following this, the model undergoes a domain-progressive fine-tuning phase, where it is systematically introduced to tasks of increasing complexity using a curriculum-based approach. This enables the model to develop specialized skills in a stepwise manner. Finally, to address the challenge of efficient inference, the mechanism incorporates a distillation-enhanced decoding method. This approach leverages knowledge distillation techniques to

accelerate the processing of long-context documents, ensuring that the models can deliver timely and accurate responses in real-time applications.

The established SOAEsV2 models represent a significant step forward in developing LLMs for the SOAE domain. They are designed to balance specialization and versatility through the optimization mechanism that focuses on gradual knowledge integration and efficient inference. The optimization process includes continual pretraining to maintain general language capabilities while adapting to SOAE-specific tasks, and a domain-progressive fine-tuning approach that exposes the model to increasingly complex tasks. This enables the model to handle both routine and strategic tasks effectively. Additionally, the model employs an advanced decoding method to speed up the processing of long documents, ensuring timely responses in real-time applications.

4.4 Discussions

Our research has established a systematic framework for developing the LLMs that address the specialization of SOAE domain through tailored data ecosystems, adaptive training strategies, and end-to-end model optimization. SOAEs-DataSuite, the hybrid training strategy, and the SOAEsV2 models collectively provide a robust foundation for creating AI tools that drive digital innovation in regulated industries. These advancements enhance the capacity of SOAEs to leverage AI for operational excellence and offer transferable methodologies for domain-specific LLM development worldwide. Therefore, our research presents a good direction to meet the challenge of LLM domain specialization.

Conclusions and Remarks

We have investigated and analysed the three major challenges of LLMs on their developments and practical applications in the society. It is clear that LLMs have made a new era of AI from natural language processing to general intelligent information processing. They have already changed the world as well as our daily life. They will quickly develop and be widely applied to various fields of AI. However, there are still many challenges for the developments of LLMs. Particularly, mathematical computation, reasoning and specialization are three major challenges to the practical applications. According to our investigations and analyses, we can meet these challenges in certain ways and improve the performances of LLMs with the suggested strategies and mechanisms. According to the

experiments and applications, LLMs can get the top results in many tasks and perform well in various practical situations. Nevertheless, there has been short of systematic theory on the LLM design and performance analysis, which really limits their further developments and applications. We should try our best to establish the theoretical foundation for LLMs as soon as possible. On the other hand, LLMs need the huge computation resource that will prevent their wide applications. The simplification of the LLMs on both the model structure and training process is another research direction in the future.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under the grant 62471175.

References

1. Rosenfeld R.: Two decades of statistical language modelling: where do we go from here? Proc IEEE 88(8):1270–1278, 2000.
2. Annepaka Y., Pakray P.: Large language models: a survey of their development, capabilities, and applications. Knowledge and Information Systems 67: 2967-3022, 2025.
3. Vaswani S.: Attention is all you need. In: Advances in neural information processing systems, 2017.
4. Dosovitskiy A., Beyer L., Kolesnikov A., et al.: An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv 2020, arXiv:2010.11929.
5. Bai S., Chen K., Liu X., et al.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923.
6. Vapnik V.: The Nature of Statistical Learning Theory. Springer-Verlag,1999.
7. Seo M., Hajishirzi H., Farhadi A., et al.: Solving geometry problems: Combining text and diagram interpretation. In: Màrquez, L., Callison-Burch, C., Su, J. (eds.) Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1466–1476. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015).
8. Lu P., Gong R., Jiang S., et al.: Inter-GPS: Interpret-able geometry problem solving with formal language and symbolic reasoning. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language

Processing (Volume 1: Long Papers). pp. 6774–6786. Association for Computational Linguistics, Online (Aug 2021).
9. Trinh T.H., Wu Y., Le Q.V., et al.: Solving Olympiad geometry without human demonstrations. Nature 625(7995), 476–482, 2024.
10. Bai S., Chen K., Liu X., et al.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923.
11. Zhang J., Li Z., Zhang M., et al.: Geoeval: benchmark for evaluating llms and multi-modal models on geometry problem-solving. arXiv preprint arXiv:2402.10104 (2024).
12. Zhang Z., Cheng J.K., Deng J., et al.: Diagram formalization enhanced multi-modal geometry problem solver. In: 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP,2025). pp. 1–5 (2025).
13. Chen J., Tang J., Qin J., et al.: GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Lin-guistics: ACL-IJCNLP 2021. pp. 513–523. Association for Computational Linguistics, Online (Aug 2021).
14. Xia R., Li M., Ye H., et al.: Geox: Geometric problem solving through unified formalized vision-language pre-training (2025), https://arxiv.org/abs/2412.11863.
15. Gao J., Pi R., Zhang J., et al.: G-llava: Solving geometric problem with multi-modal large language model (2023), https://arxiv.org/abs/2312.11370.
16. Zhang X., Zhu N., He Y., et al.: Formalgeo: An extensible formalized framework for Olympiad geometric problem solving (2024), https://arxiv.org/abs/2310.18021.
17. Bedi S., Liu Y., Orr-Ewing L., et al.: Testing and evaluation of health care applications of large language models: a systematic review. JAMA 33(4): 319-328,2025.
18. Nazi Z.A., Peng W.: Large language models in healthcare and medical domain: A review. Informatics. vol. 11, p. 57. MDPI (2024).
19. Nie Y., Kong Y., Dong X., et al.: A survey of large language models for financial applications: Progress, prospects and challenges. arXiv preprint arXiv:2406.11903 (2024).
20. Ariai F., Demartini G.: Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. arXiv preprint arXiv:2410.21306 (2024).
21. Huang, J., Deng, J., Ma, J.: SOAEs-DataSuite: Tailored pre-training corpus, fine-tuning dataset and benchmark for large language models in the state-owned assets and enterprises domain. In: 2024 IEEE 17th

International Conference on Signal Processing (ICSP). pp. 253–257. IEEE (2024).
22. Deng J., Zhang Z., Cheng J.K., et al.: Enhancing large language models on domain-specific tasks: A novel training strategy via domain adaptation and preference alignment. In: 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP,2025). pp. 1–5. IEEE (2025).
23. Deng J., Chen R., Cheng J.K., et al. Soaesv2-7b/72b: Full-pipeline optimization for state-owned enterprise llms via continual pre-training, domain-progressive SFT and distillation-enhanced speculative decoding. arXiv preprint arXiv:2505.04723 (2025).

CHAPTER TWO

MULTIMODAL ASPECT-BASED SENTIMENT CLASSIFICATION ON STRUCTURAL ENHANCEMENT AND TEXT-IMAGE ALIGNMENT FUSION[1]

FENG ZHANG[1], HUA LONG[2] AND YUBIN SHAO[1]
[1]FACULTY OF INFORMATION ENGINEERING AND AUTOMATION, KUNMING UNIVERSITY OF SCIENCE AND TECHNOLOGY, KUNMING 650500, YUNNAN, P. R. CHINA
[2]YUNNAN COMMUNICATIONS VOCATIONAL AND TECHNICAL COLLEGE, KUNMING 650500, YUNNAN, P. R. CHINA

Abstract: To address the insufficient utilization of structural information in text-image modalities and the semantic mismatch in cross-modal interactions in existing multimodal aspect-based sentiment classification methods, we propose a Structural Enhancement and Text-Image Alignment Fusion model (SETIAF), which utilizes RoBERTa to capture context and aspect representations, and Faster R-CNN to extract image features. Based on a Graph Attention Network, the proposed model enhances the structural information of both context and image representations by leveraging the syntactic structure of text and the spatial positioning of images. The semantic alignment of text and image based on aspects is achieved through Cross-Transformer and contrastive learning. Furthermore, Gate-Transformer is used to fuse modality information and enhance the consistency of modality features through gated dynamic alignment. Experiments on the Twitter-2015 and Twitter-2017 datasets demonstrate that the SETIAF model achieves aspect-based sentiment classification accuracies of 78.21% and 72.93%, with corresponding F1

[1] These authors contributed equally to this work.

scores of 73.88% and 72.25%, showing further performance improvement compared to existing baseline models.

Keywords: Multimodal Aspect-based Sentiment Classification; Graph Attention Network; Contrastive Learning; Gate-Transformer

Introduction

Aspect-based Sentiment Classification (ASC) is a more precise classification task within sentiment analysis, where the objective is to determine the sentiment orientation of a specific aspect within a sentence. These aspects are commonly entities or attributes of entities mentioned in the sentence. Consider this example from the Twitter 2017 dataset: "*Kevin Durant has more points (23) than the Splash Bros combined (22).*" Here, the aspect "*Kevin Durant*" has a positive sentiment, while "*Splash Bros*" has a negative sentiment. The proliferation of contemporary multimedia platforms has driven users to integrate multimodal content when conveying aspect-specific opinions, positioning Multimodal Aspect-based Sentiment Classification (MASC) as a critical research frontier in affective computing. This task focuses on determining the polarity of sentiment for aspects in particular situations by integrating textual and visual data information. Earlier research primarily centered on the interaction among textual modality, visual modality, and aspects. Nevertheless, image information is not always highly relevant to the aspects referenced in the text, so direct interaction may bring in interference into the learning process. Additionally, since visual and textual representations are generated through distinct pretraining frameworks, the semantic gap between them may further impact the model's accuracy.

In response to the aforementioned challenges, we put forward a structure-enhanced and image-text alignment fusion model (SETIAF), which enhances the structural information of both image and text, facilitates aspect-based image-text semantic alignment, and applies dynamic gating to align the multimodal fusion representations. Our primary contributions include:

1. The syntactic structure of the textual modality and the spatial structure of the visual modality are modeled, while a graph attention network is employed to obtain structurally enhanced representations for both text and images.
2. A Cross-Transformer is employed to obtain aspect-based image and text representations, followed by a multimodal contrastive

optimization strategy to align semantic information across the visual and textual features.
3. The Gate-Transformer is designed, which incorporates a dynamic gating mechanism into the Cross-Transformer to suppress noise expression in multimodal fusion features.
4. Experimental comparisons using two widely adopted Twitter datasets indicate that the proposed model surpasses baseline models, showcasing enhanced performance.

Related work

Aspect-based sentiment classification

In early research, researchers attempted to construct syntactic structures and combine them with machine learning models to classify aspect-level sentiments. However, the limited accuracy of conventional syntactic parsers constrained the performance advantages of such approaches. Recent advances in neural network-based syntactic dependency parsing have significantly improved the quality of generated syntactic trees, thereby enhancing the reliability of syntactic features for ASC tasks and yielding substantial performance improvements. Sun et al.[1] proposed the CDT model, which utilizes Bi-LSTM to learn word representations and performs graph convolution operations on the dependency syntax tree to facilitate the flow of information from opinion words to aspect words. Zhang et al.[2] introduced the SSEGCN modal, which integrates syntactic structures into semantic representations using a mask matrix. By constructing a syntactic dependency tree, the distance between aspect words and opinion words is effectively shortened, allowing the model to capture richer and more precise semantic information.

Multimodal aspect-based sentiment classification

Compared to the ASC task, the MASC task can integrate relevant information from other modalities to supplement and verify the text content, thereby enhancing classification performance. In the field of multimodal research, there is an emphasis on the interaction of features and the extraction of important information from images. In 2019, Yu et al.[3] proposed the TomBERT architecture, which obtains target-related image representations through a Cross-Transformer but fails to tackle image noise. In 2021, Yu et al.[4] proposed an attention-based fusion framework, ESAFN, which facilitates interactions within modalities