Prompting for creation of Trinity Models: A Survey

Swapnil Bhattacharyya
swapnil2000bhyya@gmail.com

Pushpak Bhattacharyya
pushpakbh@gmail.com

Abstract

Prompting has emerged as a powerful strategy for building domain-task-language specific models across several specialized applications. In consumer grievance chatbots, prompting enables task-oriented dialogue flows that guide users through structured complaint filing and redressal processes while ensuring domain relevance and conversational consistency. Legal decision assist tools leverage prompting for multi-step legal reasoning, material summarization, and retrieval of similar cases, offering valuable support for legal professionals in analyzing complex case files. Question Answering in Marathi presents unique challenges due to low-resource language settings, where prompt-based techniques combined with multilingual retrieval are employed to enable accurate responses from regional government datasets. Across all three tasks, prompting addresses key demands such as reasoning consistency, domain-specific knowledge integration, and multilingual adaptation, though challenges like hallucination, low-resource data scarcity, and prompt transferability remain active areas of research.

1 Introduction

The advent of large-scale pre-trained language models (LLM) such as GPT-3 (Brown et al., 2020a), PaLM (Chowdhery et al., 2022) and GPT-4 (OpenAI et al., 2024) has led to significant breakthroughs in natural language processing (NLP), demonstrating strong capabilities across various tasks ranging from text generation and summarization to question answering and code synthesis. These models achieve generalization by being pretrained on massive amounts of web-scale corpora, enabling them to acquire substantial world knowledge and linguistic competence.

However, when deployed in specialized domains such as law, medicine, finance, and scientific research, general-purpose models often face challenges due to the presence of domain-specific terminology, knowledge gaps, regulatory constraints, and complex reasoning requirements (Xu et al., 2025; Bommasani et al., 2022). To address these limitations, *Domain-Specific Language Models* (DSLMs) have emerged, either by training models from scratch on domain-relevant corpora or by adapting general models via continued pretraining or fine-tuning (Lee et al., 2019; Chalkidis et al., 2020; Yang et al., 2020). These DSLMs are tailored to handle specialized tasks, benefiting from exposure to in-domain vocabulary, structured knowledge, and nuanced reasoning patterns.

Despite the availability of domain-specific models, fully supervised fine-tuning remains costly and impractical in many real-world settings due to limited labeled data, privacy concerns, and high computational demands. In this context, *prompting* has gained prominence as a lightweight and flexible mechanism to adapt LLMs and DSLMs to downstream tasks without extensive retraining. Prompting allows models to condition their outputs based on carefully designed instructions or exemplars, effectively converting a wide range of NLP problems into text-to-text tasks (Liu et al., 2021).

1.1 Discrete and Continuous Prompting

Prompt technique can be broadly categorized into two families: *discrete prompting* and *continuous prompting*. These paradigms offer complementary capabilities for steering DSLMs, each with unique advantages, limitations, and applications in domain-specific contexts.

1.1.1 Discrete Prompting

Discrete prompting, often referred to as *textual prompting*, involves the manual or semi-automatic construction of natural language instructions that are fed directly to the model as input text. Early works demonstrated the surprising capability of LLMs to follow natural language instructions

when provided with zero-shot or few-shot exemplars (Brown et al., 2020a; Schick and Schütze, 2021). Subsequent developments such as *Chain-of-Thought (CoT)* prompting (Wei et al., 2023) further improved performance on complex reasoning tasks by encouraging models to generate intermediate reasoning steps.

In domain-specific settings, discrete prompting is particularly appealing due to its transparency and interpretability. Domain experts can craft prompts that explicitly specify task requirements, incorporate domain terminology, and impose task constraints. For instance, in legal document summarization, prompts may request extraction of *"case facts, legal provisions applied, and relief granted"*. In biomedical applications, prompts may guide models to extract *"clinical findings, diagnosis hypotheses, and treatment recommendations"* (Singhal et al., 2022a).

However, discrete prompting is highly sensitive to prompt phrasing, template design, and instruction format (Zhao et al., 2021; Lu et al., 2022). Crafting effective prompts often requires domain expertise and iterative experimentation, particularly in highly specialized domains where incorrect instructions can lead to hallucinations, omissions, or irrelevant outputs. Moreover, discrete prompts may suffer from limited generalizability across tasks or datasets, making large-scale deployment challenging.

1.1.2 Continuous Prompting

Continuous prompting, also known as *soft prompting* or *prompt tuning*, represents a complementary paradigm wherein continuous vector representations are learned and prepended to model embeddings during inference (Lester et al., 2021; Li and Liang, 2021a,b). Rather than relying on natural language instructions, continuous prompts are optimized via gradient-based learning on small amounts of task-specific data.

In domain-specific scenarios, continuous prompting offers several attractive properties. First, it enables parameter-efficient adaptation, requiring updates to only a small set of prompt parameters while keeping the vast majority of model weights frozen (Lester et al., 2021). This is particularly valuable for DSLMs, where retraining entire models may be prohibitive due to computational or privacy constraints. Second, continuous prompts can capture fine-grained domain-specific task information that may be difficult to express explicitly via discrete natural language. Third, continuous prompting can facilitate multi-task or multi-domain adaptation by learning separate prompt embeddings for different tasks (Vu et al., 2022).

Nevertheless, continuous prompting introduces trade-offs. Unlike discrete prompts, continuous prompts are not human-interpretable, limiting transparency and making debugging more difficult in high-stakes domains such as healthcare or law. Furthermore, continuous prompts may overfit to narrow task distributions if not carefully regularized or if insufficient data is available (Wang et al., 2022). Additionally, the learned prompt embeddings may not easily transfer across models or domains, reducing their reusability compared to well-designed discrete templates.

1.2 Motivation

The widespread deployment of DSLMs across numerous high-stakes domains necessitates a deeper understanding of how prompting methods can be adapted, optimized, and evaluated for domain-specific tasks. While prompting has become a core technique in general-purpose LLM applications, its application to DSLMs presents new challenges: handling long documents, ensuring factual consistency, aligning with domain ontologies, and managing domain-specific evaluation standards.

This survey aims to fill this gap by providing a comprehensive review of prompting strategies tailored for DSLMs. Specifically, our contributions include:

- Presenting a unified taxonomy of both discrete and continuous prompting approaches for DSLMs.
- Systematically analyzing domain-specific challenges in prompt design, robustness, and evaluation.
- Reviewing state-of-the-art prompting methods across major domains such as law, medicine, finance, and scientific literature.
- Identifying emerging trends, open research problems, and future directions in domain-specific prompting research.

2 Background

Prompting has emerged as a central paradigm in adapting large language models (LLMs) to downstream tasks, particularly when labeled training

data is scarce or costly to obtain. Unlike traditional supervised fine-tuning approaches, prompting leverages pretrained language models as conditional generators, steering their outputs via carefully crafted input instructions or learned representations. In the context of domain-specific language models (DSLMs), prompting provides an efficient mechanism to leverage pretrained knowledge while incorporating domain-specific constraints, reasoning patterns, and specialized knowledge.

Large language models such as GPT-3 (Brown et al., 2020a), PaLM (Chowdhery et al., 2022), and LLaMA (Touvron et al., 2023a) are trained on massive corpora using next-token prediction objectives, enabling them to learn statistical patterns of language, factual knowledge, and even rudimentary reasoning capabilities. However, these models are often trained on open-domain datasets, which may not fully capture the linguistic nuances, specialized vocabulary, or domain-specific knowledge required for legal, biomedical, financial, or scientific applications.

Domain-specific language models (DSLMs) address this gap by either:

- **Pretraining from scratch** on large domain-specific corpora (e.g., PubMedBERT (Gu et al., 2021) for biomedical literature, LegalBERT (Chalkidis et al., 2020) for legal text, or FinBERT (Yang et al., 2020) for financial reports).
- **Continued pretraining** or *domain-adaptive pretraining* (DAPT) on top of existing general-purpose models (Gururangan et al., 2020).
- **Instruction tuning** on domain-specific datasets to align model behavior to domain-specific tasks (Ouyang et al., 2022).

Even with DSLMs, explicit prompt design remains crucial for downstream performance, particularly in low-resource domains where fine-tuning data is limited. Before delving into the details of different prompting strategies we discuss two language models that we have applied in our applications.

3 Llama 3.1 and Gemma 2: Model Overview

Llama 3.1 (Dubey et al., 2024), developed by Meta, is the latest evolution in the Llama series, designed for scalable, efficient, and adaptable large language modeling. Building on its Transformer-based architecture, Llama 3.1 introduces improvements such as long-context support up to 128,000 tokens, gradient checkpointing, mixed-precision training, and dynamic scaling for efficient learning. It is offered in sizes ranging from 8 billion to 405 billion parameters, making it accessible to both large research labs and smaller organizations. The model performs strongly in zero-shot and few-shot learning, supports parameter-efficient fine-tuning methods like LoRA, and is optimized for real-time inference on diverse hardware. Its flexibility makes it well-suited for tasks such as summarization, translation, question answering, and multi-turn dialogue across domains.

Gemma 2 (Team et al., 2024), developed by Google AI, is a versatile family of large language models available in 2 billion, 9 billion, and 27 billion parameter sizes, with smaller models optimized for edge and mobile devices. Known for its computational efficiency and fast inference capabilities, Gemma 2 employs quantization and memory-efficient transformer variants to support deployment in resource-constrained environments. It is particularly strong in multilingual applications, especially for Indian languages, due to its diverse and balanced training dataset. Gemma 2 is also ethically aligned, with proactive filtering to minimize harmful content generation. Its ability to handle tasks ranging from question answering and summarization to code generation and creative writing makes it highly adaptable for research, educational, and real-world deployment scenarios.

In the next sections, we survey the literature in three different applications of trinity models.

4 Consumer Grievance chatbot

Domain specialization of large language models (LLMs) (Ling et al., 2024) is the process of customizing general-purpose LLMs according to specific domain contextual data, enhanced by domain-specific knowledge, optimized by the domain objective, and regulated by domain-specific constraints. Domain specific models such as MedPalm (Singhal et al., 2023a), saulLLM (Colombo et al., 2024) and FinGPT (Yang et al., 2023) have created a new paradigm through their comparative performance as opposed to general purpose models like GPT-3 (Brown et al., 2020b). Besides being a family of Transformer-based neural language models specialized for dialog, LaMDA (Thoppilan et al.,

2022) also paved the way for domain-grounding using system prompts. Task-oriented dialogue systems assist users in achieving specific domain-related goals through interactive conversations (Yi et al., 2024) which have made them quite popular. The different models and techniques that helped us in conceptualizing Grahaknyay (our consumer grievance chatbot) range from GPT to Llama as well as approaches in RAG and prompting. (Radford et al., 2019) suggested the importance of zero-shot prompting, reducing our reliance on data and giving importance to carefully crafted prompts. Similar to the system prompt is ghost attention (Touvron et al., 2023b), which came with Llama2, a potent open source chat model even available in small 7B version, easily tailoring to user needs. On the other hand, LaMDA (Thoppilan et al., 2022) gives enough evidence for using an external information retrieval system instead of increasing model size for groundedness. Although models like LawGPT (Zhou et al., 2024b) and LegaLlama (Chalkidis et al., 2023) have achieved domain specificity through finetuning, pre-trained models have also been prompted for domain grounding in few-shot settings as in SwitchPrompt (Goswami et al., 2023). Prompting has been effective for social conversation synthesis (Chen et al., 2023) and as a creator of high-quality conversational agents (Lee et al., 2023) in low-resource situations. RAG has also proved to be an effective mechanism for building chatbots (Kulkarni et al., 2024) to answer domain-specific questions.

We have tried to develop a domain-specific consumer grievance redressal chatbot using a system prompt and simple RAG-based framework. Taking the efficacy of LLM as a judge into consideration (Zheng et al., 2023) we have extensively performed automatic evaluation considering the limited baselines and calculated their correlation with scores given by human experts. For the decision assist tool, for extracting useful information besides efficiently engineering the system prompt, methods like self-discover (Zhou et al., 2024a) are quite useful.

5 Decision Assist Tool

Legal summarization involves condensing complex legal texts, such as court rulings, legislative documents, and contracts, into shorter, more accessible versions without losing critical legal meaning. Approaches to legal summarization can be broadly categorized into extractive, abstractive, and hybrid methods (Shukla et al., 2022a). Extractive summarization selects important sentences directly from the original text, while abstractive summarization involves generating new sentences that capture the essence of the original content (Zhang et al., 2024). Hybrid methods combine both approaches to improve the quality of the summary. Recent research has increasingly focused on using transformer-based models, which have shown significant promise in improving summarization accuracy, especially for complex legal documents (Akter et al., 2025).

Predicting similar cases and using them as legal precedents for easing has been explored quite vividly as a challenging research problem (Wu et al., 2023). Extracting legal elements from judicial documents helps enhance the interpretative and analytical capacities of legal cases, thereby facilitating a wide array of downstream applications in various domains of law (Zongyue et al., 2023). The legal elements, which typically comprise key facts in a specialized legal context, can improve the relevance matching of legal case retrieval (Deng et al., 2024). Unsupervised case retrieval methods using event extraction have also been performed recently (Joshi et al., 2023).

Prompt engineering has emerged as an indispensable technique for extending the capabilities of large language models (LLMs). This approach leverages task-specific instructions, known as prompts, to enhance model efficacy without modifying the core model parameters (Sahoo et al., 2024). Prompt techniques such as chain of thought (Wei et al., 2023) pave the way for introducing intermediate reasoning steps that help to complete the task better. Our approach uses a combination of prompting, extraction of legal elements (specifically sector information), and sector identification to obtain adequately and fluently drafted material summaries of consumer case files.

The Llama models natively support coding, reasoning, tool usage, and various NLP generation tasks. Being open source, it's widely used by the research community. Several studies emphasize the importance of human evaluation in assessing NLP models (Guzmán et al., 2015; Gillick and Liu, 2010). However, challenges such as inconsistent quality and limited reproducibility make human evaluation complex (Clark et al., 2021). Moreover, constructing large-scale reference-based datasets can be costly. Recent research demonstrates the

potential of Large Language Models (LLMs) as reference-free evaluators for Natural Language Generation (NLG) tasks. For example, Liu et al. (2023) introduced G-EVAL, which employs LLMs with chain-of-thought (CoT) reasoning and a form-filling approach, showing strong alignment with human judgments in summarization. Similarly, Chiang and Lee (2023) found that LLM-based evaluation produces results comparable to expert human assessments. Zheng et al. (2023) reported that advanced LLMs, such as GPT-4, exhibit agreement levels similar to human evaluators, and Siledar et al. (2024) validated the effectiveness of LLMs in evaluating both proprietary and open-source models for opinion summarization.

6 Enhancing transparency in OGD systems

Recent years have seen a surge in Open Government Data (OGD) initiatives worldwide, aiming to increase transparency and accountability. This section explores research on three key areas: OGD fundamentals, similar applications in other government domains, and language models for Indic languages. Researchers have defined OGD terminology, explored its impact on citizen engagement, and identified challenges in OGD implementation (Attard et al., 2015). (Peña et al., 2023) focused on using LLMs for topic classification in public documents. For Indian languages, several studies have provided datasets and pre-trained BERT models. (Dabre et al., 2022) highlighted the benefits of script unification for low-resource languages, while Haq et al. demonstrated the effectiveness of machine-translated data for improving retriever performance. Our work focuses on enhancing a QA system for Marathi government orders. By leveraging our carefully curated dataset, we aim to improve the system's ability to provide accurate and informative answers.

7 Prompting Taxonomy for DSLMs

While general-purpose prompting methods have achieved impressive success in open-domain NLP, their direct application to domain-specific language models (DSLMs) requires careful adaptation due to several unique challenges. DSLMs operate under specialized vocabularies, knowledge constraints, task formats, and reasoning patterns that differ significantly across domains such as law, medicine, finance, and scientific research. In this section, we present a taxonomy of prompting strategies specifically adapted to DSLMs but at the same time based on the prompting hierarchy used in large language models, organized along several key axes: task complexity, supervision level, reasoning depth, and domain constraints.

7.1 Zero-Shot Prompting in DSLMs

Zero-shot prompting remains a highly attractive paradigm in DSLMs when labeled training data is limited or unavailable. In zero-shot settings, the model is conditioned purely through natural language instructions, without any explicit task demonstrations.

Unlike open-domain LLMs, DSLMs often require specialized task formulations that explicitly invoke domain-specific language, templates, and evaluation criteria. For example, legal zero-shot prompts may specify extraction of *"parties involved, case facts, legal statutes cited, and judgment issued"* (Chalkidis et al., 2021), while biomedical prompts may request outputs following clinical report structures (Lehman et al., 2021).

Several studies have demonstrated that carefully engineered zero-shot prompts—often co-designed with domain experts—can achieve surprisingly strong performance in DSLMs (Zheng et al., 2021; Nori et al., 2023; Singhal et al., 2022a). However, zero-shot performance remains highly sensitive to prompt phrasing and the model's internal domain knowledge.

7.2 Few-Shot Prompting in DSLMs

Few-shot prompting enhances zero-shot setups by adding a small set of in-context examples to illustrate the desired input-output mappings (Brown et al., 2020a). In domain-specific tasks, these few-shot exemplars are often manually curated by domain experts to ensure high-quality coverage of task formats, label definitions, and complex decision boundaries.

For example, (Singhal et al., 2022b) demonstrate improved performance in biomedical question answering using carefully constructed few-shot exemplars highlighting clinical reasoning paths. Similarly, in legal judgment summarization, domain experts can provide few-shot demonstrations that illustrate how statutory provisions and facts interact to yield legal outcomes (Zheng et al., 2021; Chalkidis et al., 2020).

One notable challenge in few-shot prompting for DSLMs is the limited availability of diverse

and representative examples due to data privacy, annotation cost, and legal restrictions—especially in healthcare and finance.

7.3 Instruction Prompting in DSLMs

Instruction prompting leverages large instruction-tuned models such as FLAN-T5 (Chung et al., 2022), InstructGPT (Ouyang et al., 2022), and domain-adapted instruction models like BioMedLM (Singhal et al., 2022a) or Med-PaLM 2 (Singhal et al., 2023b). Instruction-tuning enhances DSLMs by directly optimizing models to follow domain-specific natural language instructions.

In practice, instruction prompting helps align DSLMs with professional guidelines, regulatory policies, and ethical considerations. For instance, financial instruction prompts may enforce risk disclosure rules, while medical instructions can reflect clinical guidelines or diagnostic protocols (Hu et al., 2025).

Instruction prompting is increasingly viewed as an effective middle ground between pure prompting and full fine-tuning for DSLMs, particularly when domain-annotated instruction datasets are available.

7.4 Reasoning-Augmented Prompting for DSLMs

A major advantage of prompting is its ability to elicit structured reasoning processes that closely resemble expert decision-making. DSLMs benefit substantially from reasoning-augmented prompting strategies, several of which have emerged recently:

7.4.1 Chain-of-Thought (CoT)

As introduced in (Wei et al., 2023), CoT prompting elicits intermediate reasoning steps that improve multi-step logical and arithmetic reasoning. In DSLMs, CoT is particularly effective for modeling legal reasoning chains (e.g., statute application sequences), clinical diagnostic flows, and financial portfolio evaluations (Kojima et al., 2023; Wang et al., 2023).

7.4.2 Tree-of-Thought (ToT)

Tree-of-Thought prompting (Yao et al., 2023) extends CoT by allowing exploration of multiple parallel reasoning branches with search and evaluation mechanisms. ToT has shown strong potential in complex multi-diagnosis generation, multi-party legal case analysis, and policy simulations.

7.4.3 Graph-of-Thought (GoT)

Graph-of-Thought prompting (Besta et al., 2024) models reasoning as a dynamic graph where nodes represent partial solutions and edges denote transitions. This is highly aligned with tasks requiring non-linear reasoning, such as legal multi-jurisdictional conflicts or scientific hypothesis testing across interrelated studies.

7.4.4 Buffer-of-Thought (BufT)

Buffer-of-Thought (Yang et al., 2024) introduces external self-evaluation buffers that allow DSLMs to pause, critique, and revise intermediate reasoning steps. This reflection-driven framework reduces hallucinations and inconsistency, which is critical in high-stakes domains like clinical medicine or financial forecasting.

7.4.5 Self-Discovery and Self-Refinement

Self-discovery (Zelikman et al., 2022) and self-refinement (Madaan et al., 2023) frameworks enable models to autonomously decompose tasks into subtasks and iteratively refine their solutions, allowing more robust problem solving when domain knowledge is fragmented or partially uncertain.

7.5 Retrieval-Augmented Prompting for DSLMs

DSLMS often operate in dynamic domains where up-to-date knowledge is critical. Retrieval-Augmented Prompting (RAP) combines pretrained models with retrieval mechanisms to dynamically inject relevant knowledge into the prompt (Lewis et al., 2021; Izacard and Grave, 2021).

In legal NLP, RAP retrieves case precedents, statutes, and regulatory documents at inference time (Shukla et al., 2022b); in biomedicine, RAP retrieves recent publications and guidelines (Lee et al., 2019); while in finance, RAP can inject real-time market reports or compliance updates (Yang et al., 2020).

Recent works combine RAP with CoT, ToT and BufT frameworks to simultaneously ground model reasoning and reduce hallucination (Fadeeva et al., 2024; Shi et al., 2024).

7.6 Continuous Prompting for DSLMs

As discussed earlier, continuous prompting strategies such as soft prompt tuning (Lester et al., 2021), prefix tuning (Li and Liang, 2021a), and P-tuning (Li and Liang, 2021b) enable efficient adaptation of DSLMs without modifying the full model. In

domain-specific contexts, continuous prompting allows DSLMs to absorb highly specialized task structures and domain schemas in a parameter-efficient manner (Tuan et al., 2022; Vu et al., 2022).

Hybrid approaches that combine continuous and discrete prompts are increasingly being explored for DSLMs to balance interpretability and performance (Han et al., 2021).

8 Evaluation Methods for Prompted DSLMs

Evaluation of prompted Domain-Specific Language Models (DSLMs) remains one of the most challenging aspects of applied NLP research. Unlike general-purpose benchmarks, DSLMs often operate in high-stakes, highly specialized domains where correctness, factual grounding, and reasoning validity are essential. This section reviews both traditional and emerging evaluation frameworks applicable to DSLMs.

8.1 General Evaluation Challenges in DSLMs

Prompted DSLMs introduce several unique evaluation difficulties:

- **Ambiguity of Gold Labels:** Many domain tasks (e.g., legal argumentation, medical diagnosis) have multiple valid solutions depending on jurisdiction, guidelines, or expert interpretation.
- **Hallucination Detection:** DSLMs are prone to factual errors that may not be easily identified without expert knowledge (Maynez et al., 2020).
- **Multi-Hop Reasoning Verification:** Reasoning chains must be evaluated stepwise, not just based on final answers.
- **Structured Output Comparison:** Outputs often include multiple fields (e.g., case facts, statutes, relief granted) rather than simple scalar labels.
- **Expert Human Review:** Ground truth creation often requires domain specialists, which limits dataset size and consistency (Nori et al., 2023).

8.2 Automatic Evaluation Metrics

While human evaluation remains the gold standard, several automatic metrics have been employed across domains:

8.2.1 Lexical Overlap Metrics

- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation): For summarization and content selection (lin).
- **BLEU** (Bilingual Evaluation Understudy): For generation quality, though limited for long-form outputs (Papineni et al., 2002).

8.2.2 Semantic Similarity Metrics

- **BERTScore**: Measures token-level semantic similarity using contextual embeddings (Zhang et al., 2020).
- **BLEURT**: Learned quality evaluation trained on human judgments (Sellam et al., 2020).

8.2.3 Reasoning and Consistency Metrics

- **Faithfulness Metrics:** Entity-level correctness for factual extraction tasks (Maynez et al., 2020).
- **Self-Consistency:** Agreement across multiple CoT reasoning samples (Wang et al., 2023).
- **TruthfulQA:** Specifically designed to capture hallucination susceptibility (Lin et al., 2022).

8.2.4 Prompt-Specific Metrics

- **Prompt Sensitivity Metrics:** Evaluates model stability across prompt variants (Zhao et al., 2021; Webson and Pavlick, 2022).
- **Calibration Error:** Measures confidence alignment for probabilistic predictions.

8.3 Domain-Specific Benchmarks

In recent years, domain-specific evaluation datasets have been developed to better capture DSLM task complexities:

8.3.1 Legal Domain Benchmarks

- **LegalBench** (Chalkidis et al., 2020): Comprehensive evaluation across statutory reasoning, case summarization, and legal entailment.
- **MultiEURLEX** (Chalkidis et al., 2021): Multilingual legal document classification across European legal corpora.

8.3.2 Biomedical Domain Benchmarks

- **MedQA** (Jin et al., 2020): Medical board exam QA dataset evaluating clinical reasoning.
- **PubMedQA** (Jin et al., 2019): Biomedical factual consistency benchmark.
- **RadQA and RadGraph** (Hu et al., 2025): Radiology-specific report generation and reasoning.

8.3.3 Financial Domain Benchmarks

- **FinQA** (Chen et al., 2021): Multi-step financial reasoning with tables and texts.
- **TAT-QA** (Zhu et al., 2021): Financial table-based QA.

8.3.4 Scientific Literature Benchmarks

- **SciBench** (Wang et al., 2024): Evaluates scientific multi-hop reasoning, hypothesis chaining, and cross-domain synthesis.
- **NarrativeQA Science Subset** (Kočiský et al., 2018): Multi-document synthesis for scientific narratives.

8.4 Human Evaluation Protocols

In many DSLM applications, automatic metrics fall short, and expert human evaluation remains necessary. Key considerations include:

- **Accuracy:** Does the model produce correct answers under domain standards?
- **Reasoning Validity:** Are intermediate reasoning steps logical and legally or medically valid?
- **Factual Grounding:** Does the output rely on verifiable domain knowledge or invent facts?
- **Safety and Ethics:** Particularly for medical and financial outputs, adherence to guidelines and risk disclosures is critical.
- **Inter-Annotator Agreement:** Measuring reliability across expert annotators.

While human evaluation ensures high-quality assessments, it remains costly, slow, and difficult to scale across domains. Recent efforts combine human-in-the-loop pipelines with automatic heuristics to balance scalability and expert oversight (Madaan et al., 2023; Shi et al., 2024).

9 Future Directions

While prompting for Domain-Specific Language Models (DSLMs) has achieved significant progress, several important research avenues remain open. This section outlines key future directions that can advance the field both theoretically and practically.

9.1 Unified Multi-Domain Prompting Frameworks

Current DSLMs are often trained for isolated domains (law, medicine, finance), limiting cross-domain reasoning capacity. Future work should explore:

- Joint multi-domain instruction-tuning pipelines.
- Unified meta-prompting architectures capable of adapting to task specifications across legal, biomedical, scientific, and financial domains.
- Transfer learning strategies where reasoning chains discovered in one domain are leveraged to bootstrap others.

Such unified models may also reduce model proliferation and facilitate regulatory oversight.

9.2 Automated Prompt Engineering and Optimization

Manual prompt design remains time-consuming and heavily reliant on domain expertise. Automated solutions include:

- Neural prompt search algorithms for domain-specific instruction discovery.
- Reinforcement Learning with Human Feedback (RLHF) pipelines tailored for DSLMs.
- Self-generated demonstrations and CoT self-discovery mechanisms (Zelikman et al., 2022).
- Gradient-based soft prompt optimizers for low-resource domains (Lester et al., 2021; Li and Liang, 2021a).

Reducing human dependency in prompt design will enable more scalable DSLM deployment.

9.3 Causal and Trustworthy Reasoning Chains

A critical open question is how to enforce reliable multi-step reasoning:

- Verifiable CoT pipelines that produce formally checkable intermediate steps.
- Domain-specific logic constraints embedded into reasoning paths.
- Hybrid symbolic-neural models to enforce legal, medical, or financial regulations.
- Certification of reasoning chains via external expert validators or programmatic verifiers (Madaan et al., 2023).

Such architectures will be essential for AI deployment in safety-critical systems.

9.4 Continual Domain Adaptation

DSLMS must keep pace with evolving regulations, guidelines, and discoveries. Future research should explore:

- Retrieval-augmented continual learning pipelines with dynamically updated corpora.
- Few-shot continual learning techniques that require minimal retraining.
- Lifelong prompt adaptation frameworks that allow DSLMs to track policy changes, legal amendments, or scientific breakthroughs.

Dynamic DSLMs will better serve real-world practitioners who operate in constantly shifting knowledge environments.

9.5 Benchmark Development and Evaluation Standards

Robust evaluation remains a bottleneck across DSLM prompting research. Needed improvements include:

- Expansion of existing benchmarks to cover reasoning, fairness, and cross-domain tasks.
- Expert-annotated evaluation corpora for edge-case assessment.
- Development of model-agnostic reasoning verifiers to measure CoT validity (Wang et al., 2023; Guha et al., 2023).
- Public leaderboards and open-source DSLM benchmarks to drive reproducible research.

Community-wide evaluation standards will promote transparency, replicability, and reliable DSLM development.

9.6 Policy-Aware DSLM Governance

Finally, future prompting research must actively engage with emerging AI governance frameworks:

- Regulatory compliance enforcement via policy-constrained prompting.
- Explainability mechanisms to allow regulators to audit DSLM outputs.
- Bias detection pipelines tailored to domain-specific fairness metrics.
- Multi-stakeholder governance involving AI researchers, domain experts, legal scholars, and policymakers.

Responsible prompting innovation will require both technical and institutional safeguards.

10 Conclusion

Prompting has emerged as a powerful and flexible paradigm for adapting large language models to specialized domains where traditional supervised fine-tuning remains costly or impractical. In this survey, we presented a comprehensive review of prompting strategies tailored for Domain-Specific Language Models (DSLMs), covering both discrete and continuous prompting, reasoning-augmented frameworks such as Chain-of-Thought, Tree-of-Thought, Graph-of-Thought, Buffer-of-Thought, and recent self-reflective architectures.

We systematically examined domain-specific prompting applications across legal, biomedical, financial, scientific, and industrial domains, highlighting unique challenges in prompt design, hallucination mitigation, long-context reasoning, and evaluation. We further discussed emerging trends such as retrieval-integrated prompting, soft prompt compression, automated prompt discovery, and self-consistency frameworks, which collectively point toward increasingly sophisticated and trustworthy DSLM architectures.

While DSLMs offer tremendous potential for high-stakes domains, significant challenges remain regarding prompt stability, safety, factual grounding, and regulatory compliance. We advocate for

continued research that integrates domain expertise, automated prompt optimization, reasoning verifiability, and multi-stakeholder governance to ensure responsible and impactful deployment of domain-specific language models.

We hope that this survey serves as both a reference and a roadmap for researchers, practitioners, and policymakers working at the intersection of prompting, domain adaptation, and trustworthy AI.

References

Mousumi Akter, Erion Çano, Erik Weber, Dennis Dobler, and Ivan Habernal. 2025. A comprehensive survey on legal summarization: Challenges and future directions. *Preprint*, arXiv:2501.17830.

Judie Attard, Fabrizio Orlandi, Simon Scerri, and Sören Auer. 2015. A systematic review of open government data initiatives. *Government Information Quarterly*, 32.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, and 95 others. 2022. On the opportunities and risks of foundation models.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020a. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020b. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. 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.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. Lex-files and legallama: Facilitating english multinational legal language model development. *Preprint*, arXiv:2305.07507.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. Places: Prompting language models for social conversation synthesis. *Preprint*, arXiv:2302.03269.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. pages 3697–3711.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. Saullm-7b: A pioneering large language model for law. *Preprint*, arXiv:2403.03883.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Chenlong Deng, Zhicheng Dou, Yujia Zhou, Peitian Zhang, and Kelong Mao. 2024. An element is worth a thousand words: Enhancing legal case retrieval by incorporating legal elements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2354–2365, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.

Koustava Goswami, Lukas Lange, Jun Araki, and Heike Adel. 2023. Switchprompt: Learning domain-specific gated soft prompts for classification in low-resource domains. *Preprint*, arXiv:2302.06868.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Suchin Gururangan, Ana Marasović, 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*, pages 8342–8360, Online. Association for Computational Linguistics.

Francisco Guzmán, Ahmed Abdelali, Irina Temnikova, Hassan Sajjad, and Stephan Vogel. 2015. How do humans evaluate machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 457–466, Lisbon, Portugal. Association for Computational Linguistics.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021. Ptr: Prompt tuning with rules for text classification.

Danqing Hu, Shanyuan Zhang, Qing Liu, Xiaofeng Zhu, and Bing Liu. 2025. Large language models in summarizing radiology report impressions for lung cancer in chinese: Evaluation study. *Journal of Medical Internet Research*, 27:e65547.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. pages 2567–2577.

Abhinav Joshi, Akshat Sharma, Sai Kiran Tanikella, and Ashutosh Modi. 2023. U-CREAT: Unsupervised case retrieval using events extrAcTion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13899–13915, Toronto, Canada. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. volume 6, pages 317–328, Cambridge, MA. MIT Press.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Mandar Kulkarni, Praveen Tangarajan, Kyung Kim, and Anusua Trivedi. 2024. Reinforcement learning for optimizing rag for domain chatbots. *Preprint*, arXiv:2401.06800.

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023. Prompted llms as chatbot modules for long open-domain conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace. 2021. Does BERT pretrained on clinical notes reveal sensitive data? pages 946–959.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. pages 3045–3059.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Xiang Lisa Li and Percy Liang. 2021a. Prefix-tuning: Optimizing continuous prompts for generation. *ACL*.

Xiang Lisa Li and Percy Liang. 2021b. Prefix-tuning: Optimizing continuous prompts for generation.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. pages 3214–3252.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Dhagash Mehta, Stefano Pasquali, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, and 5 others. 2024. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *Preprint*, arXiv:2305.18703.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Preprint*, arXiv:2107.13586.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alejandro Peña, Aythami Morales, Julian Fierrez, Ignacio Serna, Javier Ortega-Garcia, Íñigo Puente, Jorge Córdova, and Gonzalo Córdova. 2023. *Leveraging Large Language Models for Topic Classification in the Domain of Public Affairs*, page 20–33. Springer Nature Switzerland.

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022a. Legal case document summarization: Extractive and abstractive methods and their evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1048–1064, Online only. Association for Computational Linguistics.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022b. Legal case document summarization: Extractive and abstractive methods and their evaluation.

Tejpalsingh Siledar, Swaroop Nath, Sankara Muddu, Rupasai Rangaraju, Swaprava Nath, Pushpak Bhattacharyya, Suman Banerjee, Amey Patil, Sudhanshu Singh, Muthusamy Chelliah, and Nikesh Garera. 2024. One prompt to rule them all: LLMs for opinion summary evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12119–12134, Bangkok, Thailand. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, and 11 others. 2022a. Large language models encode clinical knowledge.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, and 11 others. 2022b. Large language models encode clinical knowledge.

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 Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023a. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

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 Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023b. Towards expert-level medical question answering with large language models.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, and 41 others. 2022. Lamda: Language models for dialog applications. *Preprint*, arXiv:2201.08239.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Yi-Lin Tuan, Sajjad Beygi, Maryam Fazel-Zarandi, Qiaozi Gao, Alessandra Cervone, and William Yang Wang. 2022. Towards large-scale interpretable knowledge graph reasoning for dialogue systems. pages 383–395.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. 2022. Spot: Better frozen model adaptation through soft prompt transfer.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang.

2024. Scibench: Evaluating college-level scientific problem-solving abilities of large language models.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022. Learning to prompt for continual learning.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12060–12075, Singapore. Association for Computational Linguistics.

Zihan Xu, Haotian Ma, Gongbo Zhang, Yihao Ding, Chunhua Weng, and Yifan Peng. 2025. Natural language processing in support of evidence-based medicine: A scoping review.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *Preprint*, arXiv:2306.06031.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E. Gonzalez, and Bin CUI. 2024. Buffer of thoughts: Thought-augmented reasoning with large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *Preprint*, arXiv:2402.18013.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning.

Haopeng Zhang, Philip S. Yu, and Jiawei Zhang. 2024. A systematic survey of text summarization: From statistical methods to large language models. *Preprint*, arXiv:2406.11289.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Zheng Zhao, Eric Wallace, Shi Wang, and 1 others. 2021. Calibrate before use: Improving few-shot performance of language models. In *ICML*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. *CoRR*, abs/2104.08671.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024a. Self-discover: Large language models self-compose reasoning structures. *Preprint*, arXiv:2402.03620.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024b. Lawgpt: A chinese legal knowledge-enhanced large language model. *Preprint*, arXiv:2406.04614.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

Xue Zongyue, Liu Huanghai, Hu Yiran, Kong Kangle, Wang Chenlu, Liu Yun, and Shen Weixing. 2023. Leec: A legal element extraction dataset with an extensive domain-specific label system. *Preprint*, arXiv:2310.01271.