Fundamental Capabilities of Large Language Models and their Applications in Domain Scenarios: A Survey

Jiawei Li[1,2] Yizhe Yang[1,2] Yu Bai[1,2] Xiaofeng Zhou[1,2] Yinghao Li[1,2] Huashan Sun[1,2] Yuhang Liu[1,2] Xingpeng Si[1,2] Yuhao Ye[1,2] Yixiao Wu[1,2] Yiguan Lin[1,2] Bin Xu[1,2] Bowen Ren[1,2] Chong Feng[1,3] Yang Gao[1,3]* Heyan Huang[1,3]

[1] School of Computer Science and Technology, Beijing Institute of Technology

[2] Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications

[3] Beijing Institute of Technology Southeast Academy of Information Technology

{jwli, yizheyang, Yubai, fengchong, gyang, hhy63}@bit.edu.cn

Abstract

Large Language Models (LLMs) demonstrate significant value in domain-specific applications, benefiting from their fundamental capabilities. Nevertheless, it is still unclear which fundamental capabilities contribute to success in specific domains. Moreover, the existing benchmark-based evaluation cannot effectively reflect the performance of real-world applications. In this survey, we review recent advances of LLMs in domain applications, aiming to summarize the fundamental capabilities and their collaboration. Furthermore, we establish connections between fundamental capabilities and specific domains, evaluating the varying importance of different capabilities. Based on our findings, we propose a reliable strategy for domains to choose more robust backbone LLMs for real-world applications.

1 Introduction

In the current research and application of artificial intelligence, the abundant acquisition of big data, breakthroughs in high-performance computing technology, and innovations in algorithm design have jointly promoted the development and deployment of LLMs (Li, 2022). LLMs are also considered to have strong potential value in specific domains, with an increasing number of industries embracing LLMs and already demonstrating outstanding performance (Gururangan et al., 2020; Ling et al., 2023b; Kaddour et al., 2023).

However, applying LLMs in specific domains has encountered a series of challenges. These challenges mainly stem from the inherent characteristics of domain tasks and data, such as the diversity of data sources, the complexity of domain-specific knowledge, and the specificity of application goals and constraints. To enable LLMs to be better applied in specific domains and address the challenges they face in these areas, this paper summarizes two key issues that need to be resolved when applying LLMs in specific domains:

Issue 1: Fundamental capabilities of LLMs and their interactions. LLMs exhibit outstanding performance in comprehending and addressing complex tasks, thus demonstrating great potential in specific domain applications. Numerous studies broadly summarize the core capabilities of LLMs as robust comprehension and generation (Huang and Chang, 2023; Ling et al., 2023a), yet fall short of assisting us in aligning the LLMs' fine-grained capabilities with the intricate requirements of real-world scenarios. Consequently, elucidating the inherent fundamental capabilities manifested by LLMs in domain-specific scenarios and the dynamics among these capabilities becomes essential.

Issue 2: The Capabilities Assessment of LLMs in Specific Domain. Due to the disparity between the capabilities evaluated in benchmarks and those required in real-world domains (Ling et al., 2023a; Kaddour et al., 2023), the excellent performance of LLMs in benchmarks may not necessarily translate to actual applications in specific domains. Therefore, conducting capabilities assessments of LLMs to establish a bridge between benchmarks and real-world domains is crucial.

Based on the above two issues, this survey aims to systematically summarize the fundamental capabilities of LLMs and clarify the capabilities assessment of LLMs. The key contributions of this survey paper are summarized below.

1. This paper summarizes the fundamental capabilities of LLMs in domain applications, including memorization, reasoning, generalization, and diversification. It provides detailed descriptions of each capability and how they collaborate to accomplish specific applications.

* Corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11116–11141

2. This paper summarizes the applications of LLMs in nine specific domains from the perspective of real scenarios. In addition, this paper summarizes the importance of fundamental abilities corresponding to each domain, addressing the issue of strong performance on benchmarks not necessarily translating to domain scenarios, and providing users with clear model selection strategies.

2 Fundamental Capabilities and Interactive Capabilities

The human brain's information processing has been extensively studied, revealing five core modules: natural language interaction, knowledge, memorization, reasoning, and generalization (Xi et al., 2023). Some research suggests that LLMs exhibit a similar information processing mechanism to the human brain (Toneva and Wehbe, 2019; Caucheteux and King, 2022). Consequently, we categorize LLM information processing into four fundamental capabilities including memorization, reasoning, generalization, and diversification. As illustrated in Figure 1, LLMs utilize short-term memory to understand task instructions and long-term memory to retrieve historical data (Cowan, 2008; Norris, 2017; Zhu et al., 2020; Zhang et al., 2023g; Cheng et al., 2016; Davis and Marcus, 2015; Dawid and LeCun, 2023). This data is processed through the reasoning module, which performs logical, commonsense, and symbolic reasoning to generate outputs (Lu et al., 2023b; Bursztyn et al., 2022; Yan et al., 2023; Banerjee et al., 2021; Hamilton et al., 2022). Throughout this process, the generalization enables LLMs to manage information across varying lengths, structures, and tasks (Davis and Marcus, 2015; Lake and Baroni, 2023; Zhao et al., 2021a), while diversification allows for tailored results (Sun et al., 2024).

The four fundamental capabilities of the LLM work together to complete complex domain applications. In the following section, we will introduce each of the fundamental capabilities in detail.

2.1 Memorization Capabilities

The memorization capabilities of LLMs play a pivotal role in their effectiveness and performance across various domains. Memory, in the context of LLMs, refers to the capacity to retain and access information over time. It can be broadly categorized into two types: long-term memory and short-term memory. Long-term memory refers to the LLM's ability to store and recall knowledge, facts, and concepts acquired during training and previous experiences. It encompasses the model's understanding of world knowledge, implicit encoding of substantial information, and its capacity to leverage both internal and external knowledge sources. On the other hand, short-term memory focuses on the LLM's in-context learning capabilities and its ability to retain and utilize information within a limited temporal context. Enhancements in short-term memory aim to overcome the contextual limitations of LLMs and enable them to generate coherent content over longer stretches.

2.1.1 Long-term Memory

The long-term memory of LLMs is intimately connected to their scale, with LLMs showcasing broader knowledge capacity and diversity. Benchmarks like KoLA critically evaluate LLMs' world knowledge across numerous tasks (Yu et al., 2023c), while studies by Tirumala et al. (2022) reveal that model size is crucial for efficient memorization. LLMs acting as knowledge bases, as reviewed by AlKhamissi et al. (2022), encode substantial information implicitly, and innovations like REALM employ latent retrievers to augment this knowledge (Guu et al., 2020). Petroni et al. (2019) suggest LLMs can serve as effective knowledge bases even without fine-tuning. Together, these studies underscore the remarkable potential of LLMs to use both internal and external knowledge sources to enhance their long-term memory, applicable across various fields.

Addressing the challenge of preserving the long-term memory of LLMs during continual learning is crucial for their application in specialized fields. Luo et al. (2023c) propose a novel framework, SCCL, which mitigates catastrophic forgetting—a common obstacle to maintaining long-term knowledge—by employing adaptive classification strategies alongside memory replay and distillation techniques. Complementing this, Luo et al. (2023d) suggests that initial training on general linguistic tasks and the adoption of a hybrid continual learning strategy can substantially reduce the loss of long-term syntactic and semantic knowledge.

2.1.2 Short-term Memory

Short-term memory in LLMs has been a focus area to enhance their in-context learning (ICL) capabilities. ICL is a capability that allows LLMs to understand and execute tasks based on the immedi-

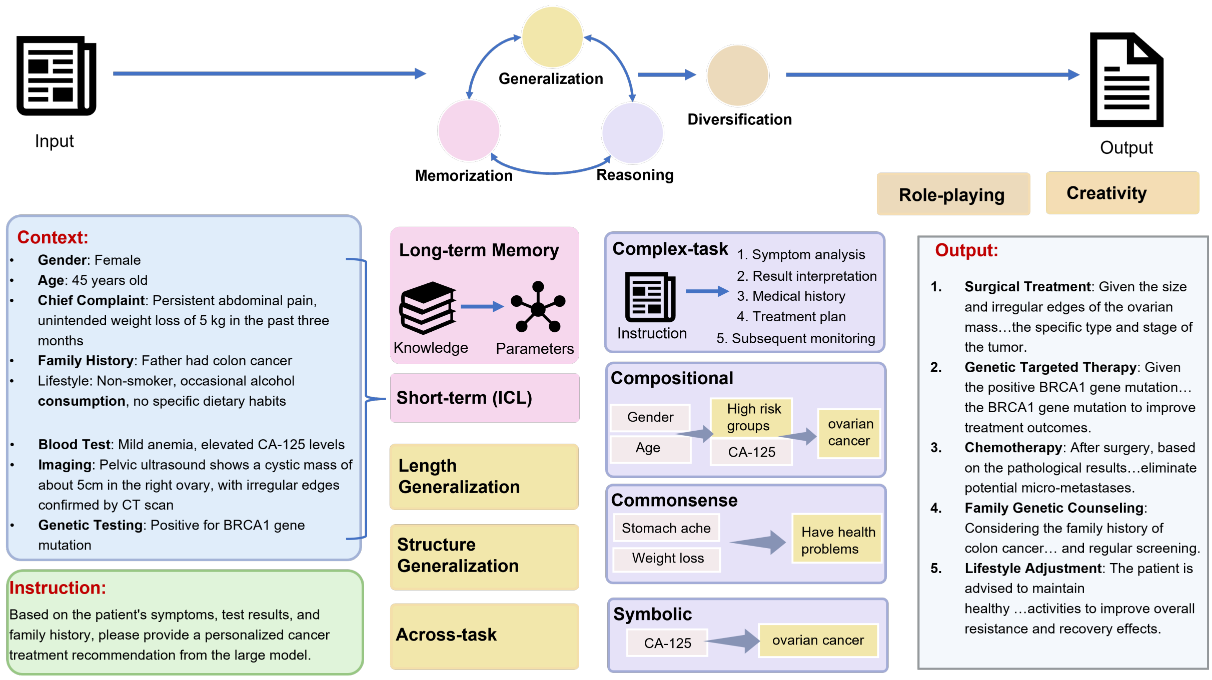

Figure 1: The relationship between the four fundamental abilities when solving complex domain applications. Take medical treatment as an example.

ate context provided within the input text (Brown et al., 2020; Dong et al., 2022). This skill set eliminates the need for extensive retraining or fine-tuning across different tasks. By analyzing a few examples included in their input, LLMs can infer the task requirements and apply learned patterns to generate accurate responses (Zhao et al., 2021b). This in-context learning ability showcases LLMs' adaptability, making them highly efficient for a broad spectrum of applications with minimal setup (Mosbach et al., 2023). Despite its effectiveness, this approach has limitations in terms of the depth and complexity of understanding it can provide, which is directly influenced by the model's design and the richness of the context provided (Lu et al., 2022; Zhang et al., 2022b). Meanwhile, the working mechanism of ICL is also a widely open question and has been investigated a lot by the commuity (Min et al., 2022; Liu et al., 2021; Olsson et al., 2022; Bhattamishra et al., 2023).

2.2 Reasoning Capabilities

The reasoning capabilities of LLMs refer to logically process information, draw conclusions, and make decisions based on available data and knowledge (Qiao et al., 2023). The reasoning capabilities of LLMs have greatly enhanced their application across various industries. For example, they apply commonsense reasoning to user interactions in customer service and healthcare, providing contextually relevant responses. Additionally, advances in symbolic reasoning allow LLMs to support software development and mathematical fields with increased accuracy and clarity. These developments mark significant progress toward more sophisticated AI systems capable of augmenting human tasks. In this section, we summarize the recent advances in the reasoning capabilities of LLMs.

2.2.1 Compositional Reasoning Capabilities

Recent advancements demonstrate that augmenting LLMs with specialized modules or training approaches enhances their compositional reasoning capabilities, surpassing traditional methods in diverse and complex tasks. Lu et al. (2023c) augments LLMs with modules for complex reasoning, achieving significant accuracy improvements on multi-modal tasks. Chen et al. (2023c) propose a novel prompting strategy, skills-in-context (SKiC), enabling LLMs to exhibit compositional reasoning by solving unseen, complex problems through the innovative composition of pre-existing skills, achieving groundbreaking success on compositional tasks. Ma et al. (2023) introduces new benchmarks for evaluating GVLMs' compositional reasoning, with a novel metric to reduce morphological bias. Compositional Task Representations (CTR), a new prompt-free approach, is proposed to learn compositional codes, surpassing prompt-based methods in zero-shot learning (SHAO et al., 2023). An LLM trained on the PLANE benchmark shows strong capacities in compositional entailment, leveraging subword representations(Bertolini et al., 2022).

2.2.2 Complex Task Decomposition

A variety of new prompting techniques such as ADaPT, chain of thought, zero-shot CoT, iterative context-aware, least-to-most, decomposed, and successive prompting have been proposed to enable LLMs to decompose and tackle complex tasks more effectively (Zhou et al., 2022; Drozdov et al., 2023; Khot et al., 2022; Dua et al., 2022). Prasad et al. (2023) introduces ADaPT, which enhances LLMs' decision-making by planning and decomposing tasks as needed, significantly improving performance on complex tasks. Wei et al. (2022) demonstrates that "chain of thought" prompting boosts LLMs' complex reasoning, achieving state-of-the-art results on the GSM8K benchmark. Kojima et al. (2022) presents Zero-shot-CoT, using simple prompts to unlock LLMs' underlying reasoning capabilities, achieving substantial gains across diverse reasoning tasks. Wang et al. (2022a) proposes an iterative prompting framework that progressively extracts PLMs knowledge for multi-step reasoning, overcoming the limitations of traditional prompting methods.

2.2.3 Commonsense Reasoning

Recent work in commonsense reasoning explores innovative approaches like integrating LLMs with search algorithms, conducting comprehensive surveys, and applying code generation models to outperform traditional methods, while also exposing the limitations of LMs in truly understanding commonsense knowledge without specific supervision. Bhargava and Ng (2022) surveys recent tasks in commonsense reasoning and generation, evaluating the capabilities and limitations of state-of-the-art pre-trained models. Zhao et al. (2023d) demonstrates that combining LLMs with MCTS for task planning leverages commonsense knowledge to enhance reasoning and efficiency in complex tasks. Madaan et al. (2022) proposes using code generation LMs for structured commonsense reasoning tasks, outperforming traditional LMs in natural language processing. Li et al. (2022) conducts a zero-shot and few-shot evaluation of LMs' commonsense knowledge, revealing limitations and the insufficiency of larger models to reach human-level performance.

2.2.4 Symbolic Reasoning

Recent work on symbolic reasoning highlights the effectiveness of novel prompting techniques and hybrid frameworks combining LLMs with symbolic solvers or distillation methods. Gaur and Saunshi (2023) explores symbolic reasoning in math word problems using LLMs, introducing a self-prompting method that aligns symbolic reasoning with numeric answers, enhancing interpretability and accuracy. Wei et al. (2022) demonstrates that chain of thought prompting significantly improves LLMs' performance on complex reasoning tasks including symbolic reasoning. Pan et al. (2023a) introduces Logic-LM, a framework that combines LLMs with symbolic solvers, resulting in substantial improvements in logical reasoning tasks. Gaur and Saunshi (2022) shows that GPT-3's performance on symbolic math word problems can be enhanced with specific prompting techniques that encourage the model to describe its reasoning process. Li et al. (2023g) reveals that even smaller models can benefit from chain-of-thought prompting through Symbolic Chain-of-Thought Distillation from larger models, leading to improved reasoning performance.

2.3 Generalization Capabilities

Generalization refers to a model's ability to apply learned knowledge from past experiences to new, unseen situations (Elangovan et al., 2021; Shen et al., 2021). This capability is essential for real-world applications where models encounter various data. In this section, we will explore the generalization capabilities of LLMs, focusing on three key aspects: length, structural, and across-task generalization.

2.3.1 Length Generalization

Length generalization in LLMs, which refers to the model's capacity to extend acquired skills to longer problem instances outside the training range, is crucial for addressing complex problems with extensive descriptions (Anil et al., 2022). Improving length generalization is key to enhancing the practical use of LLMs in diverse real-world situations.

Theoretical insights, such as Anil et al. (2022) and Xiao and Liu (2023), have examined the length generalization capabilities of transformer-based models and identified conditions for length generalization in reasoning tasks. For arithmetic tasks, Jelassi et al. (2023) introduced train set priming to improve generalization while the innovative LM-Infinite (Han et al., 2023), RASP-Generalization Conjecture (Zhou et al., 2023a) and Attention Bias Calibration (ABC) (Duan and Shi, 2023) employ different variant of attention mechanism for longer text generation. On the other hand, Awasthi and Gupta (2023) focused on multitask training with task hinting to address length generalization.

These studies collectively present a range of strategies, from practical methodologies like task hinting to theoretical frameworks, enhancing LLMs' ability to manage longer input sequences. They mark significant progress in overcoming the challenges of length generalization, paving the way for more capable and adaptable LLMs.

2.3.2 Structure Generalization

Structure generalization of LLMs refers to the capability to process and interpret complex data structures, such as graphs and tables even though the models are trained on text-only datasets. This ability is crucial for applications extending beyond traditional text-based tasks, spanning various domains including bioinformatics and social network analysis.

Numerous studies have aimed at enhancing the capabilities of LLMs to process and generate diverse data forms beyond traditional text, including graphs (Tang et al., 2023b; Guo et al., 2023a; Pan et al., 2023b; Zhang et al., 2023h; Liu et al., 2023b; Zhang et al., 2023f; Wang et al., 2023b,b), tables (Zhao et al., 2023a), and visualization charts (Wang et al., 2023c). This expansion into handling various data types is particularly notable in fields such as healthcare (Thirunavukarasu et al., 2023), recommendation (Wang et al., 2023c), question answering (Pan et al., 2023b; Jiang et al., 2023a), and biomedical science (Wang et al., 2023b; Qian et al., 2023), significantly broadening the practical applications of LLMs.

These studies collectively underscore the expanding versatility of LLMs in handling structured data, revealing a trend toward more sophisticated AI models capable of complex reasoning and diverse applications.

2.3.3 Generalization Across Tasks

Task generalization in Large Language Models (LLMs) refers to their ability to manage a wide range of tasks, especially those not seen during training. This capacity allows the models to tackle a variety of novel and unexpected challenges, showcasing their flexibility, efficiency, and versatility.

To enhance task generalization in LLMs, two prevalent strategies are employed. Firstly, fine tuning approaches, such as multi-task (Sanh et al., 2022), instruction tuning (Wei et al., 2021), and meta tuning (Zhong et al., 2021), fine-tune language models across various NLP tasks to augment their comprehension of instructions, thereby achieving significant zero-shot learning capabilities. These methods highlight LLMs' potential in managing an array of tasks through enhanced instruction understanding. However, fine-tuning parameters of large language models can be resource-intensive. In response, Ye et al. (2023) and Brown et al. (2020) investigate few-shot or in-context learning mechanisms. By providing a few task examples, LLMs can infer the task's requirements and format, allowing them to address new tasks effectively. This approach circumvents the need for extensive fine-tuning, instead leveraging examples to foster the models' abilities.

In summary, these studies highlight the generalization of LLMs through diverse methodologies, ranging from fine-tuning to prompting strategies. The overarching objective is to enhance models that not only perform proficiently on familiar tasks but also demonstrate remarkable adaptability to novel challenges, thereby creating more adaptable and intelligent systems.

2.4 Diversification Capabilities

The concept of diversification in LLMs pertains to their capability to produce unique content tailored to various contexts. This diversification arises during the inference process, where a model generates a new token, y_t, based on the previously generated tokens, y_{t-1}, and a specific condition, x, according to the formula $y_t \sim p(y_t|y_{:t-1}, x)$. Notably, the architecture of most LLMs is decoder-only, meaning that conditions such as prompts or in-context examples are incorporated as initial tokens. Thus, we regard these initial inputs as x and the sequence of generated tokens as $y_{:t-1}$. By manipulating these inputs and conditions, LLMs can produce a wide array of content.

We delve into the diversification of the expanding capabilities of LLMs in terms of role-playing and creativity. These two areas highlight the versatility of LLMs, showcasing their ability to adapt to diverse scenarios and tasks. Role-playing enhances the dynamism and context-awareness of LLMs, enabling more nuanced interactions across different scenarios by utilizing role profiles as the condition x.Furthermore, creativity plays a crucial role in unlocking the potential of LLMs for generating innovative and valuable content. This is achieved by modifying the generation process, specifically the sequence of previously generated tokens, y_{t-1}.

2.4.1 Role-playing

Role-playing in LLMs represents a significant advancement in the field of natural language processing and artificial intelligence. It involves LLMs assuming specific characters or personas, enabling them to engage in more dynamic, context-rich, and human-like interactions. By embodying different roles, i.e. x we defined before, LLMs can offer tailored responses based on character-specific knowledge and behavior patterns, enhancing the relevance and engagement of user interactions.

Wei et al. (2023a) investigate multi-party conversations, revealing that LLMs can significantly improve group dynamics when trained on datasets like MultiLIGHT. Wang et al. (2023f)'s RoleLLM framework enhances role-playing in LLMs, leading to advancements in English and Chinese models. Shanahan et al. (2023) discusses the importance of role play in understanding dialogue agents' behaviors, focusing on aspects like deception and self-awareness. Li et al. (2023a) develop ChatHaruhi, demonstrating enhanced role-playing in mimicking anime characters. Personalization in LLMs is the focus of Salemi et al. (2023)'s LaMP benchmark, which improves model outputs by incorporating user profiles. Finally, Li et al. (2023c) explore autonomous cooperation among LLMs through role-playing, showcasing the potential of inception prompting in multi-agent systems.

These studies collectively represent a significant stride in enhancing the role-playing capabilities of LLMs. They demonstrate how role-playing can transform LLMs into more adaptable, engaging, and effective conversational partners, capable of nuanced interactions across various domains by adjusting the condition x.

2.4.2 Creativity

The creativity in LLMs is gaining traction, emphasizing their potential to generate novel and valuable content. Emphasizing creativity in LLMs is key to developing AI systems that not only replicate human language but also exhibit a degree of ingenuity akin to human creativity.

Recent studies in this area offer diverse insights. Chakrabarty et al. (2023) develop a framework for evaluating the creativity of LLMs, revealing their current limitations compared to human writers. Franceschelli and Musolesi (2023) explore LLMs' creative writing potential, examining their development through various creativity theories and considering their societal impact. Summers-Stay et al. (2023) demonstrate that LLMs can enhance their creativity by mimicking human brainstorming techniques. Swanson et al. (2021) introduce tools to assist creative writers in leveraging LLMs' capabilities, while Sinha et al. (2023) propose a model to balance creativity with factual accuracy in LLM outputs. Bhavya et al. (2023) focus on creative analogy mining using PLMs, underscoring the role of LLMs in augmenting human creativity.

2.5 Interactive Capabilities

In addition to the four fundamental capabilities, LLMs also possess strong interactive capabilities during domain applications. Interactive capabilities refer to the capacity of LLMs to enhance performance by acquiring external information, planning and making decisions regarding the environment, and utilizing external tools (Xi et al., 2023). For example, integrating specialized tools can overcome limitations of LLM in domain tasks (Qin et al., 2023b; Patil et al., 2023), while interaction with environments such as web pages, communities, and databases expands application domains (Yao et al., 2022; Team, 2023a). Appendix A provides a detailed overview of the interactive capabilities of large models.

3 The Capabilities Assessment of LLMs in Specific Domains

LLMs have different applications in different domain scenarios. For instance, in the medical and ledge domains, they may function as domain experts engaged in dialogues or summarizing documents (Shi et al., 2023; Tang et al., 2023a; Choi et al., 2023; Pettinato Oltz, 2023). Systematically summarizing the application methods of LLMs in various domains facilitates to combine these models with specific scenarios more efficiently. However, some research is often classified and summarized from the perspective of NLP tasks (Ling et al., 2023a; Kaddour et al., 2023). Kaddour et al.

Domains	Memorization	Reasoning	Generalization	Diversification	Interaction	Total
Medicine	16	15	12	10	5	25
Law	11	10	1	2	9	19
Computational Biology	17	20	17	13	2	20
Finance	21	21	3	4	5	24
Social and Psychology	6	20	2	3	4	22
Programming	12	21	1	3	1	23
Robots and Agents	22	22	3	13	6	22
AI for Disciplines	9	13	7	3	2	15
Creative Work	4	16	2	15	2	16

Table 1: We analyzed the fundamental capabilities across various domains in this table. For example, we analyzed 19 papers in the law domain. Among these 19 papers, 11 focused on memorization capabilities, 10 on reasoning capabilities, 1 on generalization capabilities, 2 on diversification capabilities, and 9 on interaction capabilities. Based on this table, we construct the radar chart in Figure 2.

(2023) classify applications in medical scenarios into medical question answering and comprehension, and medical information retrieval. However, LLMs may participate in medical diagnosis, diagnostic assistance, and other scenarios. The differences make it difficult for research results to be directly applied to real scenarios. We enumerates articles from nine domains we have summarized in Figure 3, including medicine, law, computational biology, finance, social sciences and psychology, computer programming and software engineering, robots and agents, AI for disciplines, and creative work. We will provide a detailed summary of the real-world applications and roles played by LLMs in these domains in the future.

3.1 Fundamental Capabilities Assessment and Application

The performance of LLMs in specific domains is closely related to their fundamental capabilities. However, we often evaluate LLMs based on their performance on benchmarks, but their strong performance on benchmarks may not necessarily translate to domain scenarios (Guo et al., 2023c; Zhou et al., 2023b). For example, while InstructBLIP exhibits outstanding performance in image caption tests, its performance significantly diminishes in online interactive evaluations closer to real-world scenarios (Dai et al., 2023).

Guo et al. (2023c) highlights that the range of model capabilities assessed by different benchmarks varies, leading to discrepancies between benchmarks and the model's performance in domain scenarios. Therefore, the quantitative assessment of fundamental capabilities within specific domains is crucial for users in choosing the most appropriate benchmarks. We employ a case study approach to conduct a case-by-case statistical analysis of the articles in Appendix B, deriving quantitative values for the important capabilities in each domain through expert evaluation. Taking the medical field as an example, among the 25 papers categorized for this study, methods involving memorization capabilities are present in 16 papers, reasoning capabilities in 15, generalization capabilities in 12, diversification capabilities in 10, and interactive capabilities in 5. Therefore, memorization capabilities emerge as the most critical in the medical domain, accounting for 64% of the focus. In table 1, we list the number of papers covered in each domain and analyze the fundamental capabilities demonstrated by these papers across various domains. Based on this data, we create radar charts for each domain.

As shown in Figure 2, we constructed radar charts to illustrate the relative importance of different fundamental capabilities in various domains. Based on these radar charts, researchers can quantify the differences between benchmarks and real scenarios. In the following chapters, we will introduce our selection strategy in medical and computer programming domains as examples.

3.1.1 Medical

According to the radar chart, memorization capabilities (64%) and reasoning capabilities (60%) are important fundamental capabilities in the medical domain. In scenarios like Medical Diagnosis and Knowledge Acquisition, LLMs need to engage in dialogues with patients using their medical knowledge. In this context, long-term memory related to the domain knowledge and reasoning capabilities to assist in answering questions are crucial for model performance. In scenarios of Diagnostic Assistance and Medical Report Generation, LLMs typically assist doctors in reading patient case in-

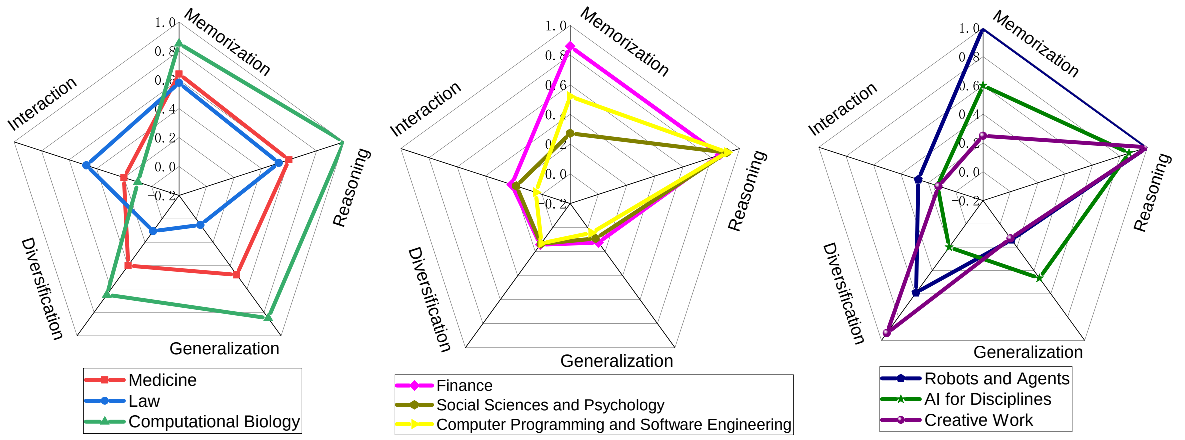

Figure 2: The radar charts of LLMs' fundamental capabilities in various domains.

MedQA	
Model	Acc(%)
LLAMA2-70B (Chen et al., 2023h)	61.5
FLAN-PaLM (Singhal et al., 2022)	67.6
Meditron-70B (Chen et al., 2023h)	70.2
Med-PaLM 2 (Singhal et al., 2023)	85.4
GPT4 (Nori et al., 2023b)	90.2
MedMCQA	
PubmedBERT (Pal et al., 2022)	41.0
BioMedGPT-10B (Luo et al., 2023b)	51.4
Codex (Liévin et al., 2022)	62.7
VOD (Liévin et al., 2023)	62.9
Med-PaLM 2 (Singhal et al., 2023)	72.3
PubMedQA	
BioGPT (Luo et al., 2022)	78.2
Flan-PaLM (Singhal et al., 2022)	79.0
Med-PaLM 2 (Singhal et al., 2023)	79.2
BioGPT-Large (Luo et al., 2022)	81.0
Meditron-70B (Chen et al., 2023h)	81.6

Table 2: The performance of different LLMs on MedQA, MedMCQA, and PubMedQA.

formation and generating treatment plans. In this context, short-term memory capabilities and reasoning capabilities play a crucial role.

Therefore, we recommend applying LLMs that excel in memorization capabilities and reasoning abilities benchmarks to the medical domain. Here, we provide three recommended medical benchmarks, with Table 2 summarizing the performance of different LLMs on these three benchmarks.

MedQA (Jin et al., 2021) is a medical text question and answer dataset in a multiple-choice format. It aims to test the professional knowledge and clinical decision-making abilities of LLMs. The examination of professional knowledge mainly targets the **memorization capabilities** of LLMs, while the assessment of clinical decision-making abilities primarily focuses on the **reasoning and generalization capabilities** of LLMs.

MedMCQA (Pal et al., 2022) is a large-scale multiple-choice question and answer dataset, with data sourced from All India Institute of Medical Sciences (AIIMS) and National Eligibility cum Entrance Test (NEET PG). Different from the MedQA dataset, besides directly examining the **memorization capabilities** of LLMs, the MedMCQA dataset includes detailed explanations for each answer, requiring LLMs to possess deep language **reasoning capabilities**.

PubMedQA (Jin et al., 2019) is a biomedical question answering dataset collected from PubMed abstracts. The task involves generating an answer in a multiple-choice format of yes/no given a question. This dataset demands **reasoning** over biomedical research texts, especially the capabilities to understand and analyze quantitative content, in order to answer questions.

In order to demonstrate the effectiveness of our proposed method for selecting robust backbone LLM, we chose the medical domain as a case study to further illustrate the practicality and effectiveness of our approach.

As outlined in above, we identiy memorization capabilities and reasoning capabilities as the most crucial fundamental capabilities in the medical domain. Based on this, we recommend models that perform well on benchmarks (MedQA, MedMCQA, and PubMedQA) focused on these capabilities. Through this process, we discover that Med-PaLM 2 excels in these benchmarks. This finding

Model	pass@1
XwinCoder-34B (Team, 2023b)	75.6
Magicoder-6.7B (Wei et al., 2023b)	76.8
Coderllama-34B (Rozière et al., 2023)	77.4
WizardCoder-33B (Luo et al., 2023e)	79.9
DeepSeek-Coder-33B (Guo et al., 2024)	81.1
GPT-4-Turbo (OpenAI, 2023)	88.4
MBPP	
XwinCoder-34B (Team, 2023b)	67.7
Coderllama-70B (Rozière et al., 2023)	75.4
WizardCoder-33B (Luo et al., 2023e)	78.9
DeepSeek-Coder-33B (Guo et al., 2024)	78.7
GPT-4-Turbo (OpenAI, 2023)	83.5

Table 3: The performance of different LLMs on HumanEval and MBPP.

is consistent with the industry recognition that the model has received in the medical domain. Specifically, renowned organizations such as HCA Healthcare, BenchSci, Accenture, and Deloitte have deployed the Med-PaLM 2 model across various medical scenarios, validating its value in real-world applications. In contrast, although RobotGPT-30B and jianpeiGPT performed well on CMB benchmark, their performance in real-world applications does not match that of the former, further proving the effectiveness of our selection methods.

3.1.2 Computer Programming

According to the radar chart, reasoning capabilities (91%) are considered the most crucial skill in computer programming and software development, followed by memorization capabilities (52%). Since code is a symbolic, hierarchical, and logic-driven language commonly used for handling complex tasks, the reasoning capabilities of LLMs are applied in various code scenarios. Short-term memory helps LLMs understand requirements and gather contextual information in code generation and automatic program repair.

Evaluation criteria for programming-related tasks evolve from single-type code language and static metrics to multi-type code languages and metrics (Zhang et al., 2023i; Zan et al., 2023). Among these, evaluation standards involving multi-language, multi-type metrics require models to possess stronger reasoning capabilities. In this paper, we recommend HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) benchmarks as the basis for selecting LLMs for programming-related scenarios. Table 3 presents the performance of some LLMs on these benchmarks.

4 Conclusion

In this paper, we summarize the fundamental capabilities of LLMs in domain applications and illustrate how they collaborate. Simultaneously, we summarize the applications of LLMs in various domains from real-world perspectives. Furthermore, we outline the key capabilities emphasized in different domains, aiding users in more accurately applying LLMs in domain applications.

5 Limitations

LLMs have found extensive applications across various fields. Although we aim to summarize the applications of LLMs in all domains, our work does not claim to exhaustively cover all possible application scenarios. Furthermore, although every attempt was made to provide readers with a comprehensive overview of each domain, the literature cited and discussed in this document does not constitute a fully exhaustive collection.

Additionally, in assessing the fundamental capabilities focused on by LLMs in various domains, we conduct in-depth analyses of each piece of literature to manually identify these capabilities. However, it is possible that some other relevant works were overlooked, potentially leading to inaccuracies in our analysis.

Acknowledgements

We extend our sincere gratitude to the anonymous reviewers for their helpful feedback and the diligent efforts of the conference committees. This work has received partial funding from the Major Research Plan of the National Natural Science Foundation of China (Grant No.92370110) and the Joint Funds of National Natural Science Foundation of China (No.U21B2009).

References

Alaa Abd-alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, Padraig Mark Healy, Syed Latifi, Sarah Aziz, Rafat Damseh, Sadam Alabed Alrazak, and Javaid Sheikh. 2023. Large language models in medical education: Opportunities, challenges, and future directions. *JMIR Med Educ*, 9:e48291.

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, and Mengyuan Yan. 2022. Do as I can, not as I say: Grounding language in robotic affordances. *CoRR*, abs/2204.01691.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. Playing repeated games with large language models. *CoRR*, abs/2305.16867.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv: 2204.06031*.

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. 2022. Exploring length generalization in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 38546–38556. Curran Associates, Inc.

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. 2021. Program synthesis with large language models. *CoRR*, abs/2108.07732.

Pranjal Awasthi and Anupam Gupta. 2023. Improving length-generalization in transformers via task hinting.

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L. Yuille, Trevor Darrell, Jitendra Malik, and Alexei A. Efros. 2023. Sequential modeling enables scalable learning for large vision models. *CoRR*, abs/2312.00785.

Katie Bainbridge, Candace A. Walkington, Armon Ibrahim, Iris Zhong, Debshila Basu Mallick, Julianna Washington, and Richard G. Baraniuk. 2023. A case study using large language models to generate metadata for math questions. In *Proceedings of the Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation 2023 co-located with 24th International Conference on Artificial Intelligence in Education (AIED 2023), Tokyo, Japan, July 7, 2023*, volume 3487 of *CEUR Workshop Proceedings*, pages 34–42. CEUR-WS.org.

Pratyay Banerjee, Swaroop Mishra, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2021. Commonsense reasoning with implicit knowledge in natural language. In *3rd Conference on Automated Knowledge Base Construction, AKBC 2021, Virtual, October 4-8, 2021*.

Lorenzo Bertolini, Julie Weeds, and David Weir. 2022. Testing large language models on compositionality and inference with phrase-level adjective-noun entailment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4084–4100, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pretrained language models: A survey. *AAAI Conference on Artificial Intelligence*.

Satwik Bhattamishra, Arkil Patel, Phil Blunsom, and Varun Kanade. 2023. Understanding in-context learning in transformers and llms by learning to learn discrete functions. *arXiv preprint arXiv:2310.03016*.

Bhavya Bhavya, Jinjun Xiong, and Chengxiang Zhai. 2023. Cam: A large language model-based creative analogy mining framework. In *Proceedings of the ACM Web Conference 2023*, pages 3903–3914.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2023. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning? *arXiv preprint arXiv:2302.06100*.

Michael Bommarito II and Daniel Martin Katz. 2022. Gpt takes the bar exam. *arXiv preprint arXiv:2212.14402*.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools.

Boyd Branch, Piotr Mirowski, and Kory W. Mathewson. 2021. Collaborative storytelling with human actors and AI narrators. In *Proceedings of the Twelfth International Conference on Computational Creativity, México City, México (Virtual), September 14-18, 2021*, pages 96–101. Association for Computational Creativity (ACC).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Victor S. Bursztyn, David Demeter, Doug Downey, and Larry Birnbaum. 2022. Learning to perform complex tasks through compositional fine-tuning of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1676–1686. Association for Computational Linguistics.

Alex Calderwood, Noah Wardrip-Fruin, and Michael Mateas. 2022. Spinning coherent interactive fiction through foundation model prompts. In *Proceedings of the 13th International Conference on Computational Creativity, Bozen-Bolzano, Italy, June 27 - July 1, 2022*, pages 44–53. Association for Computational Creativity (ACC).

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2023. Art or artifice? large language models and the false promise of creativity. *arXiv preprint arXiv:2309.14556*.

Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022. Help me write a poem - instruction tuning as a vehicle for collaborative poetry writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6848–6863. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *CoRR*, abs/2308.07201.

Bo Chen, Xingyi Cheng, Yangli ao Geng, Shengyin Li, Xin Zeng, Bo Wang, Jing Gong, Chiming Liu, Aohan Zeng, Yuxiao Dong, Jie Tang, and Leo T. Song. 2023a. xtrimopglm: Unified 100b-scale pre-trained transformer for deciphering the language of protein. *bioRxiv*.

Bo Chen, Ziwei Xie, Jiezhong Qiu, Zhaofeng Ye, Jinbo Xu, and Jie Tang. 2023b. Improved the heterodimer protein complex prediction with protein language models. *Briefings Bioinform.*, 24(4).

Jiaao Chen, Xiaoman Pan, Dian Yu, Kaiqiang Song, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2023c. Skills-in-context prompting: Unlocking compositionality in large language models. *arXiv preprint arXiv: 2308.00304*.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, Defu Lian, and Enhong Chen. 2023d. When large language models meet personalization: Perspectives of challenges and opportunities.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, Yuanzheng Ci, Bin Li, Xiaokang Yang, and Wanli Ouyang. 2023e. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *CoRR*, abs/2304.02948.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023f. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *CoRR*, abs/2308.10848.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *CoRR*, abs/2211.12588.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Huimin Zheng, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, and Xiangmin Xu. 2023g. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *CoRR*, abs/2310.15896.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023h. MEDITRON-70B: scaling medical pretraining for large language models. *CoRR*, abs/2311.16079.

Zheng Chen. 2023. PALR: personalization aware llms for recommendation. *CoRR*, abs/2305.07622.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & deep learning for recommender systems. In *Proceedings* *of the 1st Workshop on Deep Learning for Recommender Systems, DLRS@RecSys 2016, Boston, MA, USA, September 15, 2016*, pages 7–10. ACM.

Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. 2023. Chatgpt goes to law school. *Available at SSRN*.

Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Charlotte Rochereau, George M. Church, Peter K. Sorger, and Mohammed Alquraishi. 2021. Single-sequence protein structure prediction using language models from deep learning. *bioRxiv*.

The UniProt Consortium. 2009. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research*, 38(suppl$_1$) : $D142 - -D148$.

Nelson Cowan. 2008. What are the differences between long-term, short-term, and working memory? *Progress in brain research*, 169:323–338.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Anna Dawid and Yann LeCun. 2023. Introduction to latent variable energy-based models: A path towards autonomous machine intelligence. *CoRR*, abs/2306.02572.

Yinlin Deng, Chunqiu Steven Xia, Chenyuan Yang, Shizhuo Dylan Zhang, Shujing Yang, and Lingming Zhang. 2023. Large language models are edge-case fuzzers: Testing deep learning libraries via fuzzgpt. *CoRR*, abs/2304.02014.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How ready are pre-trained abstractive models and llms for legal case judgement summarization? *arXiv preprint arXiv:2306.01248*.

Jingzhe Ding, Yan Cen, and Xinyuan Wei. 2023. Using large language model to solve and explain physics word problems approaching human level. *CoRR*, abs/2309.08182.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv: 2301.00234*.

Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2023. Self-collaboration code generation via chatgpt. *CoRR*, abs/2304.07590.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palme: An embodied multimodal language model. *CoRR*, abs/2303.03378.

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, O. Bousquet, and Denny Zhou. 2023. Compositional semantic parsing with large language models. *International Conference on Learning Representations*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *CoRR*, abs/2305.14325.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. *Conference on Empirical Methods in Natural Language Processing*.

Shaoxiong Duan and Yining Shi. 2023. From interpolation to extrapolation: Complete length generalization for arithmetic transformers.

Naoki Egami, Musashi Jacobs-Harukawa, Brandon M Stewart, and Hanying Wei. 2023. Using large language model annotations for valid downstream statistical inference in social science: Design-based semi-supervised learning. *arXiv preprint arXiv:2306.04746*.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization: quantifying data leakage in nlp performance evaluation. *arXiv preprint arXiv:2102.01818*.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas B. Fehér, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. 2020. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *bioRxiv*.

Zachary Englhardt, Richard Li, Dilini Nissanka, Zhihan Zhang, Girish Narayanswamy, Joseph Breda, Xin Liu, Shwetak N. Patel, and Vikram Iyer. 2023. Exploring and characterizing large language models for embedded system development and debugging. *CoRR*, abs/2307.03817.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. 2022. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *NeurIPS*.

Xiaomin Fang, Fan Wang, Lihang Liu, Jingzhou He, Dayong Lin, Yingfei Xiang, Xiaonan Zhang, Hua Wu, Hui Li, and Le Song. 2022. Helixfold-single: Msa-free protein structure prediction by using protein language model as an alternative. *CoRR*, abs/2207.13921.

Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun R. Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Layoutgpt: Compositional visual planning and generation with large language models. *CoRR*, abs/2305.15393.

Lorenzo Jaime Yu Flores, Heyuan Huang, Kejian Shi, Sophie Chheang, and Arman Cohan. 2023. Medical text simplification: Optimizing for readability with unlikelihood training and reranked beam search decoding. *CoRR*, abs/2310.11191.

Giorgio Franceschelli and Mirco Musolesi. 2023. On the creativity of large language models. *arXiv preprint arXiv:2304.00008*.

Guanghui Fu, Qing Zhao, Jianqiang Li, Dan Luo, Changwei Song, Wei Zhai, Shuo Liu, Fan Wang, Yan Wang, Lijuan Cheng, Juan Zhang, and Bing Xiang Yang. 2023. Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals. *CoRR*, abs/2308.15192.

Pablo Gainza, Freyr Sverrisson, F. Monti, Emanuele Rodolà, D. Boscaini, Michael M. Bronstein, and Bruno E. Correia. 2019. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17:184 – 192.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S^3: Social-network simulation system with large language model-empowered agents. *CoRR*, abs/2307.14984.

Mu Gao, Hongyi Zhou, and Jeffrey Skolnick. 2019. Destini: A deep-learning approach to contact-driven protein structure prediction. *Scientific Reports*, 9.

Vedant Gaur and Nikunj Saunshi. 2022. Symbolic math reasoning with language models. *2022 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pages 1–5.

Vedant Gaur and Nikunj Saunshi. 2023. Reasoning in large language models through symbolic math word problems. *arXiv preprint arXiv: 2308.01906*.

Mingyang Geng, Shangwen Wang, Dezun Dong, Haotian Wang, Ge Li, Zhi Jin, Xiaoguang Mao, and Xiangke Liao. 2024. Large language models are few-shot summarizers: Multi-intent comment generation via in-context learning.

Vladimir Gligorijević, P. Douglas Renfrew, Tomasz Kosciólek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C. Taylor, Ian Fisk, Hera Vlamakis, Ramnik J. Xavier, Rob Knight, Kyunghyun Cho, and Richard Bonneau. 2021. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12.

Tomas Goldsack, Zhihao Zhang, Chen Tang, Carolina Scarton, and Chenghua Lin. 2023. Enhancing biomedical lay summarisation with external knowledge graphs. *CoRR*, abs/2310.15702.

Lewis D. Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly T. Mai, Maria Vau, Matthew Caldwell, and Augustine Marvor-Parker. 2023. Susceptibility to influence of large language models. *CoRR*, abs/2303.06074.

Sophia Gu. 2023. Llms as potential brainstorming partners for math and science problems. *CoRR*, abs/2310.10677.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021. Long text generation by modeling sentence-level and discourse-level coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6379–6393. Association for Computational Linguistics.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. Deepseek-coder: When the large language model meets programming - the rise of code intelligence. *CoRR*, abs/2401.14196.

Jiayan Guo, Lun Du, Hengyu Liu, Mengyu Zhou, Xinyi He, and Shi Han. 2023a. Gpt4graph: Can large language models understand graph structured data ? an empirical evaluation and benchmarking.

Yue Guo, Zian Xu, and Yi Yang. 2023b. Is chatgpt a financial expert? evaluating language models on financial natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 815–821. Association for Computational Linguistics.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023c. Evaluating large language models: A comprehensive survey. *CoRR*, abs/2310.19736.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Z. Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *International Conference on Machine Learning*.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2022. Machine intuition: Uncovering human-like intuitive decision-making in GPT-3.5. *CoRR*, abs/2212.05206.

Kyle Hamilton, Aparna Nayak, Bojan Bozic, and Luca Longo. 2022. Is neuro-symbolic AI meeting its promise in natural language processing? A structured review. *CoRR*, abs/2202.12205.

Sil Hamilton. 2023. Blind judgement: Agent-based supreme court modelling with gpt. *arXiv preprint arXiv:2301.05327*.

Chi Han, Qifan Wang, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2023. Lm-infinite: Simple on-the-fly length generalization for large language models.

Lois Haruna-Cooper and Mohammed Ahmed Rashid. 2023. Gpt-4: the future of artificial intelligence in medical school assessments. *Journal of the Royal Society of Medicine*, 116(6):218–219. PMID: 37318843.

Joy He-Yueya, Gabriel Poesia, Rose E. Wang, and Noah D. Goodman. 2023. Solving math word problems by combining language models with symbolic solvers. *CoRR*, abs/2304.09102.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1772–1791. Association for Computational Linguistics.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *CoRR*, abs/2308.00352.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1049–1065. Association for Computational Linguistics.

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022. Inner monologue: Embodied reasoning through planning with language models. *CoRR*, abs/2207.05608.

Kwan Yuen Iu and Vanessa Man-Yi Wong. 2023. Chatgpt by openai: The end of litigation lawyers? *Available at SSRN*.

Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D. Bocarsly, Andres M. Bran, Stefan Bringuier, L. Catherine Brinson, Kamal Choudhary, Defne Circi, Sam Cox, Wibe A. de Jong, Matthew L. Evans, Nicolas Gastellu, Jerome Genzling, María Victoria Gil, Ankur K. Gupta, Zhi Hong, Alishba Imran, Sabine Kruschwitz, Anne Labarre, Jakub Lála, Tao Liu, Steven Ma, Sauradeep Majumdar, Garrett W. Merz, Nicolas Moitessier, Elias Moubarak, Beatriz Mouriño, Brenden Pelkie, Michael Pieler, Mayk Caldas Ramos, Bojana Rankovic, Samuel G. Rodriques, Jacob N. Sanders, Philippe Schwaller, Marcus Schwarting, Jiale Shi, Berend Smit, Ben E. Smith, Joren Van Heck, Christoph Völker, Logan T. Ward, Sean Warren, Benjamin Weiser, Sylvester Zhang, Xiaoqi Zhang, Ghezal Ahmad Zia, Aristana Scourtas, K. J. Schmidt, Ian T. Foster, Andrew D. White, and Ben Blaiszik. 2023. 14 examples of how llms can transform materials science and chemistry: A reflection on a large language model hackathon. *CoRR*, abs/2306.06283.

Samy Jelassi, Stéphane d'Ascoli, Carles Domingo-Enrich, Yuhuai Wu, Yuanzhi Li, and François Charton. 2023. Length generalization in arithmetic transformers.

Zhenlan Ji, Pingchuan Ma, Zongjie Li, and Shuai Wang. 2023. Benchmarking and explaining large language model-based code generation: A causality-centric approach. *CoRR*, abs/2310.06680.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023a. StructGPT: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251, Singapore. Association for Computational Linguistics.

Shuyang Jiang, Yuhao Wang, and Yu Wang. 2023b. Selfevolve: A code evolution framework via large language models. *CoRR*, abs/2306.02907.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2567–2577. Association for Computational Linguistics.

Nicolas Jonason, Luca Casini, Carl Thomé, and Bob L. T. Sturm. 2023. Retrieval augmented generation of symbolic music with llms. *CoRR*, abs/2311.10384.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh J. Ramanathan, Wei Xu, Byron C. Wallace, and Junyi Jessy Li. 2023. Multilingual simplification of medical texts. *CoRR*, abs/2305.12532.

Harshit Joshi, José Pablo Cambronero Sánchez, Sumit Gulwani, Vu Le, Gust Verbruggen, and Ivan Radicek. 2023. Repair is nearly generation: Multilingual program repair with llms. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI*

2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, pages 5131–5140. AAAI Press.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *CoRR*, abs/2307.10169.

Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Maheswaran Sathiamoorthy, Lichan Hong, Ed H. Chi, and Derek Zhiyuan Cheng. 2023. Do llms understand user preferences? evaluating llms on user rating prediction. *CoRR*, abs/2305.06474.

Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, Dor Muhlgay, Noam Rozen, Erez Schwartz, Gal Shachaf, Shai Shalev-Shwartz, Amnon Shashua, and Moshe Tennenholtz. 2022. MRKL systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *CoRR*, abs/2205.00445.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 passes the bar exam. *Available at SSRN 4389233*.

Tushar Khot, H. Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *CoRR*, abs/2302.02083.

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille ElepaĆ±o, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2(2):1–12.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *CoRR*, abs/2307.11991.

Brenden M. Lake and Marco Baroni. 2023. Human-like systematic generalization through a meta-learning neural network. *Nat.*, 623(7985):115–121.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023a. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *CoRR*, abs/2306.00890.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023c. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.

Hang Li. 2022. Language models: past, present, and future. *Commun. ACM*, 65(7):56–63.

Haonan Li, Yu Hao, Yizhuo Zhai, and Zhiyun Qian. 2023d. The hitchhiker's guide to program analysis: A journey with large language models. *CoRR*, abs/2308.00245.

Jia Li, Ge Li, Yongmin Li, and Zhi Jin. 2023e. Structured chain-of-thought prompting for code generation. *arXiv preprint arXiv:2305.06599*.

Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate neural template explanations for recommendation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 755–764. ACM.

Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4947–4957. Association for Computational Linguistics.

Lei Li, Yongfeng Zhang, and Li Chen. 2023f. Personalized prompt learning for explainable recommendation. *ACM Trans. Inf. Syst.*, 41(4):103:1–103:26.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023g. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. *Annual Meeting of the Association for Computational Linguistics*.

Qi Li. 2023. Harnessing the power of pre-trained vision-language models for efficient medical report generation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 1308–1317. ACM.

Siyu Li, Jin Yang, and Kui Zhao. 2023h. Are you in a masquerade? exploring the behavior and impact of large language model driven social bots in online social networks. *CoRR*, abs/2307.10337.

Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2022. A systematic investigation of commonsense knowledge in large language models. In *Proceedings of the 2022 Conference on Empirical Methods* in Natural Language Processing, pages 11838–11855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xin-Ye Li, Jiang-Tian Xue, Zheng Xie, and Ming Li. 2023i. Think outside the code: Brainstorming boosts large language models in code generation. *CoRR*, abs/2305.10679.

Yuan Li, Yixuan Zhang, and Lichao Sun. 2023j. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *CoRR*, abs/2310.06500.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. 2023k. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *CoRR*, abs/2303.14070.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. 2023. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *CoRR*, abs/2305.13655.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023a. Encouraging divergent thinking in large language models through multi-agent debate. *CoRR*, abs/2305.19118.

Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, Yun Wang, Linjun Shou, Ming Gong, and Nan Duan. 2023b. Taskmatrix.ai: Completing tasks by connecting foundation models with millions of apis. *CoRR*, abs/2303.16434.

Zhiding Liang, Jinglei Cheng, Rui Yang, Hang Ren, Zhixin Song, Di Wu, Xuehai Qian, Tongyang Li, and Yiyu Shi. 2023c. Unleashing the potential of llms for quantum computing: A study in quantum architecture design.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *CoRR*, abs/2207.08143.

Valentin Liévin, Andreas Geert Motzfeldt, Ida Riis Jensen, and Ole Winther. 2023. Variational open-domain question answering. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 20950–20977. PMLR.

Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. 2023. Agentsims: An open-source sandbox for large language model evaluation. *CoRR*, abs/2308.04026.

Zeming Lin, Halil Akin, Roshan Rao, Brian L. Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. 2022. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, and Liang Zhao. 2023a. Domain specialization as the key to make large language models disruptive: A comprehensive survey.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Carl J. Yang, and Liang Zhao. 2023b. Beyond one-model-fits-all: A survey of domain specialization for large language models. *CoRR*, abs/2305.18703.

Chao Liu, Xuanlin Bao, Hongyu Zhang, Neng Zhang, Haibo Hu, Xiaohong Zhang, and Meng Yan. 2023a. Improving chatgpt prompt for code generation. *CoRR*, abs/2305.08360.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, L. Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; Deep Learning Inside Out*.

Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S. Yu, and Chuan Shi. 2023b. Towards graph foundation models: A survey and beyond.

June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023c. Chatcounselor: A large language models for mental health support. *CoRR*, abs/2309.15461.

Xiao-Yang Liu, Guoxuan Wang, and Daochen Zha. 2023d. Fingpt: Democratizing internet-scale data for financial large language models. *CoRR*, abs/2307.10485.

Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2023e. How AI processing delays foster creativity: Exploring research question co-creation with an llm-based agent. *CoRR*, abs/2310.06155.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4513–4519. ijcai.org.

Junru Lu, Jiazheng Li, Byron C. Wallace, Yulan He, and Gabriele Pergola. 2023a. Napss: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1049–1061. Association for Computational Linguistics.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023b. Chameleon: Plug-and-play compositional reasoning with large language models. *CoRR*, abs/2304.09842.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023c. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv: 2304.09842*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *ACL*.

Ling Luo, Jinzhong Ning, Yingwen Zhao, Zhijun Wang, Zeyuan Ding, Peng Chen, Weiru Fu, Qinyu Han, Guangtao Xu, Yunzhi Qiu, Dinghao Pan, Jiru Li, Hao Li, Wenduo Feng, Senbo Tu, Yuqi Liu, Zhihao Yang, Jian Wang, Yuanyuan Sun, and Hongfei Lin. 2023a. Taiyi: A bilingual fine-tuned large language model for diverse biomedical tasks. *CoRR*, abs/2311.11608.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinform.*, 23(6).

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023b. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *CoRR*, abs/2308.09442.

Yun Luo, Xiaotian Lin, Zhen Yang, Fandong Meng, Jie Zhou, and Yue Zhang. 2023c. Mitigating catastrophic forgetting in task-incremental continual learning with adaptive classification criterion. *arXiv preprint arXiv: 2305.12270*.

Yun Luo, Zhen Yang, Xuefeng Bai, Fandong Meng, Jie Zhou, and Yue Zhang. 2023d. Investigating forgetting in pre-trained representations through continual learning. *arXiv preprint arXiv: 2305.05968*.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023e. Wizardcoder: Empowering code large language models with evol-instruct. *CoRR*, abs/2306.08568.

Teli Ma, Rong Li, and Junwei Liang. 2023. An examination of the compositionality of large generative vision-language models. *arXiv preprint arXiv: 2308.10509*.

Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. *Conference on Empirical Methods in Natural Language Processing*.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13641–13660. Association for Computational Linguistics.

Vaibhav Mavi, Abulhair Saparov, and Chen Zhao. 2023. Retrieval-augmented chain-of-thought in semi-structured domains. *CoRR*, abs/2310.14435.

Muhammad Miftahul Amri and Urfa Khairatun Hisan. 2023. Incorporating ai tools into medical education: Harnessing the benefits of chatgpt and dall-e. *Journal of Novel Engineering Science and Technology*, 2(02):34–39.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv: 2202.12837*.

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, Lisanna Paladin, Shriya Raj, Lorna J. Richardson, Robert D. Finn, and Alex Bateman. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res.*, 49(Database-Issue):D412–D419.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, D. Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *Annual Meeting of the Association for Computational Linguistics*.

Yida Mu, Ben P. Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. *CoRR*, abs/2305.14310.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

Muhammad U. Nasir, Sam Earle, Julian Togelius, Steven James, and Christopher Cleghorn. 2023. Llmatic: Neural architecture search via large language models and quality diversity optimization.

Michel Nass, Emil Alégroth, and Robert Feldt. 2023. Improving web element localization by using a large language model. *CoRR*, abs/2310.02046.

John J Nay. 2022. Law informs code: A legal informatics approach to aligning artificial intelligence with humans. *Nw. J. Tech. & Intell. Prop.*, 20:309.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. 2023. Climax: A foundation model for weather and climate. In *International Conference on Machine Learning, ICML 2023,* 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, pages 25904–25938. PMLR.

Ansong Ni, Srini Iyer, Dragomir Radev, Veselin Stoyanov, Wen-Tau Yih, Sida I. Wang, and Xi Victoria Lin. 2023. LEVER: learning to verify language-to-code generation with execution. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 26106–26128. PMLR.

Kole Norberg, Husni Almoubayyed, Stephen E. Fancsali, Logan De Ley, Kyle Weldon, April Murphy, and Steven Ritter. 2023. Rewriting math word problems with large language models. In *Proceedings of the Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation 2023 co-located with 24th International Conference on Artificial Intelligence in Education (AIED 2023), Tokyo, Japan, July 7, 2023*, volume 3487 of *CEUR Workshop Proceedings*, pages 163–172. CEUR-WS.org.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of GPT-4 on medical challenge problems. *CoRR*, abs/2303.13375.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolò Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023b. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *CoRR*, abs/2311.16452.

Dennis Norris. 2017. Short-term memory and long-term memory are still different. *Psychological bulletin*, 143(9):992.

Kolby Nottingham, Prithviraj Ammanabrolu, Alane Suhr, Yejin Choi, Hannaneh Hajishirzi, Sameer Singh, and Roy Fox. 2023. Do embodied agents dream of pixelated sheep: Embodied decision making using language guided world modelling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 26311–26325. PMLR.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Carlos Outeiral and Charlotte M. Deane. 2022. Codon language embeddings provide strong signals for protein engineering. *bioRxiv*.

Siru Ouyang, Zhuosheng Zhang, Bing Yan, Xuan Liu, Jiawei Han, and Lianhui Qin. 2023. Structured chemistry reasoning with large language models. *CoRR*, abs/2311.09656.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023a. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv: 2305.12295*.

Shirui Pan, Yizhen Zheng, and Yixin Liu. 2023b. Integrating graphs with large language models: Methods and prospects.

Bhargavi Paranjape, Scott M. Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Túlio Ribeiro. 2023. ART: automatic multi-step reasoning and tool-use for large language models. *CoRR*, abs/2303.09014.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.

Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *The 35th Annual ACM Symposium on User Interface Software and Technology, UIST 2022, Bend, OR, USA, 29 October 2022 - 2 November 2022*, pages 74:1–74:18. ACM.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2023. Gorilla: Large language model connected with massive apis. *CoRR*, abs/2305.15334.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, A. Bakhtin, Yuxiang Wu, Alexander H. Miller, and S. Riedel. 2019. Language models as knowledge bases? *Conference on Empirical Methods in Natural Language Processing*.

Tammy Pettinato Oltz. 2023. Chatgpt, professor of law. *Professor of Law (February 4, 2023)*.

Laura Plein, Wendkûuni C. Ouédraogo, Jacques Klein, and Tegawendé F. Bissyandé. 2023. Automatic generation of test cases based on bug reports: a feasibility study with large language models. *CoRR*, abs/2310.06320.

Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. 2023. Adapt: As-needed decomposition and planning with language models. *arXiv preprint arXiv: 2311.05772*.

Chen Qian, Huayi Tang, Zhirui Yang, Hong Liang, and Yong Liu. 2023. Can large language models empower molecular property prediction?

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5368–5393. Association for Computational Linguistics.

Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. 2023a. Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media. *CoRR*, abs/2305.05138.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023b. Tool learning with foundation models. *CoRR*, abs/2304.08354.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023c. Toolllm: Facilitating large language models to master 16000+ real-world apis. *CoRR*, abs/2307.16789.

Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian D. Reid, and Niko Sünderhauf. 2023. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. *CoRR*, abs/2307.06135.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton-Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. Code llama: Open foundation models for code. *CoRR*, abs/2308.12950.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *ICLR 2022-Tenth International Conference on Learning Representations*.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4). *arXiv preprint arXiv:2306.09525*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *CoRR*, abs/2302.04761.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 843–861. Association for Computational Linguistics.

Andrew W. Senior, Richard Evans, John M. Jumper, James Kirkpatrick, L. Sifre, Tim Green, Chongli Qin, Augustin Zídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. 2020. Improved protein structure prediction using potentials from deep learning. *Nature*, 577:706–710.

Chantal Shaib, Millicent L. Li, Sebastian Joseph, Iain James Marshall, Junyi Jessy Li, and Byron C. Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1387–1407. Association for Computational Linguistics.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, pages 1–6.

NAN SHAO, Zefan Cai, Hanwei xu, Chonghua Liao, Yanan Zheng, and Zhilin Yang. 2023. Compositional task representations for large language models. In *The Eleventh International Conference on Learning Representations*.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Xiaoming Shi, Zeming Liu, Chuan Wang, Haitao Leng, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2023. Midmed: Towards mixed-type dialogues for medical consultation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8145–8157. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *CoRR*, abs/2212.13138.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *CoRR*, abs/2305.09617.

Ritwik Sinha, Zhao Song, and Tianyi Zhou. 2023. A mathematical abstraction for balancing the trade-off between creativity and reality in large language models. *arXiv preprint arXiv:2306.02295*.

Alexey Skrynnik, Zoya Volovikova, Marc-Alexandre Côté, Anton Voronov, Artem Zholus, Negar Arabzadeh, Shrestha Mohanty, Milagro Teruel, Ahmed Awadallah, Aleksandr Panov, Mikhail Burtsev, and Julia Kiseleva. 2022. Learning to solve voxel building embodied tasks from pixels and natural language instructions. *CoRR*, abs/2211.00688.

Yifan Song, Weimin Xiong, Dawei Zhu, Cheng Li, Ke Wang, Ye Tian, and Sujian Li. 2023. Restgpt: Connecting large language models with real-world applications via restful apis. *CoRR*, abs/2306.06624.

Evangelia Spiliopoulou, Artidoro Pagnoni, Yonatan Bisk, and Eduard H. Hovy. 2022. Events realm: Event reasoning of entity states via language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1982–1997. Association for Computational Linguistics.

Douglas Summers-Stay, Clare R Voss, and Stephanie M Lukin. 2023. Brainstorm, then select: a generative language model improves its creativity score. In *The AAAI-23 Workshop on Creative AI Across Modalities*.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, and Yue Zhao. 2024. Trustllm: Trustworthiness in large language models. *CoRR*, abs/2401.05561.

Freyr Sverrisson, Jean Feydy, Bruno E. Correia, and Michael M. Bronstein. 2020. Fast end-to-end learning on protein surfaces. *bioRxiv*.

Ben Swanson, Kory Mathewson, Ben Pietrzak, Sherol Chen, and Monica Dinalescu. 2021. Story centaur: Large language model few shot learning as a creative writing tool. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 244–256.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric P. Xing, and Zhiting Hu. 2021. Progressive generation of long text with pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4313–4324. Association for Computational Linguistics.

Chen Tang, Shun Wang, Tomas Goldsack, and Chenghua Lin. 2023a. Improving biomedical abstractive summarisation with knowledge aggregation from citation papers. *CoRR*, abs/2310.15684.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. 2023b. Graphgpt: Graph instruction tuning for large language models.

DB-GPT Team. 2023a. DB-GPT.

Xwin-LM Team. 2023b. Xwin-lm.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2023. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences. *CoRR*, abs/2311.06025.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Neural Information Processing Systems*.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *CoRR*, abs/2310.05984.

Dietrich Trautmann, Alina Petrova, and Frank Schilder. 2022. Legal prompt engineering for multilingual legal judgement prediction. *arXiv preprint arXiv:2212.02199*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John M. Pauly, and Akshay S. Chaudhari. 2023. Clinical text summarization: Adapting large language models can outperform human experts. *CoRR*, abs/2309.07430.

Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. 2023. Chatgpt for robotics: Design principles and model abilities. *CoRR*, abs/2306.17582.

Boshi Wang, Xiang Deng, and Huan Sun. 2022a. Iteratively prompt pre-trained language models for chain of thought. *Conference on Empirical Methods in Natural Language Processing*.

Cunxiang Wang, Pai Liu, and Yue Zhang. 2021a. Can generative pre-trained language models serve as knowledge bases for closed-book qa? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3241–3251. Association for Computational Linguistics.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *CoRR*, abs/2305.16291.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. 2023b. Can language models solve graph problems in natural language?

Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, Kun Yu, Yuxing Yuan, Yinghao Zou, Jiquan Long, Yudong Cai, Zhenxiang Li, Zhifeng Zhang, Yihua Mo, Jun Gu, Ruiyi Jiang, Yi Wei, and Charles Xie. 2021b. Milvus: A purpose-built vector data management system. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, pages 2614–2627. ACM.

Lei Wang, Songheng Zhang, Yun Wang, Ee-Peng Lim, and Yong Wang. 2023c. Llm4vis: Explainable visualization recommendation using chatgpt.

Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023d. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *CoRR*, abs/2302.07257.

Siyuan Wang, Bo Peng, Yichao Liu, and Qi Peng. 2023e. Fine-grained medical vision-language representation learning for radiology report generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15949–15956. Association for Computational Linguistics.

Wenkai Wang, Zhenling Peng, and Jianyi Yang. 2022b. Single-sequence protein structure prediction using supervised transformer protein language models. *bioRxiv*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023f. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023g. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *CoRR*, abs/2307.05300.

Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023h. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *ArXiv*, abs/2302.01560.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023a. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv:2304.13835*.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. 2023b. Magicoder: Source code is all you need. *CoRR*, abs/2312.02120.

Yuxiang Wei, Chunqiu Steven Xia, and Lingming Zhang. 2023c. Copiloting the copilots: Fusing large language models with completion engines for automated program repair. *CoRR*, abs/2309.00608.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Towards generalist foundation model for radiology. *CoRR*, abs/2308.02463.

Rui Min Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, and Jian Peng. 2022. High-resolution de novo structure prediction from primary sequence. *bioRxiv*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann. 2023b. Bloomberggpt: A large language model for finance. *CoRR*, abs/2303.17564.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huan, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *CoRR*, abs/2309.07864.

Chunqiu Steven Xia and Lingming Zhang. 2023a. Conversational automated program repair. *CoRR*, abs/2301.13246.

Chunqiu Steven Xia and Lingming Zhang. 2023b. Keep the conversation going: Fixing 162 out of 337 bugs for $0.42 each using chatgpt. *CoRR*, abs/2304.00385.

Changnan Xiao and Bing Liu. 2023. Conditions for length generalization in learning reasoning skills.

Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. 2023a. The wall street neophyte: A zero-shot analysis of chatgpt over multimodal stock movement prediction challenges. *CoRR*, abs/2304.05351.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023b. PIXIU: A large language model, instruction data and evaluation benchmark for finance. *CoRR*, abs/2306.05443.

Zhuokui Xie, Yinghao Chen, Chen Zhi, Shuiguang Deng, and Jianwei Yin. 2023c. Chatunitest: a chatgpt-based automated unit test generation tool. *CoRR*, abs/2305.04764.

Zhenchang Xing, Qing Huang, Yu Cheng, Liming Zhu, Qinghua Lu, and Xiwei Xu. 2023. Prompt sapper: Llm-empowered software engineering infrastructure for ai-native services. *CoRR*, abs/2306.02230.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. 2023a. Protst: Multi-modality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*.

Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Hong Yu, James A. Hendler, Anind K. Dey, and Dakuo Wang. 2023b. Leveraging large language models for mental health prediction via online text data. *CoRR*, abs/2307.14385.

Junbing Yan, Chengyu Wang, Taolin Zhang, Xiaofeng He, Jun Huang, and Wei Zhang. 2023. From complex to simple: Unraveling the cognitive tree for reasoning with small language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 12413–12425. Association for Computational Linguistics.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023a. DOC: improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3378–3465. Association for Computational Linguistics.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4393–4479. Association for Computational Linguistics.

Yizhe Yang, Huashan Sun, Jiawei Li, Runheng Liu, Yinghao Li, Yuhang Liu, Heyan Huang, and Yang Gao. 2023b. Mindllm: Pre-training lightweight large language model from scratch, evaluations and domain applications. *CoRR*, abs/2310.15777.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. In *NeurIPS*.

Seonghyeon Ye, Hyeonbin Hwang, Sohee Yang, Hyeongu Yun, Yireun Kim, and Minjoon Seo. 2023. In-context instruction learning. *arXiv preprint arXiv:2302.14691*.

Caiyang Yu, Xianggen Liu, Wentao Feng, Chenwei Tang, and Jiancheng Lv. 2023a. Gpt-nas: Evolutionary neural architecture search with the generative pre-trained model.

Dingyao Yu, Kaitao Song, Peiling Lu, Tianyu He, Xu Tan, Wei Ye, Shikun Zhang, and Jiang Bian. 2023b. Musicagent: An AI agent for music understanding and generation with large language models. *CoRR*, abs/2310.11954.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Nianyi Lin, Kaifeng Yun, Linlu Gong, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, and Juanzi Li. 2023c. Kola: Carefully benchmarking world knowledge of large language models.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. In *IUI 2022: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, March 22 - 25, 2022*, pages 841–852. ACM.

Haoqi Yuan, Chi Zhang, Hongcheng Wang, Feiyang Xie, Penglin Cai, Hao Dong, and Zongqing Lu. 2023. Plan4mc: Skill reinforcement learning and planning for open-world minecraft tasks. *CoRR*, abs/2303.16563.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Wei Lin, et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Daoguang Zan, Bei Chen, Fengji Zhang, Dianjie Lu, Bingchao Wu, Bei Guan, Yongji Wang, and Jian-Guang Lou. 2023. Large language models meet nl2code: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7443–7464. Association for Computational Linguistics.

Bowen Zhang and Harold Soh. 2023. Large language models as zero-shot human models for human-robot interaction. *CoRR*, abs/2303.03548.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023a. Huatuogpt, towards taming language model to be a doctor. *CoRR*, abs/2305.15075.

Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023b. Recommendation as instruction following: A large language model empowered recommendation approach. *CoRR*, abs/2305.07001.

Kechi Zhang, Zhuo Li, Jia Li, Ge Li, and Zhi Jin. 2023c. Self-edit: Fault-aware code editor for code generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 769–787. Association for Computational Linguistics.

Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Siyuan Cheng, Haosen Hong, Shumin Deng, Jiazhang Lian, Qiang Zhang, and Huajun Chen. 2022a. Ontoprotein: Protein pretraining with gene ontology embedding. *ArXiv*, abs/2201.11147.

Xuanyu Zhang and Qing Yang. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 4435–4439. ACM.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022b. Active example selection for in-context learning. *Conference on Empirical Methods in Natural Language Processing*.

Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I. Jordan, and Jianmin Wang. 2023d. Skilful nowcasting of extreme precipitation with nowcastnet. *Nat.*, 619(7970):526–532.

Yuwei Zhang, Zhi Jin, Ying Xing, and Ge Li. 2023e. STEAM: simulating the interactive behavior of programmers for automatic bug fixing. *CoRR*, abs/2308.14460.

Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, Yijian Qin, Simin Wu, and Wenwu Zhu. 2023f. Llm4dyg: Can large language models solve problems on dynamic graphs?

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023g. How do large language models capture the ever-changing world knowledge? A review of recent advances. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 8289–8311. Association for Computational Linguistics.

Ziwei Zhang, Haoyang Li, Zeyang Zhang, Yijian Qin, Xin Wang, and Wenwu Zhu. 2023h. Large graph models: A perspective. *arXiv preprint arXiv:2308.14522*.

Ziyin Zhang, Chaoyu Chen, Bingchang Liu, Cong Liao, Zi Gong, Hang Yu, Jianguo Li, and Rui Wang. 2023i. A survey on language models for code. *CoRR*, abs/2311.07989.

Zuobai Zhang, Minghao Xu, Arian R. Jamasb, Vijil Chenthamarakshan, Aurélie C. Lozano, Payel Das, and Jian Tang. 2022c. Protein representation learning by geometric structure pretraining. *ArXiv*, abs/2203.06125.

Bowen Zhao, Changkai Ji, Yuejie Zhang, Wen He, Yingwen Wang, Qing Wang, Rui Feng, and Xiaobo Zhang. 2023a. Large language models are complex table parsers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14786–14802, Singapore. Association for Computational Linguistics.

Guosheng Zhao, Yan Yan, and Zijian Zhao. 2023b. Normal-abnormal decoupling memory for medical report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1962–1977. Association for Computational Linguistics.

Junjie Zhao, Xiang Chen, Guang Yang, and Yiheng Shen. 2023c. Automatic smart contract comment generation via large language models and in-context learning. *CoRR*, abs/2311.10388.

Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. 2021a. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6277–6286.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021b. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2023d. Large language models as commonsense knowledge for large-scale task planning. *arXiv preprint arXiv: 2305.14078*.

Mingkai Zheng, Xiu Su, Shan You, Fei Wang, Chen Qian, Chang Xu, and Samuel Albanie. 2023. Can gpt-4 perform neural architecture search?

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2856–2878.

Denny Zhou, Nathanael Scharli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, D. Schuurmans, O. Bousquet, Quoc Le, and E. Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *International Conference on Learning Representations*.

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Josh Susskind, Samy Bengio, and Preetum Nakkiran. 2023a. What algorithms can transformers learn? a study in length generalization.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023b. Don't make your LLM an evaluation benchmark cheater. *CoRR*, abs/2311.01964.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix X. Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *CoRR*, abs/2012.00363.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *CoRR*, abs/2305.03514.

A Interactions of the LLMs

The significant advantages of LLMs are not only manifested in their typical capabilities, but also in their strong interactive capabilities. The interactive capabilities are essentially a perfect integration of the model's inherent capabilities with external information. It is primarily reflected in enhancing the model's performance by acquiring information from the outside environments, planning and decision-making for external environments, and using external tools (Xi et al., 2023). In this section, we will focus on discussing the capabilities of LLMs in terms of use tools and environment interaction, as well as personalized and customized interaction.

A.1 Use tools and environment interaction

Integrating specialized tools with LLMs can fully leverage their unique advantages, addressing the limitations of LLMs in specific domain tasks (Qin et al., 2023b). There are primarily two modes of interaction between large models and tools: First, external tools can continuously modify and refine the instructions for LLMs, enabling LLMs to perform more complex tasks. ToolFormer (Schick et al., 2023) utilizes prompts to guide the model to generate candidate texts that meet the instructions' requirements, followed by an automated process to filter high-quality results. Additionally, ART (Paranjape et al., 2023) employs a specific program syntax to build a task repository. When a new task emerges, it retrieves similar tasks from this repository to add to the prompt. Moreover, LLMs can also play a coordinating role in the system, issuing outlines for solving tasks and automatically matching sub-tasks outlined in the framework with APIs, systems and models that have specific functionalities to complete tasks (Patil et al., 2023; Liang et al., 2023b; Qin et al., 2023c).

LLMs significantly expand their application scope by interacting with external environments through unified natural language interfaces and tool use. For instance, WebGPT (Nakano et al., 2021) interacts with a text-based web browsing environment, enabling end-to-end optimization search and aggregation through imitation and reinforcement learning. WebShop (Yao et al., 2022) trains LLMs using real-world product information and crowd-sourced textual instructions, enabling navigation and various operations on e-commerce websites. HuggingGPT interacts with the Huggingface community, utilizing ChatGPT to process user requests, selecting models based on function descriptions within the community, and executing AI tasks with the chosen models. The interaction of LLMs with database environments adds capabilities such as knowledge base management, unified data vectorization storage and indexing, and automated prompt generation and optimization. This ensures complete control over sensitive data and environments, preventing any data privacy breaches or security risks (Team, 2023a). Vector databases provide large models with expanded memory storage space and enhanced capabilities for advanced query processing (Wang et al., 2021b).

A.2 Personalized and customized interaction

The enhancement of LLMs' capabilities has transformed the interaction between humans and personalized systems. Unlike traditional recommendation systems and search engines that passively filter information, LLMs provide a foundation for proactive user participation (Chen et al., 2023d). Firstly, LLMs extend the capability of fact retrieval into explicit knowledge bases, offering a more comprehensive knowledge source for recommendation systems (Jiang et al., 2020; Heinzerling and Inui, 2021; Wang et al., 2021a). This allows for a broader and more accurate understanding of user queries and preferences. Secondly, the instructions tailored for recommendation scenarios can make LLMs significantly outperform traditional recommenders (Kang et al., 2023; Zhang et al., 2023b). The characteristics of users and their interaction history can be efficiently transformed into natural language instructions for input to LLMs (Chen, 2023). Furthermore, the robust interpretability of LLMs enables the creation of precise, natural, and user-preference-aligned custom explanations, alleviating the limitations of traditional, formulaic explanations (Li et al., 2020, 2021, 2023f). Lastly, LLMs with strong reasoning and decision-making capabilities, such as GPT-NAS (Yu et al., 2023a), GENIUS (Zheng et al., 2023), and LLMatic (Nasir et al., 2023), provide enhanced support for personalized customization services. These models leverage their advanced cognitive capabilities to deliver more accurate and user-centric recommendations, enhancing the overall personalization experience.

B Summary of Capabilities in Various Domains

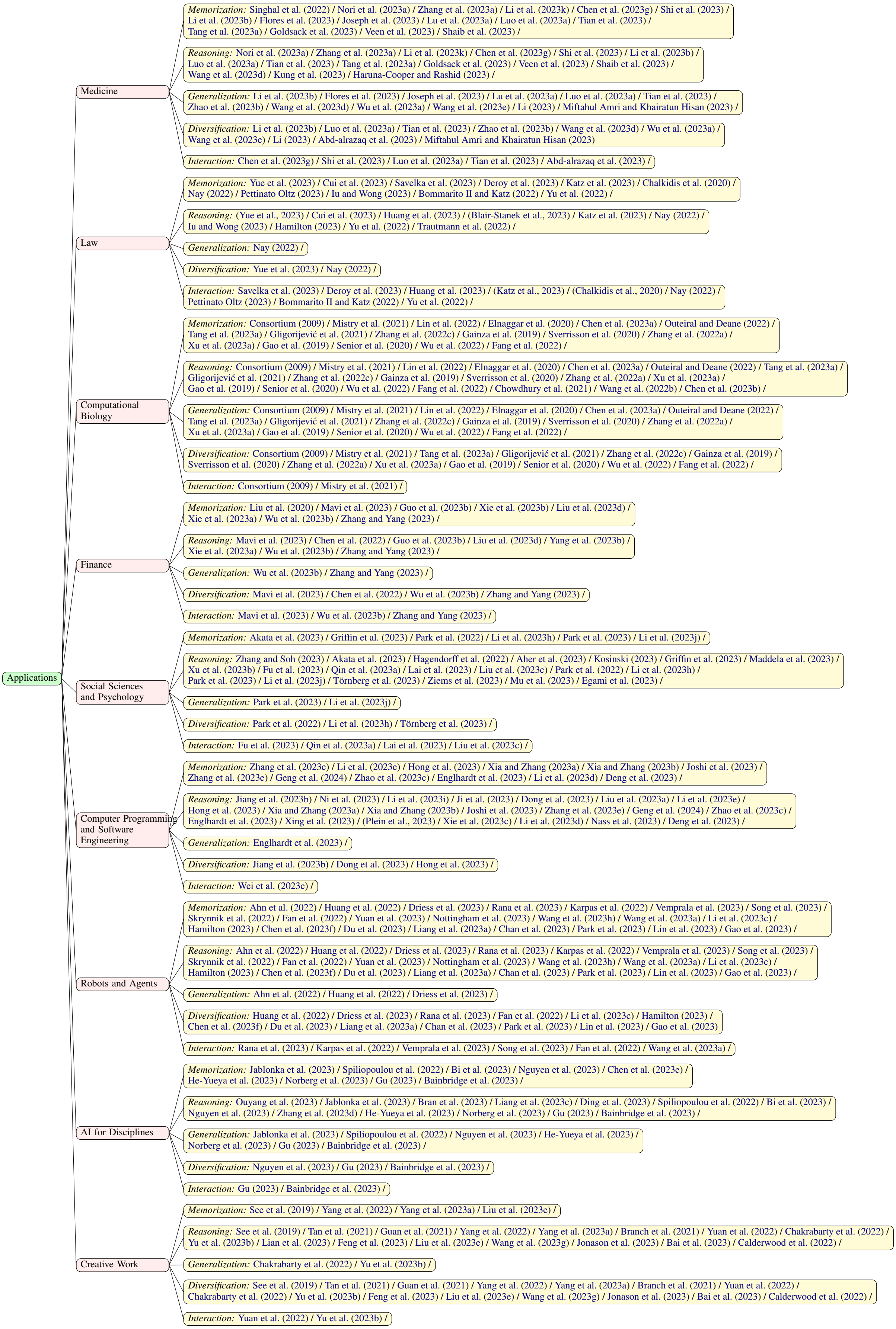

Figure 3: Correspondence between domains and fundamental capabilities in this paper.