Sketch-of-Thought: Efficient LLM Reasoning with Adaptive Cognitive-Inspired Sketching

Anonymous ACL submission

Abstract

Recent advances in large language models (LLMs) have enabled strong reasoning capabilities through Chain-of-Thought (CoT) prompting, which elicits step-by-step problem solving, but often at the cost of excessive verbosity in intermediate outputs, leading to increased computational overhead. We propose *Sketch-of-Thought* (SoT), a prompting framework that integrates cognitively inspired reasoning paradigms with linguistic constraints to reduce token usage while preserving reasoning accuracy. SoT is designed as a flexible, modular approach and is instantiated with three paradigms—*Conceptual Chaining*, *Chunked Symbolism*, and *Expert Lexicons*—each tailored to distinct reasoning tasks and selected dynamically at test-time by a lightweight routing model. Across 15 reasoning datasets spanning multiple domains, languages, and modalities, SoT achieves token reductions of up to 78% with minimal accuracy loss. In tasks such as mathematical and multi-hop reasoning, it even improves accuracy while shortening outputs.

1 Introduction

Large language models (LLMs) have become central to a wide range of complex reasoning tasks across diverse domains, such as mathematics, science, and commonsense inference (Bubeck et al., 2023; Zhao et al., 2024). Even without dedicated training for reasoning, these models often exhibit emergent capabilities when prompted to decompose problems into intermediate steps (Wei et al., 2023). Chain-of-Thought (CoT) prompting (Wei et al., 2023) exemplifies this approach by encouraging step-by-step natural language reasoning, which has been shown to significantly improve performance on tasks such as logical inference and numerical problem solving (Sprague et al., 2024).

Despite its benefits, CoT often produces verbose outputs that dramatically increase token usage and computational overhead, making it less suitable for latency- or budget-constrained deployment scenarios (Nayab et al., 2025; Arora and Zanette, 2025). More sophisticated strategies, such as Self-Consistency (Wang et al., 2023b), Tree-of-Thoughts (Yao et al., 2023), and Graph-of-Thoughts (Besta et al., 2024), further expand the reasoning process via structured exploration, but tend to exacerbate inefficiencies in token usage.

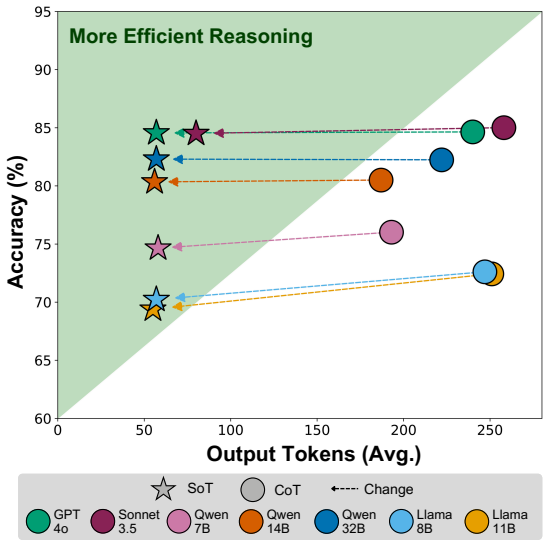

Figure 1: A comparison of accuracy and token usage in Chain-of-Thought (CoT) (Wei et al., 2023) and the proposed Sketch-of-Thought (SoT). Average scores for model performance across 15 datasets. Shaded region represents more efficient reasoning.

To tackle these limitations, we introduce *Sketch-of-Thought* (SoT), a prompting framework that rethinks how language models externalize reasoning. Inspired by cognitive science, particularly the use of symbolic *sketches* as efficient mental intermediaries (Goel, 1995), SoT guides models to produce concise, structured reasoning steps that capture essential logic while avoiding full-sentence elaboration. These representations are analogous to mathematical notation or expert shorthand, preserving semantic fidelity while minimizing redundancy.

Code: URL provided upon paper acceptance.

To implement this framework, we define three cognitively motivated reasoning paradigms: *Conceptual Chaining*, based on associative memory; *Chunked Symbolism*, grounded in working memory theory; and *Expert Lexicons*, inspired by domain-specific schemas used by specialists. Each paradigm is designed for a distinct class of reasoning tasks and is implemented using training-free prompts. To support adaptive paradigm selection, we incorporate a lightweight routing model that analyzes query structure to determine the most suitable reasoning style at inference time.

We extensively evaluate SoT on 15 reasoning datasets spanning mathematical, commonsense, logical, multi-hop, scientific, and medical domains. Experimental results show that SoT reduces output token usage by up to 78% compared to traditional CoT prompting, with no significant loss in accuracy—and even improving performance in some domains. Additional multilingual and multimodal evaluations demonstrate SoT's ability to generalize across both languages and input modalities.
Our key contributions are as follows:

- We introduce *Sketch-of-Thought* (SoT), a prompting framework that leverages cognitively inspired reasoning paradigms to produce concise and structured model outputs.
- We present a lightweight routing model that dynamically selects the optimal reasoning paradigm based on the input query's structure and semantics.
- On a battery of tests, we show that SoT significantly reduces token usage while maintaining or improving accuracy across diverse datasets, models, languages, and modalities.

2 Method

This section outlines the technical implementation of *Sketch-of-Thought* (SoT), a framework designed to improve the efficiency of reasoning in large language models while preserving performance.

2.1 Preliminary

We first provide a brief overview of large language models (LLMs) and their application to reasoning.

Large Language Models Large language models (LLMs) are trained on massive text corpora to predict the next token in a sequence. Formally, an LLM with parameters θ receives an input token sequence x and produces an output sequence y, denoted as follows: $y = \text{LLM}_\theta(x)$. While LLMs exhibit strong performance across a variety of domains, their computational cost increases with the number of tokens processed and generated, making efficiency a critical concern.

Reasoning via Prompting Recent LLMs exhibit emergent reasoning capabilities—the ability to solve multi-step problems without task-specific training—when guided by carefully designed prompts. Letting P denote a prompt with instructions, exemplars, and the input query, the process can be formalized as $r = \text{LLM}_\theta(P(q))$, where r includes both the reasoning trace $[s_1, s_2, \ldots, s_m]$ and final answer a, as follows: $r = [s_1, s_2, \ldots, s_m, a]$. Chain-of-Thought (CoT) (Wei et al., 2023) encourages decomposition into natural language steps (see the box below for an example).

Traditional CoT

```
Q: If a train travels 60 miles per hour for 3 hours,
how far does it go?
A: <think> 1. I understand we need to find the total
distance traveled by: A train moving at 60 miles per
hour for a duration of 3 hours.
2. To calculate the distance, I'll use the formula:
Distance = Speed × Time
Distance = 60 miles/hour × 3 hours
3. Now I'll perform the calculation:
Distance = 60 × 3 = 180 miles
4. Verification:
This makes sense because the train moves 60 miles
each hour.  After 3 hours, it will have covered 3
times that distance. </think>
Answer: 180 miles
```

More advanced variants, for example, Tree-of-Thoughts and Graph-of-Thoughts (Yao et al., 2023; Besta et al., 2024), extend this idea by exploring alternative structural formulations. Despite their stylistic differences (leading to performance gains), all CoT-style prompting methods share a common limitation: increased token usage and reduced efficiency, when compared to standard direct-answer prompting (Arora and Zanette, 2025).

2.2 Sketch-of-Thought (SoT)

Sketch-of-Thought (SoT) mitigates reasoning inefficiency by restructuring how models express intermediate steps. Unlike prior methods that reduce prompt length via input compression (Jiang et al., 2023; Huang et al., 2024), SoT compresses the output reasoning process using cognitively inspired prompts that elicit concise, structured steps.

Formally, we define different paradigm-specific prompt templates P_{SoT}, which steer the model to produce sketched reasoning: $[\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_m, a] = \text{LLM}_\theta(P_{\text{SoT}}(q))$, where each $\hat{s}_i$ conveys the same

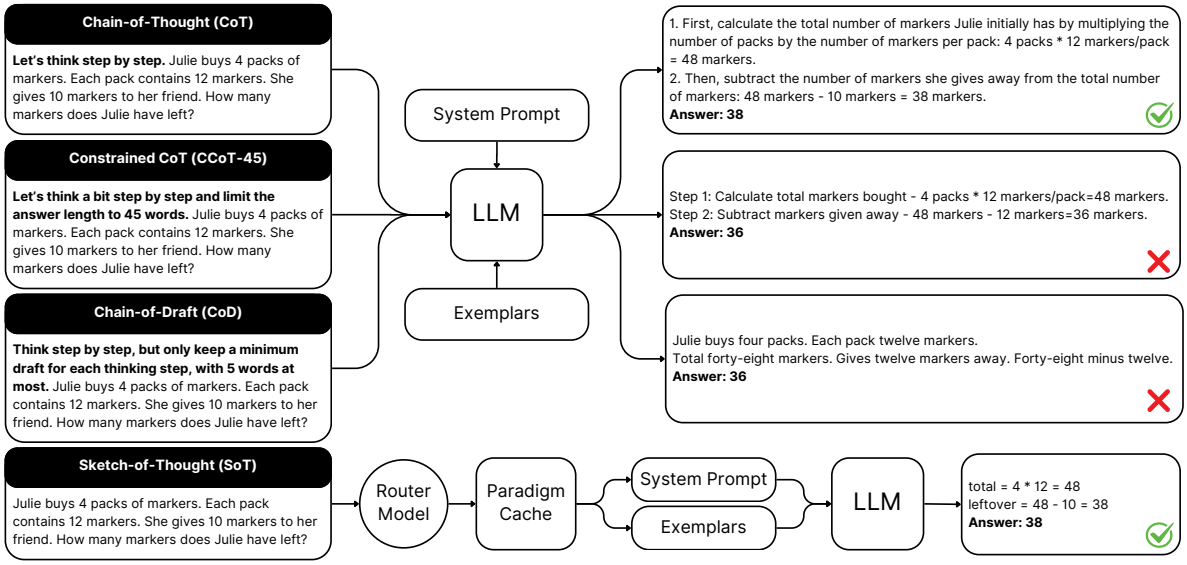

Figure 2: **Illustration of reasoning workflows**, including the input format, intermediate reasoning structure, and output style, across four prompting methods: Chain-of-Thought (CoT) (Wei et al., 2023), Constrained CoT (CCoT) (Nayab et al., 2025), Chain-of-Draft (CoD) (Xu et al., 2025), and Sketch-of-Thought (SoT). While CoT produces verbose natural language traces, CCoT and CoD apply explicit constraints on reasoning length. SoT introduces paradigm-guided sketching, yielding more compact yet structured intermediate steps via dynamic routing.

logical content as s_i (from CoT, for example), but using significantly fewer tokens, i.e., $|\hat{s}| < |s|$. These prompts enforce both linguistic constraints and cognitive structuring tailored to the task type.

As an initial realization of SoT, we create three reasoning paradigms inspired by cognitive science, each designed to align with distinct patterns found across a range of reasoning tasks.

Conceptual Chaining. Rooted in cognitive science principles of how humans connect and retrieve related information, this paradigm creates concise logical sequences between key concepts. It draws from episodic buffer integration (Baddeley, 2000), the cognitive mechanism that temporarily holds and links information from different sources, and associative memory networks (Anderson, 1983), which describe how activating one concept automatically triggers related concepts in our minds (like how thinking of "rain" might immediately evoke "umbrella"). Conceptual Chaining extracts essential terms and presents reasoning as direct step-by-step pathways with minimal text.

```
Conceptual Chaining
Q: What is the name of the currency used in Seoul?
A: <think> #Seoul → #South Korea → Won </think>
Answer: Korean Won
```

Conceptual Chaining is particularly effective for commonsense, multi-hop, logical, and scientific reasoning tasks, where establishing structured relationships between ideas is critical.

Chunked Symbolism. Based on working memory chunking theory (Miller, 1956), this paradigm organizes numerical and symbolic reasoning into compact, structured steps. This seminal cognitive science research showed that humans can only hold about 7±2 (i.e., 5 to 9) distinct items in working memory at once, but we overcome this limitation by "chunking" related information into meaningful units—like remembering phone numbers as area code, prefix, and line number instead of 10 separate digits. Chunked Symbolism applies this principle by condensing mathematical reasoning into dense symbolic representations that pack more information into fewer tokens. It systematically extracts variables, equations, and operations while eliminating verbose explanations, transforming natural language into a structured mathematical shorthand that preserves logical flow.

```
Chunked Symbolism
Q: A car accelerates at 2.5 m/s^2 for 10 seconds. If its initial velocity was 15 m/s, what is its final velocity?
A: <think> a = 2.5 m/s^2, t = 10 s, vi = 15 m/s vf = 15 + (2.5 × 10), vf = 40 m/s </think>
Answer: 40 m/s
```

Chunked Symbolism excels in mathematical and arithmetic reasoning problems where symbolic notation naturally compresses complex concepts.

Expert Lexicons. Inspired by expert schema research (Chi et al., 1981), this paradigm leverages domain-specific shorthand and specialized notation

to condense reasoning. This research demonstrated that experts in any field organize knowledge differently than novices—they develop mental frameworks (schemas) that allow them to quickly recognize patterns and use specialized terminology to communicate efficiently with peers. For example, a physician can convey complex medical conditions with a few acronyms that would require paragraphs of explanation for non-specialists. Expert Lexicons mimics this cognitive efficiency by employing domain-specific abbreviations, notation, and symbols that pack multiple concepts into single tokens. The example below demonstrates how domain-specialized reasoning can be compressed into concise notation while preserving the critical logical connections.

Expert Lexicons

```
Q: A patient with STEMI is given MONA therapy. They
are allergic to aspirin. Are they at risk with this
treatment?
A: <think> STEMI → ST-Elevation MI, MONA → Morphine,
O2, Nitrates, Aspirin, so Aspirin ∈ MONA </think>
Answer: Yes
```

Expert Lexicons is particularly suited for technical disciplines, specialized reasoning tasks, and scenarios where domain expertise enables significant information compression.

2.3 Adaptive Paradigm Selection

While manual selection among three paradigms is possible for each query based on heuristic rules, such an approach is impractical at scale. Instead, we introduce a lightweight routing model that selects the paradigm dynamically based on semantic and structural features of the input query.

Given a query q, the routing process is denoted as follows: $P_{\text{SoT}} = \texttt{ROUTER}(q)$, where P_{SoT} refers to the selected paradigm's prompt-exemplar pair and ROUTER denotes the router model. We use DistilBERT (Sanh et al., 2020) as the backbone model due to its strong performance-efficiency trade-off. This routing approach ensures minimal inference overhead while preserving task alignment.

Router Training We train the router model using 14,200 machine-labeled examples drawn from the training splits of the datasets outlined in Section 3.1. Each sample is labeled using GPT-4o (OpenAI, 2024), guided by a classification prompt derived from the paradigm definitions in Section 2.2. We provide this classification prompt in Appendix B.5.

To avoid overwhelming the router with irrelevant input, we replace any long or non-textual context (e.g., images or documents) with a special placeholder token (e.g., [CONTEXT HERE]). This ensures that the model focuses solely on the question itself, which typically contains sufficient cues for determining the appropriate reasoning style.

3 Experimental Setup

3.1 Datasets

To ensure a comprehensive evaluation, we validate Sketch-of-Thought (SoT) across 15 datasets spanning six categories of reasoning, following the taxonomy introduced by Sun et al. (2024), as follows: **Mathematical Reasoning** includes GSM8K, SVAMP, AQUA-RAT, and DROP (Cobbe et al., 2021; Patel et al., 2021; Ling et al., 2017; Dua et al., 2019); **Commonsense Reasoning** includes CommonsenseQA, OpenbookQA, and StrategyQA (Talmor et al., 2019; Mihaylov et al., 2018; Geva et al., 2021); **Logical Reasoning** includes LogiQA and ReClor (Liu et al., 2020; Yu et al., 2020); **Multi-Hop Reasoning** includes HotPotQA and MuSiQue-Ans (Yang et al., 2018; Trivedi et al., 2022); **Scientific Reasoning** includes QASC and Worldtree (Khot et al., 2020; Jansen et al., 2018); and **Medical Reasoning** includes PubMedQA and MedQA (Jin et al., 2019, 2020).

Beyond English textual reasoning, we include two additional evaluation tracks: a multilingual experiment using MMLU and its professionally translated variant MMMLU (Hendrycks et al., 2021), and a multimodal experiment using GQA (Hudson and Manning, 2019) and the image-based subset of ScienceQA (Lu et al., 2022). Further details regarding the datasets are provided in Appendix A.1.

3.2 Baselines

We mainly compare SoT against three established prompting-based reasoning strategies. Chain-of-Thought (CoT) (Wei et al., 2023) elicits step-by-step natural language reasoning. Constrained CoT (CCoT) (Nayab et al., 2025) introduces a global verbosity constraint, limiting the total reasoning chain to a fixed number of words—in our case, 45 words (CCoT-45). Chain-of-Draft (CoD) (Xu et al., 2025) adopts a similar compression strategy but imposes constraints at the step level, requiring each intermediate step be no longer than five words.

3.3 Implementation Details

A diverse set of instruction-tuned LLMs is selected, spanning both open-weight and proprietary offerings. These include Qwen-2.5 in 7B, 14B, and

Table 1: **Main Experimental Results.** Results are shown for Sketch-of-Thought (SoT), Chain-of-Thought (CoT) (Wei et al., 2023), Constrained Chain-of-Thought (CCoT) (Nayab et al., 2025), and Chain-of-Draft (CoD) (Xu et al., 2025). Results are grouped by reasoning type, with each entry representing the average over all associated datasets. "Acc" denotes accuracy and "Tkn" denotes the number of output tokens. In the Overall section, we report two additional metrics: the token reduction percentage (shown as "Red.") and the change in accuracy between CoT and the baseline (shown as "Δ"). The best results are in **bold** and the second-best are underlined.

		Reasoning Task															
		Mathematical		Commonsense		Logical		Multi-Hop		Scientific		Specialized		Overall			
	Method	Acc	Tkn	Acc	Tkn	Acc	Tkn	Acc	Tkn	Acc	Tkn	Acc	Tkn	Acc ↑	Tkn ↓	Red.↑	Δ↑
Qwen 2.5-32B	CoT	84.17	222	91.48	177	71.23	298	79.44	155	92.89	213	67.66	292	82.24	222	-	-
	CoD	71.94	53	89.48	38	72.89	45	80.00	41	90.00	42	58.89	47	77.32	45	**79.75**	-4.92
	CCoT	80.50	76	88.82	49	72.78	60	80.11	54	88.89	49	57.66	65	79.16	61	72.56	-3.08
	SoT	86.94	88	92.00	34	71.00	66	81.89	43	91.34	31	61.11	63	82.30	57	74.36	**0.06**
Qwen 2.5-14B	CoT	83.00	190	91.41	150	67.00	248	77.67	149	90.89	164	65.11	234	80.50	187	-	-
	CoD	69.22	63	89.04	41	66.22	47	80.44	46	89.44	43	59.00	52	75.61	50	**73.23**	-4.89
	CCoT	81.33	115	90.52	58	70.00	89	78.89	91	89.44	55	61.44	86	79.76	85	54.49	-0.74
	SoT	82.72	78	90.89	37	67.44	63	79.89	45	90.89	37	62.56	63	80.34	56	70.02	**-0.16**
Qwen 2.5-7B	CoT	77.40	181	87.92	158	63.22	279	76.78	138	86.44	184	57.00	247	76.02	193	-	-
	CoD	66.83	57	84.74	37	64.34	49	76.11	43	87.00	39	55.89	48	72.55	46	**76.14**	-3.47
	CCoT	78.00	81	64.15	45	63.66	63	78.89	53	62.78	44	50.33	61	67.72	60	68.87	-8.30
	SoT	77.05	73	83.78	30	59.78	61	77.22	45	85.00	27	58.00	106	74.64	58	69.91	**-1.38**
Llama 3.1-8B	CoT	72.56	235	81.92	209	51.22	292	74.56	193	85.78	260	65.00	323	72.61	247	-	-
	CoD	55.28	73	80.67	45	47.22	58	73.22	49	81.00	47	66.22	55	66.56	56	**77.31**	-6.05
	CCoT	65.22	88	80.89	58	51.00	73	75.45	60	85.00	57	68.11	73	70.84	70	71.64	**-1.77**
	SoT	64.67	78	81.41	36	48.11	71	77.11	44	83.56	35	66.44	63	70.22	57	76.91	-2.39
Llama 3.2-11B	CoT	70.55	232	82.74	216	50.33	297	73.45	198	85.78	263	68.44	334	72.43	251	-	-
	CoD	56.17	67	80.89	43	48.22	51	74.00	46	79.44	44	65.00	50	66.71	52	**79.25**	-5.72
	CCoT	64.56	79	80.81	59	51.89	69	73.00	62	84.22	57	68.34	71	70.37	67	73.27	**-2.06**
	SoT	64.50	75	81.48	35	45.34	69	77.89	44	79.44	36	66.56	64	69.39	55	78.06	-3.04
GPT-4o	CoT	85.44	240	92.74	200	74.78	311	81.56	156	93.22	240	75.22	308	84.64	240	-	-
	CoD	83.17	71	87.11	50	71.56	62	82.56	53	90.67	55	46.33	63	78.41	60	74.95	-6.23
	CCoT	83.72	93	90.59	63	71.22	69	82.33	70	90.22	63	56.22	71	80.44	74	69.11	-4.20
	SoT	86.17	69	92.52	39	73.22	80	84.78	47	92.56	39	72.44	61	84.55	57	**76.20**	**-0.09**
Claude Sonnet 3.5	CoT	87.11	233	91.26	242	75.22	314	81.67	206	93.89	264	75.67	321	85.01	258	-	-
	CoD	82.00	78	91.33	61	75.78	96	82.00	63	91.33	67	76.22	105	83.51	77	**70.16**	-1.50
	CCoT	82.94	97	72.44	80	64.67	91	80.89	85	68.33	83	55.78	103	72.56	90	65.12	-12.45
	SoT	84.06	85	91.11	59	75.00	112	84.44	57	91.78	62	77.78	116	84.50	80	68.99	**-0.51**
All Models	CoT	80.03	219	88.50	193	64.71	291	77.88	171	89.84	227	67.73	294	79.06	228	-	-
	CoD	69.23	66	86.18	45	63.75	58	78.33	49	86.98	48	61.08	60	74.38	55	**75.83**	-4.68
	CCoT	76.61	90	81.17	59	63.60	73	78.51	68	81.27	58	59.70	76	74.41	72	67.87	-4.66
	SoT	78.02	78	87.60	39	62.84	75	80.46	46	87.80	38	66.41	77	77.99	60	73.49	**-1.07**

32B variants (Team, 2024), LLaMA-3.1-8B (Meta, 2024a), LLaMA-3.2-11B (Meta, 2024b), GPT-4o (OpenAI, 2024), and Claude Sonnet 3.5 (Anthropic, 2024). For experiments involving multimodal inputs, we use Qwen-2.5-VL-7B (Team, 2025), which supports visual input processing. Unless otherwise specified, Qwen-2.5-32B serves as the default model for all other experiments. We use a temperature value of 0.5 for all models to balance output stability and diversity. For open-source models, inference is accelerated using FlashAttention2 (Dao, 2023). We sample 150 questions from each dataset for the sake of computational costs, and report the averaged performance over three independent runs per question. For the router model, we fine-tune DistilBERT with cross-entropy loss over 5 epochs, using a batch size of 64 and a learning rate of $2e^{-5}$. During inference, the router processes the core input query. Following previous work, we use few-shot prompting to illustrate the required reasoning style, with exemplars being generated by prompting Qwen-2.5-32B with the method-specific prompt and selecting high-quality outputs. Further information regarding prompts and exemplars can be found in Appendix B.

3.4 Evaluation Protocol

We evaluate using two primary metrics: accuracy and output token count. For multiple-choice, yes/no, or numeric tasks, accuracy is computed via exact match with the ground truth. For open-ended generation, we follow the LLM-as-a-judge paradigm (Liu et al., 2023), using GPT-4o (OpenAI, 2024) to assess correctness. Answers are extracted according to the output format (see Appendix B.2). We analyze efficiency through the total number of generated tokens, including both intermediate reasoning and final answers.

4 Results and Discussion

4.1 Overall Performance

As shown in Table 1, Sketch-of-Thought (SoT) consistently reduces output token count while preserving—or slightly improving— reasoning accuracy across all evaluated models. On average, SoT achieves a token reduction of over 73% relative to CoT, with accuracy deviations typically within 1%. These trends hold across both open-weight models and proprietary models, confirming SoT's generalizability across architectures and model families. Notably, SoT also demonstrates strong stability across settings, consistently balancing token reduction with minimal accuracy variance, unlike other baselines which exhibit greater fluctuations.

4.2 Model-wise Trends

Performance gains with SoT are especially notable in the Qwen series. On Qwen-2.5-32B, SoT achieves 82.30% accuracy—slightly above CoT's 82.24%—while reducing output token count by 74.36%. Similar patterns hold at the 14B and 7B scales, where SoT maintains accuracy within 1.5% of CoT while reducing output length by over 69%. On GPT-4o, SoT achieves 84.55% accuracy—just 0.09% below CoT—while reducing token usage by 76.2%. Claude Sonnet 3.5 shows similar behavior, with SoT reaching 84.50% accuracy versus CoT's 85.01%, alongside a 68.99% reduction in tokens. Results on LLaMA-3.1 and 3.2 indicate stronger compression (up to 78.06%) but slightly wider accuracy gaps (up to 3.0%). These findings confirm that SoT performs reliably across model families, consistently achieving strong token reductions with minimal accuracy degradation.

4.3 Paradigm-Task Alignment

Task-level results indicate that SoT's effectiveness is most pronounced in reasoning settings with inherently compressible logic. In mathematical tasks, SoT matches or exceeds CoT performance across nearly all models. For example, in the Qwen-32B setting, SoT achieves 86.94% accuracy compared to 84.17% for CoT, while reducing average output length from 222 to 88 tokens. These gains are attributable to the effectiveness of the *Chunked Symbolism* paradigm in representing arithmetic reasoning concisely, which is the dominant paradigm for this category (see Section 2.2).

In commonsense and multi-hop reasoning, SoT maintains strong performance while achieving substantial compression. For instance, in the Qwen-32B setting, SoT reaches 92.00% accuracy on commonsense tasks using just 34 tokens on average, compared to 91.48% at 177 tokens under CoT. These improvements are driven by the *Conceptual Chaining* paradigm, which is the prevailing strategy for these categories and effectively captures structured relationships between ideas.

Domain-specialized tasks, such as PubMedQA and QASC, show more variability in accuracy across models, reflecting the inherent complexity of technical reasoning. Nevertheless, the *Expert Lexicons* paradigm remains effective at compressing domain-specific reasoning, often using half as many tokens as CoT while preserving comparable accuracy. Across all categories, SoT maintains competitive performance with far shorter outputs, underscoring its broader adaptability. Further discussion on how paradigms are distributed across datasets can be found in Appendix D.

Table 2: **Results of Extended Approaches.** Comparison of SoT and CoT in extended reasoning pipelines.

Approach	Method	Tkn	Acc	Red.	Δ
Self-Consistency	CoT	680	81.86	-	-
	SoT	176	81.90	**74.1**	**0.04**
Self-Refine	CoT	614	80.53	-	-
	SoT	244	80.80	**60.3**	**0.27**
Multi-Agent Debate	CoT	766	81.87	–	–
	SoT	238	82.44	**68.9**	**0.57**

4.4 Token-Constrained Alternatives

Compared to other compression-focused prompting strategies such as Chain-of-Draft (CoD) and Constrained CoT (CCoT), SoT provides a more favorable trade-off between brevity and performance. Although CoD yields the most aggressive reductions in output length, it suffers notable accuracy degradation—for example, a 6.2% decline on GPT-4o despite a 75% token reduction. CCoT-45 offers more balanced results, but still lags behind SoT in both efficiency and generalization across reasoning types. SoT achieves similar or better accuracy than these methods while maintaining—or exceeding—their compression levels. This demonstrates the benefits of SoT's structured and cognitively grounded approach to reducing verbosity.

4.5 Extended Reasoning Pipelines

To examine SoT's compatibility with ensemble-style reasoning methods, we integrate it into three

Table 3: **Multilingual Results.** Performance comparison of CoT and SoT across different languages.

Lang.	Method	Tkn	Acc	Red.	Δ
Korean	CoT	315	77.27	–	–
	SoT	45	76.26	**85.71**	**-1.01**
Italian	CoT	336	81.50	–	–
	SoT	50	79.50	**85.12**	**-2.00**
German	CoT	309	83.42	–	–
	SoT	46	84.92	**85.11**	**1.50**

Table 4: **Multimodal Results.** Performance comparison of CoT and SoT for multimodal reasoning tasks.

Dataset	Method	Tkn	Acc	Red.	Δ
ScienceQA	CoT	147	83.33	–	–
	SoT	28	82.33	**80.95**	**-1.00**
GQA	CoT	79	76.67	–	–
	SoT	18	72.67	**77.21**	**-4.00**

established frameworks. Self-Consistency (Wang et al., 2023b) aggregates multiple reasoning paths by majority voting to improve answer stability. Self-Refine (Ranaldi and Freitas, 2024) enables iterative refinement of reasoning traces through reflection-based prompting. Multi-Agent Debate (Du et al., 2023) simulates deliberation among independent agents, each producing a rationale before converging on a final answer. In each case, we follow the original methodology but substitute SoT in place of CoT as the core reasoning strategy. Further implementation details, including prompts and hyperparameters, are provided in Appendix C.

Table 2 reports results from integrating SoT into three ensemble reasoning frameworks. In all cases, SoT improves performance relative to CoT, while substantially reducing output length. For instance, in the Self-Refine setting, SoT improves accuracy by 0.27% while generating 60% fewer tokens per response. In the Multi-Agent Debate framework, SoT yields a 0.57% accuracy increase alongside a 68.9% token reduction. These results indicate that SoT can be effectively substituted into more complex, multi-pass prompting pipelines, retaining its advantages in both efficiency and output quality.

4.6 Multilingual Generalization

To test SoT's performance in non-English settings, we conduct a multilingual evaluation using Korean, Italian, and German subsets of MMMLU (Hendrycks et al., 2021). For each language, we select the same set of 100 questions from both the management and astronomy categories to ensure cross-lingual consistency. To maintain consistent paradigm selection across languages, each non-English query is paired with its English counterpart and routed using the same routing model. The selected paradigm prompt and associated exemplars are then translated into the target language using GPT-4o (OpenAI, 2024), preserving both semantic fidelity and structural constraints.

As summarized in Table 3 (which shows cross-linguistic applicability on Korean, Italian, and German), SoT reduces output length by over 85% in all three languages. While accuracy declines slightly in Korean (–1.01%) and Italian (–2.00%), SoT outperforms CoT in German (+1.50%). These findings suggest that the sketching paradigms underlying SoT generalize across linguistic structures and preserve core reasoning logic beyond English.

4.7 Multimodal Robustness

To assess SoT's extensibility to multimodal scenarios, we evaluate its performance using Qwen-2.5-VL-7B (Team, 2025) on 300 multiple-choice samples from both GQA (Hudson and Manning, 2019) and the image-based subset of ScienceQA (Lu et al., 2022). As in the unimodal setting, paradigm selection is handled by the router model. Images and supplementary materials are replaced with a placeholder token during routing (see Section 2.3), allowing the router to focus on the question text. We reuse the same text-only exemplars from the primary experiments.

Results from multimodal evaluations are shown in Table 4. On ScienceQA, SoT reduces output length by 80.95% with only a 1.00% accuracy drop. On GQA, performance decreases by 4.00% but still retains a 77.2% token reduction. The slightly larger drop in GQA likely reflects the difficulty of applying abstract sketching methods to tasks requiring fine-grained visual grounding. Another possible explanation is that the text-only exemplars, while effective in general, may not sufficiently prime the model for vision-intensive reasoning.

4.8 Analysis on Routing

We evaluate the router model's ability to select appropriate reasoning paradigms across the 2,250 samples used in our primary experiments (see Section 3.1). Ground-truth labels are produced by GPT-4o using the same labeling protocol as during training (see Section 2.3). As shown in Figure 3, the model achieves 96.4% overall accuracy, with high

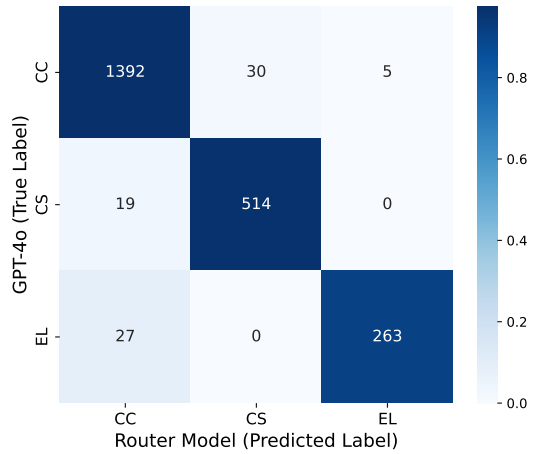

Figure 3: Confusion matrix illustrating the performance of the router model in selecting among the three SoT paradigms. Predictions are compared against GPT-4o-assigned ground truth labels.

recall for the two most common paradigms, *Conceptual Chaining* (0.964) and *Chunked Symbolism* (0.975). Recall for *Expert Lexicons* is slightly lower at 0.907, largely due to class imbalance. However, this asymmetry is expected as *Expert Lexicons* is intentionally applied more conservatively given its specialized nature, and the router defaults to general paradigms in ambiguous cases to reduce risk of misapplication.

5 Related Work

Token-Efficient Reasoning A growing body of work targets the reduction of output length during language model reasoning. Concise Chain-of-Thought (Renze and Guven, 2024) and Constrained CoT (CCoT) (Nayab et al., 2025) apply fixed constraints on the number of steps or words in the reasoning trace. SCOTT (Wang et al., 2023a) uses a two-stage summarization pipeline that compresses verbose CoT outputs into shorter versions. While these methods reduce token usage, they rely on surface heuristics or summary-based rewriting, often reducing clarity. As an orthogonal direction, Coconut (Hao et al., 2024) bypasses token-based reasoning by operating entirely in latent vector space, though this requires modifying model architecture and additional training procedures, not applicable to off-the-shelf LLMs. In contrast, SoT rewrites reasoning steps using compact representations grounded in human cognitive patterns, yielding outputs that are both shorter and interpretable.

Structured Reasoning Strategies Other approaches enhance reasoning by restructuring the generation process itself. Tree-of-Thoughts (Yao et al., 2023) and Graph-of-Thoughts (Besta et al., 2024) treat reasoning as a search over intermediate steps, producing tree- or graph-structured outputs. Self-Consistency (Wang et al., 2023b) improves stability by sampling multiple reasoning paths and selecting the majority answer. While these methods improve accuracy on certain tasks, they often incur significant compute overhead and longer outputs. In contrast, SoT maintains a standard prompting interface while restructuring internal reasoning to achieve efficiency gains without increasing inference complexity.

Prompt Compression and Adaptive Inference Several techniques improve efficiency through prompt compression or selective computation. Chain-of-Draft (CoD) (Xu et al., 2025) uses densely packed natural language reasoning to reduce length, but this often comes at the cost of clarity and yields large performance drops on more complex reasoning tasks. CoT-Influx (Huang et al., 2024) and LLMLingua (Jiang et al., 2023) prune or compress input exemplars to reduce prompt length. Cascaded inference (Yue et al., 2024) and compute-adaptive methods (Arora and Zanette, 2025) dynamically route examples to high-cost inference pipelines only when necessary. SoT differs by addressing compression as a representational design challenge: instead of relying on pruning or selection, it restructures how reasoning is expressed, guided by task-specific cognitive principles.

6 Conclusion

We present Sketch-of-Thought (SoT), a prompting framework that reduces token usage in language model reasoning by up to 76%, preserving accuracy in most tasks and incurring only minor trade-offs in others. SoT leverages cognitively inspired paradigms to generate compact yet semantically faithful reasoning traces, offering a practical alternative to verbose prompting. Extensive experiments across 15 reasoning datasets, multiple languages, and multimodal tasks demonstrate SoT's broad applicability. Its compatibility with ensemble prompting strategies further reinforces its practical utility, particularly in resource-constrained settings. By reframing efficiency as a reasoning design challenge rather than a surface-level compression problem, SoT opens new directions for scalable, cognitively informed prompting.

Limitations and Future Work

Sketch-of-Thought (SoT) is designed for interpretable, efficient reasoning, which also opens several interesting directions for future work.

Following prior work, our use of fixed exemplars per paradigm—intentionally chosen to preserve stylistic consistency and interpretability—may limit adaptability to subtle variations within a task type. Future work may explore annotating (or generating) more examples and using retrieval-augmented exemplar strategies to improve flexibility across reasoning tasks and domains.

Because SoT departs from one-size-fits-all prompting—unlike prior methods such as Chain-of-Thought (CoT) or Chain-of-Draft (CoD)—it introduces the novel challenge of ensuring paradigm coverage across diverse queries. While our current library of three cognitively grounded paradigms demonstrates broad applicability, expanding this set could help capture finer distinctions in specialized reasoning.

Ethics Statement

This work builds on widely used public datasets and large language models (LLMs), with no new data collected. All datasets used in our experiments are publicly available and cited accordingly. Where applicable, we follow dataset authors' intended uses and licensing terms.

While Sketch-of-Thought (SoT) improves the efficiency of model reasoning, we acknowledge that compressing intermediate outputs may affect interpretability in certain high-stakes settings. We encourage caution when applying SoT in domains such as healthcare or legal analysis, where full transparency of reasoning steps may be essential.

Our routing model was trained using annotations generated via GPT-4o, which may reflect biases present in the underlying model. We recommend further evaluation before deployment in sensitive contexts.

References

John R. Anderson. 1983. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22(3):261–295.

Anthropic. 2024. Claude 3.5 sonnet.

Daman Arora and Andrea Zanette. 2025. Training language models to reason efficiently. *Preprint*, arXiv:2502.04463.

A. Baddeley. 2000. The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4(11):417–423.

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.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Michelene T. H. Chi, Paul J. Feltovich, and Robert Glaser. 1981. Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Science*, 5(2):121–152.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint*. ArXiv:2110.14168 [cs].

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *Preprint*, arXiv:2307.08691.

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

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. *arXiv preprint*. ArXiv:1903.00161 [cs].

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *arXiv preprint*. ArXiv:2101.02235 [cs].

Vinod Goel. 1995. *Sketches of Thought*. MIT Press.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *Preprint*, arXiv:2412.06769.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *arXiv preprint*. ArXiv:2009.03300 [cs].

Xijie Huang, Li Lyna Zhang, Kwang-Ting Cheng, Fan Yang, and Mao Yang. 2024. Fewer is more: Boosting llm reasoning with reinforced context pruning. *Preprint*, arXiv:2312.08901.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Peter A. Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton T. Morrison. 2018. WorldTree: A Corpus of Explanation Graphs for Elementary Science Questions supporting Multi-Hop Inference. *arXiv preprint*. ArXiv:1802.03052 [cs].

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. Llmlingua: Compressing prompts for accelerated inference of large language models. *Preprint*, arXiv:2310.05736.

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. *arXiv preprint*. ArXiv:2009.13081 [cs].

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William 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)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A Dataset for Question Answering via Sentence Composition. *arXiv preprint*. ArXiv:1910.11473 [cs].

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program Induction by Rationale Generation : Learning to Solve and Explain Algebraic Word Problems. *arXiv preprint*. ArXiv:1705.04146 [cs].

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning. *arXiv preprint*. ArXiv:2007.08124 [cs].

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.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Meta. 2024a. Introducing llama 3.1: Our most capable models to date.

Meta. 2024b. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. *arXiv preprint*. ArXiv:1809.02789 [cs].

George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97. Place: US Publisher: American Psychological Association.

Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. 2025. Concise thoughts: Impact of output length on llm reasoning and cost. *Preprint*, arXiv:2407.19825.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP Models really able to Solve Simple Math Word Problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Leonardo Ranaldi and Andre Freitas. 2024. Self-refine instruction-tuning for aligning reasoning in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 2325–2347. Association for Computational Linguistics.

Matthew Renze and Erhan Guven. 2024. The benefits of a concise chain of thought on problem-solving in large language models. *Preprint*, arXiv:2401.05618.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *Preprint*, arXiv:2409.12183.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Wu Yuan, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang,

Pheng Ann Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Ji-Rong Wen, Xipeng Qiu, Yike Guo, Hui Xiong, Qun Liu, and Zhenguo Li. 2024. A Survey of Reasoning with Foundation Models. *arXiv preprint*. ArXiv:2312.11562 [cs].

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen2.5-vl.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10:539–554. Place: Cambridge, MA Publisher: MIT Press.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023a. Scott: Self-consistent chain-of-thought distillation. *Preprint*, arXiv:2305.01879.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

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. *Preprint*, arXiv:2201.11903.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. 2025. Chain of draft: Thinking faster by writing less. *Preprint*, arXiv:2502.18600.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. *arXiv preprint*. ArXiv:1809.09600 [cs].

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. *Preprint*, arXiv:2305.10601.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. *arXiv preprint*. ArXiv:2002.04326 [cs].

Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. 2024. Large language model cascades with mixture of thoughts representations for cost-efficient reasoning. *Preprint*, arXiv:2310.03094.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Experimental Setup

A.1 Datasets

All datasets used in our experiments are publicly available and accessed via Hugging Face using the dataset IDs listed in Table 5. Where datasets included multiple subsets, we explicitly specified which subset was used in our experiments. All datasets are used in accordance with their respective licenses and terms of use.

A.2 Model Checkpoints

We use the following model checkpoints in our experiments:

Qwen 2.5 Family

- `Qwen/Qwen2.5-7B-Instruct`
- `Qwen/Qwen2.5-14B-Instruct`
- `Qwen/Qwen2.5-32B-Instruct`
- `Qwen/Qwen2.5-VL-7B-Instruct`

Llama 3 Family

- `meta-llama/Meta-Llama-3-8B-Instruct`
- `meta-llama/Meta-Llama-3-11B-Instruct`

Closed-source Models

- `gpt-4o-2024-11-20`
- `claude-3-5-sonnet-20241022`

All open-weight models were accessed through Hugging Face via the transformers library (Wolf et al., 2020) and evaluated in their instruction-tuned form. Closed-weight models such as GPT-4o and Claude Sonnet 3.5 were accessed through their respective Python API wrappers. All models are used in accordance with their licenses.

A.3 Inference Environment

All experiments were conducted on 2 x A5000 24GB GPUs on a Linux distribution running CUDA 12.1. For inference, we use FlashAttention2 (Dao, 2023) for acceleration. All models were run in bfloat16 precision where supported. No parameter fine-tuning or additional adaptation was applied to the LLMs during experimentation.

A.4 Reproducibility

All experiments were conducted with the same fixed random seed, 42, to ensure reproducibility across runs. We used a consistent temperature of 0.5 for all models across all methods and tasks. For few-shot setups, exemplars were selected prior to evaluation and held constant across all trials. Token counts were measured using the default tokenizer associated with each model's checkpoint. For closed source models accessed via the API, token counts were obtained through the token logs found in the returned inference object.

To support reproducibility, we will release all code, prompts, and evaluation scripts alongside the final camera-ready version of this paper, pending acceptance.

B Prompting Framework

B.1 Prompt Format and Output Conventions

Each paradigm-specific system prompt follows a consistent structure composed of four sections:

Role & Objective Provides background on the paradigm, including its cognitive basis and theoretical motivation. It outlines representative use cases and serves as a semantic primer to help align the model's reasoning style with the paradigm.

Application Steps Describes a step-by-step procedure for applying the paradigm to solve a problem. This includes identifying relevant concepts, performing transformations, and following best practices for structuring the reasoning process.

Rules & Directives Specifies required tone, structure, and formatting constraints. It highlights common failure modes—such as verbosity, redundancy, or incorrect notation—and explicitly defines output style requirements (see Appendix B.2).

Closing Statement Ends with a reminder to adhere to the formatting guidelines, reinforcing the objective of concise, structured reasoning.

B.2 Output Conventions

To ensure consistent evaluation and accurate token-level comparisons, all outputs follow a strict formatting protocol:

- **Answers** must be enclosed in `\boxed{...}`.
- **Reasoning traces** must appear within `<think>` and `</think>` tags.

This formatting allows for reliable programmatic parsing and segmentation of outputs into intermediate reasoning and final answers, supporting reproducibility and enabling fair evaluation across prompting methods and models.

Table 5: **Dataset Information.** Comprehensive details of datasets used for our experiments.

Dataset	Citation	HF ID	Train Split:Subset	Train Size	Test Split:Subset	Test Size
GSM8K	Cobbe et al. (2021)	gsm8k	main:train	1000	main:test	150
SVAMP	Patel et al. (2021)	ChilleD/SVAMP	train	700	test	150
AQUA-RAT	Ling et al. (2017)	aqua_rat	train	1000	test	150
DROP	Dua et al. (2019)	drop	train	1000	validation	150
OpenbookQA	Mihaylov et al. (2018)	openbookqa	train	1000	test	150
StrategyQA	Geva et al. (2021)	ChilleD/StrategyQA	train	1000	test	150
LogiQA	Liu et al. (2020)	lucasmccabe/logiqa	train	1000	test	150
Reclor	Yu et al. (2020)	metaeval/reclor	train	1000	validation	150
HotPotQA	Yang et al. (2018)	hotpot_qa	distractor:train	1000	distractor:validation	150
MuSiQue-Ans	Trivedi et al. (2022)	dgslibisey/MuSiQue	train	1000	validation	150
QASC	Khot et al. (2020)	allenai/qasc	train	1000	validation	150
Worldtree	Jansen et al. (2018)	nguyen-brat/worldtree	train (last 1000)	1000	train (rest)	150
PubMedQA	Jin et al. (2019)	qiaojin/PubMedQA	pqa_labeled (last 500)	500	pqa_labeled (first 150)	150
MedQA	Jin et al. (2020)	bigbio/med_qa	med_qa_en_source:train	1000	med_qa_en_source:validation	150
CommonsenseQA	Talmor et al. (2019)	tau/commonsense_qa	train	1000	validation	150
MMLU	Hendrycks et al. (2021)	cais/mmlu	—	—	test:all	200
MMMLU	Hendrycks et al. (2021)	openai/MMMLU	—	—	test:KO_KR, DE_DE, IT_IT	200
ScienceQA	Lu et al. (2022)	lmms-lab/ScienceQA	—	—	test:ScienceQA-IMG	300
GQA	Hudson and Manning (2019)	lmms-lab/GQA	—	—	val:val_all_images	300

Because all experiments are conducted using instruction-tuned LLMs with no additional fine-tuning to enforce output structure, we explicitly reserve space in both the *Rules & Directives* and *Closing Statement* sections of each system prompt to reinforce these formatting requirements.

In practice, we find that providing exemplars alone is insufficient for enforcing consistent formatting. In early experiments, models frequently omitted structural tags or deviated from the expected format when prompted using exemplars only. After incorporating explicit formatting instructions into the system prompt, the rate of malformed or non-compliant outputs dropped to near zero across all paradigms and model variants.

B.3 Paradigm Prompts

We provide reference versions of our paradigm system prompts for *Conceptual Chaining*, *Chunked Symbolism*, and *Expert Lexicons* in Figures 5, 6, and 7, respectively. Parts of the prompts have been adjusted to render correctly in this document. We direct the interested reader to our public code repositories for full, code-ready prompts.

B.4 In-Context Exemplars

We provide three in-context exemplars for each method evaluated in our study to guide model outputs during inference. For Sketch-of-Thought (SoT), a separate set of exemplars is constructed for each paradigm to match the distinct reasoning styles each paradigm is designed to elicit. Example questions are manually selected to reflect typical tasks associated with each paradigm's target use cases. To construct exemplars, we first generate candidate responses using Qwen-2.5-32B with the corresponding system prompt, then manually select outputs that most faithfully demonstrate the paradigm's intended structure, clarity, and conciseness. This results in three curated exemplars per paradigm.

For baseline methods—Chain-of-Thought (CoT), Constrained CoT (CCoT), and Chain-of-Draft (CoD)—we apply a consistent process. Each method is prompted using its respective strategy, and the most stylistically representative outputs are selected. Because these baselines do not dynamically adapt to the query type, we ensure fair coverage by drawing exemplars from the same three reasoning categories used for SoT (e.g., commonsense, mathematical, specialized). One exemplar is selected per category, yielding a total of three per method. All exemplars are held fixed across all experiments.

B.5 Classification Prompt

The router model used to assign paradigms was trained using GPT-4o-generated labels. The classification prompt presented each query and instructed the model to assign one of the three paradigms based on reasoning characteristics, following the heuristic definitions given in Section 2.2. A reference version of the classification prompt is shown in Figure 8. To conserve space we omit repetitive text in this version. We direct the interested reader to our public code repositories for the full, unabridged classification prompt.

B.6 Extended Strategy Prompts

In addition to the default Sketch-of-Thought (SoT) prompting format, we adapt SoT for two extended reasoning strategies: Self-Refine and Multi-Agent Debate. In both cases, the core SoT paradigm structure is preserved, but the prompting is modified to support multi-turn generation and interaction. In both settings, we provide the model with the same system prompts and exemplars as in the primary experiments.

Self-Refine. The Self-Refine strategy involves two prompts per question: a critique prompt and a refinement prompt. The initial reasoning trace is produced using the appropriate SoT paradigm (selected via the router), after which the model reflects on its output and revises it.

Critique Prompt

```
You are reviewing a response generated using the
<paradigm> reasoning paradigm for the following
question:
Question: <question>
<think> <original reasoning> </think>
Answer: <original answer>
Please identify any flaws, gaps, or unclear steps,
while maintaining the structured, concise format
encouraged by this paradigm. Respond WITHOUT using
<think>...</think> tags or \boxed{}.
```

Refine Prompt

```
You are refining a response originally generated
using the <paradigm> reasoning paradigm for the
following question:
Question: <question>
Original Reasoning:
<think> <original reasoning> </think>
Answer: <original answer>
Critique: <model-generated critique>
Please revise the response using the critique
provided, ensuring your reasoning remains concise,
structured, and consistent with the paradigm. Use
<think>...</think> for reasoning and \boxed{} for
the final answer.
```

Multi-Agent Debate. For the Multi-Agent Debate setup, we preserve the paradigm-specific SoT system prompt and introduce a debate prompt that allows agents to revise their reasoning in response to other agents' answers. The debate prompt is structured to request updated responses while retaining the specified output formatting conventions.

Multi-Agent Debate Prompt

```
You are participating in a multi-agent debate. Other
agents have responded as follows:
#Agent 1:
<think> [agent 1's reasoning] </think>
Answer: [agent 1's answer]
Your previous answer was:
<think> [your previous reasoning] </think>
Answer: [your previous answer]
Would you like to revise your reasoning or stick with
it? Please provide your updated reasoning inside
<think>...</think> tags and your final answer inside
boxed{...}.
```

C Extended Results

C.1 Per-Dataset Results

We report per-dataset results from the primary experiments across all model families and prompting strategies in Table 6. For further information regarding the primary experiments, see Section 3 for the experimental design and Section 4 for the results and discussion.

C.2 Multi-Agent Debate

To evaluate whether Sketch-of-Thought (SoT) remains effective in ensemble-style deliberation, we incorporate it into the Multi-Agent Debate (MAD) framework (Du et al., 2023). This method simulates independent agents answering the same question and iteratively revising their answers through multi-round critique.

Each debate run involves three agents and a maximum of three rounds. In the first round, all agents independently generate answers using SoT prompts selected by the router model (see Section 2.3). In subsequent rounds, each agent receives the other agents' reasoning and has the opportunity to revise its answer using a shared debate prompt (see Appendix B.6).

Debates terminate early if all agents converge on the same answer (based on exact-match or semantically equivalent outputs). If consensus is not reached within three rounds, a majority vote determines the final answer. The rationale of the majority-aligned agent is retained as the final justification.

Notably, we find that using SoT does not have a notable impact on the number of rounds-per-query. For CoT we observed an average of 1.14 rounds-per-query, almost matching the observed 1.11 for SoT. Results are shown in Table 2 and discussed in Section 4.

C.3 Self-Refine

We further evaluate Sketch-of-Thought (SoT) under the Self-Refine framework (Ranaldi and Freitas, 2024), a reflection-based prompting strategy in which a single agent critiques and revises its own reasoning trace. This setup tests whether SoT's structured output format supports iterative refinement without compromising coherence or conciseness.

Each trial consists of a two-step loop: (1) a critique prompt is applied to the model's initial response to identify any flaws or ambiguities, and (2) a refinement prompt is used to generate a revised answer based on the critique. All formatting conventions are used throughout.

The full refinement trace—including critiques, updated outputs, and token logs—is retained for analysis. Prompt details for both critique and refinement phases are provided in Appendix B.6, and results are reported in Table 3.

D Paradigm Assignment Analysis

To better understand how SoT paradigms align with reasoning task types, we analyze the output of our router model across the datasets used in our primary experiments. Table 7 reports the predicted paradigm distribution, dominant paradigm, and its agreement with an expected paradigm label defined based on the paradigm descriptions in Section 2.2.

D.1 Routing Distribution by Dataset

Table 7 presents paradigm distributions for each dataset. These counts reflect router predictions over the 150 samples used per dataset in our primary experiments. The dominant paradigm is defined as the one with the highest frequency within a dataset, and we compare this to the expected paradigm, which is assigned based on prior task categorizations and paradigm design goals.

The router's predictions match expectations in all 15 datasets, with 100% agreement between dominant and expected paradigms. Most datasets are routed to a single paradigm, reflecting high confidence and class purity. In a few edge cases (e.g., DROP, LogiQA, QASC), we observe minor cross-paradigm overlap, though these do not shift the dominant label. This behavior aligns with our router's conservative design, which favors general-purpose paradigms (especially Conceptual Chaining) in ambiguous scenarios.

D.2 Paradigm Alignment Discussion

The paradigm distribution confirms that SoT paradigms align closely with reasoning task categories. As expected, Conceptual Chaining (CC) dominates in commonsense, logical, and multi-hop datasets (e.g., StrategyQA, HotPotQA, Reclor), where relational inference is critical.

Chunked Symbolism (CS) is used exclusively in mathematical tasks (e.g., GSM8K, AQUA, SVAMP), where symbolic notation offers the clearest compression benefit. In DROP, which mixes symbolic and textual reasoning, some samples are routed to CC, reflecting hybrid reasoning patterns.

Expert Lexicons (EL) is most common in domain-specific datasets like PubMedQA and MedQA. Occasional routing to CC in these cases reflects the router's conservative fallback behavior, favoring general-purpose paradigms when confidence is low—a design choice that reduces the risk of applying technical conventions in inappropriate contexts.

D.3 Paradigm Assignment Examples

To illustrate how the SoT router assigns paradigms to diverse questions, we present four representative examples below—one from each SoT paradigm, along with an edge case that demonstrates the system's conservative fallback behavior. Each example includes the query as processed by the router and the assigned paradigm.

Chunked Symbolism (GSM8K)

Query:
Darrell and Allen's ages are in the ratio of 7:11. If their total age now is 162, calculate Allen's age 10 years from now.

Assigned Paradigm:
Chunked Symbolism

Conceptual Chaining (OpenbookQA)

Query:
Polar bears require
Choices:
A. a tropical environment
B. a frigid environment
C. a tepid environment
D. a warm environment

Assigned Paradigm:
Conceptual Chaining

Expert Lexicons (PubMedQA)

Query:
Is the holmium:YAG laser the best intracorporeal lithotripter for the ureter?
Choices: Yes, No, Maybe

Assigned Paradigm:
Expert Lexicons

Edge Case: Conceptual Chaining (PubMedQA)

Query:
[Context Here] Question: Birth characteristics and risk of low intellectual performance in early adulthood: are the associations confounded by socioeconomic factors in adolescence or familial effects?
Choices: Yes, No, Maybe

Assigned Paradigm:
Conceptual Chaining

E Output Examples

Figure 4 presents representative input–output examples for each of the three SoT paradigms alongside outputs from baseline prompting strategies including Chain-of-Thought (CoT), Chain-of-Draft (CoD), and Constrained CoT (CCoT). Compared to baselines, SoT responses are significantly more compact while maintaining logical structure and semantic completeness. While CoD and CCoT reduce length relative to CoT, they rely solely on shortened natural language, often resulting in compressed but less interpretable text.

Table 6: **Full results across models.**

Reasoning Type	Dataset	Method	Qwen-7B		Qwen-14B		Qwen-32B		LLaMA-3.1-8B		LLaMA-3.2-11B		GPT-4o		Claude-3.5	
			Acc	Tkn	Acc	Tkn	Acc	Tkn	Acc	Tkn	Acc	Tkn	Acc	Tkn	Acc	Tkn
Mathematical	GSM8K	CoT	86.22	211	93.55	215	94.89	263	82.89	236	77.78	229	94.67	255	98.00	245
		CCoT	88.67	92	88.89	135	83.78	86	71.11	92	70.22	83	93.56	102	90.00	105
		CoD	59.55	66	64.89	70	67.78	58	54.22	80	55.78	73	89.78	84	89.11	80
		SoT	83.33	80	92.89	87	95.78	103	69.78	83	69.11	78	94.67	78	90.22	98
	AQUA	CoT	65.78	228	78.67	267	76.89	289	63.78	323	60.66	324	80.67	362	83.56	308
		CCoT	68.89	111	76.22	155	74.00	97	51.11	126	48.67	111	75.78	121	80.44	113
		CoD	61.56	72	59.33	79	64.00	64	41.78	107	39.78	99	74.22	94	75.56	106
		SoT	69.78	108	77.56	116	82.22	138	51.33	119	51.56	114	77.33	106	82.67	117
	SVAMP	CoT	88.22	146	92.44	136	92.22	181	82.89	192	82.44	187	92.89	180	92.67	191
		CCoT	84.89	61	89.33	86	88.67	61	78.89	72	79.33	65	88.00	74	84.89	82
		CoD	80.89	43	85.78	51	85.33	44	69.11	54	73.56	50	90.44	50	87.56	60
		SoT	86.22	51	88.89	53	94.22	58	76.45	56	76.44	54	94.00	42	86.67	61
	DROP	CoT	69.40	138	67.33	141	72.67	155	60.67	190	61.33	187	73.56	164	74.22	189
		CCoT	69.55	59	70.89	86	75.56	60	59.78	61	60.00	59	77.56	74	76.44	87
		CoD	65.33	48	66.89	51	70.66	47	56.00	51	55.56	47	78.22	55	75.78	66
		SoT	68.89	54	71.55	58	75.55	55	61.11	54	60.89	53	78.67	52	76.67	63
Commonsense	CommonsenseQA	CoT	84.44	175	85.78	158	85.33	188	72.44	220	72.00	231	86.44	215	84.00	250
		CCoT	36.22	44	87.33	53	82.67	48	71.11	56	72.44	57	85.33	62	48.44	81
		CoD	77.55	38	82.89	40	83.11	38	70.67	44	71.56	42	85.78	49	83.33	61
		SoT	83.33	25	86.00	33	86.22	29	73.33	31	72.89	31	85.33	34	83.56	58
	OpenbookQA	CoT	87.11	170	95.11	154	95.33	186	86.22	226	88.67	230	97.78	209	98.22	251
		CCoT	68.00	44	93.78	54	93.56	49	84.00	58	85.33	59	96.67	62	74.67	81
		CoD	88.89	38	94.22	41	94.67	40	83.78	44	81.78	43	97.56	51	97.78	63
		SoT	86.89	29	93.55	38	95.11	32	84.44	36	85.56	36	97.33	38	97.56	60
	StrategyQA	CoT	92.22	130	93.33	139	93.78	158	87.11	180	87.56	186	94.00	176	91.56	225
		CCoT	88.22	46	90.44	67	90.22	51	87.55	59	84.67	60	89.78	65	94.22	80
		CoD	87.78	35	90.00	42	90.66	38	87.56	45	89.33	43	78.00	49	92.89	59
		SoT	81.11	37	93.11	42	94.67	40	86.45	41	86.00	39	94.89	46	92.22	58
Logical	LogiQA	CoT	53.78	288	56.22	265	60.67	306	44.44	312	42.89	315	63.33	330	61.11	322
		CCoT	53.78	68	60.22	104	63.11	63	42.00	80	42.67	75	55.56	76	53.33	95
		CoD	53.11	53	54.22	52	63.11	47	38.89	67	41.33	58	58.22	66	61.78	102
		SoT	50.67	77	56.00	75	60.22	79	40.00	88	35.11	85	59.56	90	62.67	122
	Reclor	CoT	72.67	270	77.78	231	81.78	289	58.00	273	57.78	279	86.22	292	89.33	305
		CCoT	73.55	59	79.78	75	82.45	57	60.00	66	61.11	63	86.89	63	76.00	88
		CoD	75.56	45	78.22	42	82.67	43	55.56	50	55.11	44	84.89	58	89.78	91
		SoT	68.89	46	78.89	52	81.78	53	56.22	55	55.56	54	86.89	70	87.33	101
Multi-Hop	HotPotQA	CoT	91.63	125	90.00	135	92.22	143	89.78	164	89.33	165	90.89	145	90.00	197
		CCoT	91.56	49	89.56	84	93.33	51	88.89	58	86.89	58	91.78	69	87.33	85
		CoD	90.22	41	89.78	44	91.11	39	85.55	47	87.78	45	92.89	52	89.11	62
		SoT	87.11	43	90.22	42	94.00	41	86.44	41	88.89	41	93.33	44	90.00	54
	MuSiQue	CoT	61.93	150	65.33	163	66.67	167	59.33	222	57.56	231	72.22	167	73.33	216
		CCoT	66.22	57	68.22	98	66.89	57	62.00	63	59.11	65	72.89	71	74.44	84
		CoD	62.00	45	71.11	47	68.89	43	60.89	52	60.22	47	72.22	55	74.89	64
		SoT	67.33	47	69.55	49	69.78	46	67.78	48	66.89	48	76.22	51	78.89	60
Scientific	QASC	CoT	78.00	182	83.78	163	87.33	222	75.34	284	76.89	287	86.89	258	88.89	264
		CCoT	29.11	44	82.89	54	79.11	50	75.34	58	74.89	57	81.56	65	53.33	83
		CoD	76.89	38	82.00	45	81.33	43	71.11	49	67.55	46	81.78	59	83.78	67
		SoT	72.22	26	82.89	36	84.22	30	75.78	33	69.56	34	85.33	36	84.89	57
	Worldtree	CoT	94.89	185	98.00	166	98.45	204	96.22	236	94.67	239	99.56	223	98.89	264
		CCoT	96.44	44	96.00	56	98.67	49	94.66	57	93.56	57	98.89	62	83.33	82
		CoD	97.11	39	96.89	40	98.67	40	90.89	44	91.33	42	99.56	51	98.89	66
		SoT	97.78	29	98.89	39	98.45	33	91.33	38	89.33	38	99.78	42	98.67	66
Specialized	PubMedQA	CoT	64.67	206	70.00	221	72.22	257	75.33	296	73.78	306	65.11	260	65.78	284
		CCoT	63.56	53	64.22	87	58.00	59	76.22	72	76.67	70	27.78	69	69.33	88
		CoD	66.00	39	64.67	46	59.11	42	76.89	50	76.89	47	12.00	57	70.22	69
		SoT	64.45	67	69.33	60	59.11	60	78.00	61	76.89	60	61.33	59	69.56	108
	MedQA	CoT	49.33	287	60.22	248	63.11	327	54.67	350	63.11	362	85.33	357	85.56	358
		CCoT	37.11	68	58.67	84	57.33	70	60.00	74	60.00	72	84.67	74	42.22	118
		CoD	45.78	57	53.33	58	58.67	52	55.56	60	53.11	54	80.67	69	82.22	141
		SoT	51.56	145	55.78	65	63.11	65	54.89	66	56.22	67	83.56	63	86.00	124

Table 7: **Paradigm Distribution by Dataset**. For each dataset, we show the counts of examples under each paradigm, as selected by the router model. Additionally, we report the dominant paradigm, the expected paradigm based on heuristic categorization, and whether the dominant paradigm aligns with the expected one. This data reflects the samples from the primary experiments in Section 3.

Reasoning Type	Dataset	Paradigm Label	Count	Dominant Paradigm	Expected Paradigm	Expected is Dominant?
Mathematical	GSM8K	Chunked Symbolism	150	Chunked Symbolism	Chunked Symbolism	✓
		Conceptual Chaining	0			
		Expert Lexicons	0			
	AQUA	Chunked Symbolism	150	Chunked Symbolism	Chunked Symbolism	✓
		Conceptual Chaining	0			
		Expert Lexicons	0			
	SVAMP	Chunked Symbolism	150	Chunked Symbolism	Chunked Symbolism	✓
		Conceptual Chaining	0			
		Expert Lexicons	0			
	DROP	Chunked Symbolism	76	Chunked Symbolism	Chunked Symbolism	✓
		Conceptual Chaining	74			
		Expert Lexicons	0			
Commonsense	CommonsenseQA	Chunked Symbolism	0	Conceptual Chaining	Conceptual Chaining	✓
		Conceptual Chaining	150			
		Expert Lexicons	0			
	OpenbookQA	Chunked Symbolism	1	Conceptual Chaining	Conceptual Chaining	✓
		Conceptual Chaining	149			
		Expert Lexicons	0			
	StrategyQA	Chunked Symbolism	2	Conceptual Chaining	Conceptual Chaining	✓
		Conceptual Chaining	148			
		Expert Lexicons	0			
Logical	LogiQA	Chunked Symbolism	15	Conceptual Chaining	Conceptual Chaining	✓
		Conceptual Chaining	134			
		Expert Lexicons	1			
	Reclor	Chunked Symbolism	0	Conceptual Chaining	Conceptual Chaining	✓
		Conceptual Chaining	150			
		Expert Lexicons	0			
Multi-Hop	HotPotQA	Chunked Symbolism	0	Conceptual Chaining	Conceptual Chaining	✓
		Conceptual Chaining	150			
		Expert Lexicons	0			
	MuSiQue	Chunked Symbolism	0	Conceptual Chaining	Conceptual Chaining	✓
		Conceptual Chaining	150			
		Expert Lexicons	0			
Scientific	QASC	Chunked Symbolism	0	Conceptual Chaining	Conceptual Chaining	✓
		Conceptual Chaining	148			
		Expert Lexicons	2			
	Worldtree	Chunked Symbolism	0	Conceptual Chaining	Conceptual Chaining	✓
		Conceptual Chaining	150			
		Expert Lexicons	0			
Specialized	PubMedQA	Chunked Symbolism	0	Expert Lexicons	Expert Lexicons	✓
		Conceptual Chaining	52			
		Expert Lexicons	98			
	MedQA	Chunked Symbolism	0	Expert Lexicons	Expert Lexicons	✓
		Conceptual Chaining	2			
		Expert Lexicons	148			

Sketch-of-Thought (SoT) – Conceptual Chaining
26 Output Tokens

Question: Green parts of a life form absorb
Choices: A. carbon dioxide, B. light, C. oxygen, D. water

Model Output:
#green_parts → #chloroplasts → absorb_light → photosynthesis
Answer: B

Sketch-of-Thought (SoT) – Chunked Symbolism
65 Output Tokens

Question: After transferring to a new school, Amy made 20 more friends than Lily. If Lily made 50 friends, how many friends do Lily and Amy have together?

Model Output:
Lily_friends = 50, Amy_friends = Lily_friends + 20 = 50 + 20 = 70, Total_friends = Lily_friends + Amy_friends = 50 + 70 = 120
Answer: 120

Sketch-of-Thought (SoT) – Expert Lexicons
67 Output Tokens

Question: [CONTEXT HERE] Biomechanical and wound healing characteristics of corneas after excimer laser keratorefractive surgery: is there a difference between advanced surface ablation and sub-Bowman's keratomileusis?
Choices: Yes, No, Maybe

Model Output:
PRK, ASA → Surface Ablation, SBK → Sub-Bowman's Keratomileusis, Histologic, Ultrastructural Eval → Diff Wound Healing, Cohesive Tensile Strength → PRK, ASA > SBK > LASIK
Answer: Yes

Constrained CoT (CCoT)
51 Output Tokens

Question: Green parts of a life form absorb
Choices: A. carbon dioxide, B. light, C. oxygen, D. water

Model Output:
Green parts, primarily leaves, contain chlorophyll which absorbs light for photosynthesis. While they also take in CO2 and water, the key function of green parts is light absorption to produce energy.
Answer: B

Chain-of-Draft (CoD)
29 Output Tokens

Question: Green parts of a life form absorb
Choices: A. carbon dioxide, B. light, C. oxygen, D. water

Model Output:
Green parts photosynthesis function. Absorb light, convert energy. Light essential for process.
Answer: B

Chain-of-Thought (CoT)
206 Output Tokens

Question: Green parts of a life form absorb
Choices: A. carbon dioxide, B. light, C. oxygen, D. water

Model Output:
1. The question is asking about what green parts of a life form absorb. 2. Green parts of a life form typically refer to leaves or chloroplasts in plants, which contain chlorophyll. 3. Chlorophyll is responsible for the green color and plays a crucial role in photosynthesis. 4. During photosynthesis, plants absorb certain substances to produce energy and grow. 5. Let's consider each option: - A. Carbon dioxide: Plants absorb CO2 during photosynthesis. - B. Light: Plants do absorb light, but this is more about capturing energy rather than absorbing it in the same way they take in gases. - C. Oxygen: Plants release oxygen during photosynthesis; they don't absorb it. - D. Water: While plants do absorb water, it's primarily through their roots, not the green parts. 6. Among the choices, the most accurate and specific to the green parts' function is carbon dioxide
Answer: B

Figure 4: **Representative reasoning traces for each prompting strategy.** SoT variants demonstrate structurally distinct and compact reasoning across paradigms (Conceptual Chaining, Chunked Symbolism, Expert Lexicons), while baseline methods (CCoT, CoD, CoT) rely on natural language with varying degrees of verbosity and clarity. For the same question, SoT achieves the shortest trace with clearer logic, while CoT is the most verbose.

Conceptual Chaining System Prompt

Role & Objective
You are a reasoning expert specializing in structured concept linking by connecting essential ideas in a logical sequence. Your goal is to extract key terms and present reasoning in clear, stepwise chains while minimizing unnecessary explanation.

This reasoning method follows a conceptual chaining approach, where information is linked in structured steps to establish relationships between ideas. This process integrates associative recall (direct lookups) and multi-hop reasoning (sequential dependencies) into a unified framework.

This method is most effective for:
- Commonsense reasoning (quickly linking familiar ideas)
- Multi-hop inference (tracing logical or causal dependencies)
- Fact-based recall (retrieving knowledge with minimal cognitive load)

—

How to Apply Conceptual Chaining
1. Extract Key Concepts → Identify the most relevant words or entities.
2. Use Minimal Words → Keep each reasoning step concise and direct.
3. Link Steps Sequentially → Maintain a clear and meaningful progression between concepts.
4. Avoid Full Sentences → Responses should use structured keyword connections.
5. Follow the Required Format → Present answers using stepwise chains for clarity.

—

Rules & Directives
1. Use Structured Concept Linking
- Each step must be logically connected.
- Use arrows (‘→‘) to show dependencies.

2. Avoid Unnecessary Text
- Do not restate the question.
- Do not use full sentences.

3. Maintain Logical Flow
- Concepts must be meaningfully ordered.
- Ensure each step contributes to the reasoning process.

4. Output Format
- Use the exact structured format:
‘ ‘ ‘
<think>
[shorthand reasoning]
</think>
boxed[Final answer]
‘ ‘ ‘
- The final answer must be boxed.
- If the question is multiple-choice, return the correct letter option inside the box.
- Use minimal words in your response.

Conceptual Chaining Exemplars

Q: What is the name of the currency used in Seoul?
A: <think> #Seoul → #South_Korea → Won </think> **Answer: Korean Won**

Q: Which planet has the highest surface temperature?
Choices: A) Mercury B) Venus C) Mars D) Jupiter
A: <think> #heat_trap Mercury → no atmosphere → loses heat Venus → thick CO2 → traps heat → hottest Mars → thin CO2 → cold Jupiter → no solid surface </think> **Answer: B**

Q: Which vitamin is essential for blood clotting?
A: <think> #blood_clotting → #vitamin_K </think> **Answer: Vitamin K**

Figure 5: **Conceptual Chaining system prompt and exemplars.**

Chunked Symbolism System Prompt

Role & Objective
You are a reasoning expert specializing in Chunked Symbolism, a cognitive reasoning technique that organizes numerical reasoning into structured steps. Your goal is to utilize chunked symbolism by representing information through equations, variables, and step-by-step arithmetic, while using minimal words.

Chunked Symbolism is inspired by the cognitive science principle of chunking—the idea that humans process information more efficiently when grouped into meaningful units. Instead of solving problems in a free-form manner, Chunked Symbolism breaks down complex operations into smaller, structured steps.

This method is particularly effective for:
- Mathematical problems (arithmetic, algebra, physics, engineering)
- Symbolic reasoning (logic-based computations, formula derivations)
- Technical calculations (financial modeling, physics simulations, unit conversions)

—

How to Apply Chunked Symbolism
Step-by-Step Guide
1. Identify Variables – Extract relevant numerical values and define variables.
2. Write Equations – Represent the solution using explicit mathematical formulas.
3. Perform Step-by-Step Computations – Solve in small, logical steps, keeping each line clear.
4. Label Units – Maintain consistent unit representation to prevent ambiguity.
5. Final Answer Formatting – Present the answer in the provided format for clarity.

—

Rules & Directives
1. Use Equations & Variables
- Define variables before computation.
- Always use explicit equations to represent reasoning.

2. Avoid Redundant Text
- Do not restate the problem; go directly to calculations.
- Use minimal context only if it aids understanding.

3. Apply Step-by-Step Arithmetic
- Break operations into small, structured steps.
- Ensure each line contains only one computation for clarity.

4. Output Format
- Use the exact structured format:
‘ ‘ ‘
<think>
[shorthand reasoning]
</think>
boxed[Final answer]
‘ ‘ ‘
- The final answer must be boxed.
- If the question is multiple-choice, return the correct letter option inside the box.
- Use minimal words in your response.

Chunked Symbolism Exemplars

Q: A car accelerates at 2.5 m/s^2 for 10 seconds. If its initial velocity was 15 m/s, what is its final velocity?
A: <think> a = 2.5 m/s^2 t = 10 s vi = 15 m/s vf = 15 + (2.5 × 10) vf = 40 m/s </think> **Answer: 40**

Q: If a product costs $120 and there is a 15% discount, what is the final price?
Choices: A) $10 B) $97 C) 102
A: <think> op = 120 d = 15% dp = 120 × (15 / 100) = 18 fp = 120 - 18 = 102 </think> **Answer: C**

Q: Question: A circuit has a voltage of 12V and a resistance of 4Ω. What is the current?
A: <think> V = 12V R = 4Ω I = 12 / 4 = 3A </think> **Answer: 3**

Figure 6: **Chunked Symbolism system prompt and exemplars.**

Expert Lexicons System Prompt

Role & Objective
You are a reasoning expert specializing in Expert Lexicons, a cognitive reasoning technique that leverages domain-specific shorthand, technical symbols, and jargon to ensure precise and efficient communication. Your goal is to compress reasoning into high-information expressions while maintaining technical accuracy and clarity.

Expert Lexicons is based on the principle that domain experts communicate using shorthand and structured notation. Instead of full explanations, this method condenses reasoning into compact, high-density expressions using technical symbols and field-specific abbreviations.

This method is particularly effective for:
- Technical disciplines (science, engineering, medicine, mathematics, and coding)
- Symbolic and formulaic reasoning (using field-specific notation and logical expressions)
- Maximizing efficiency (conveying information in the fewest possible tokens)

–

How to Apply Expert Lexicons
Step-by-Step Guide
1. Use Technical Symbols → Replace common terms with mathematical, logical, or scientific notation where applicable.
2. Leverage Abbreviations → Use domain-specific shorthand to condense reasoning.
3. Prioritize Information Density → Only include essential reasoning elements.
4. Follow Standardized Notation → Adhere to widely recognized conventions within each field.
5. Maintain Structural Precision → Ensure answers are formatted using compact, industry-specific expressions.

–

Rules & Directives
1. Use Domain-Specific Notation
- Mathematical & Logical Reasoning → ‘Σ,therefore,α, →‘
- Scientific Disciplines → ‘mol, J, Hz, pH, Vmax‘
- Medical & Engineering Fields → ‘CHF, OOP, PID, μm, dB‘

2. Eliminate Redundant Text
- No full sentences – responses must be in structured notation.
- No restating the question – directly express the solution.

3. Keep Responses Ultra-Compact
- Prioritize brevity while maintaining technical precision.
- Follow industry standards for notation and structured reasoning.

4. Output Format
- Use the exact structured format:
‘ ‘ ‘
<think>
[Shorthand reasoning using expert notation]
</think>
boxed[Final answer]
‘ ‘ ‘
- The final answer must be boxed.
- If the question is multiple-choice, return the correct letter option inside the box.
- Use minimal words in your response.

Expert Lexicons Exemplars

Q: Context: The discovery of the first interstellar object passing through the Solar System, 1I/2017 U1 (’Oumuamua), provoked intense and continuing interest from the scientific community and the general public.
Question: The interstellar object 1I/2017 U1 (’Oumuamua) exhibited unusual characteristics that led to various hypotheses about its origin. What does the designation "1I/2017 U1" signify?
Choices:
A) 1st Intergalactic object detected in 2017, classified under category U1
B) 1st Interstellar object cataloged, detected in 2017, following IAU naming conventions
C) 1st Independent Unclassified body observed beyond Neptune in 2017
A: <think> 1I → 1st interstellar object 2017 → Year detected U1 → Sequence ID IAU → Naming rules so 1st cataloged interstellar object (2017) </think>
Answer: B

Q: A patient with STEMI is given MONA therapy. They have a history of being allergic to aspirin. Are they at risk with this treatment?
A: <think> STEMI → ST-Elevation MI MONA → Morphine, O2, Nitrates, Aspirin. so Aspirin $\in$ MONA </think>
Answer: Yes

Q: What does EBITDA measure?
A: <think> EBITDA → Earnings Before Interest, Taxes, Depreciation, Amortization so Measures Core Profitability </think>
Answer: Core Profitability

Figure 7: **Expert Lexicons system prompt and exemplars.**

Classification System Prompt

You are an advanced language model tasked with classifying reasoning questions into one of three cognitive-inspired paradigms based on their linguistic structure and reasoning style.

Task:
Given a question, classify it into one of the following paradigms:
- conceptual_chaining → Used for multi-hop reasoning, structured fact-based recall, and sequential dependencies.
- chunked_symbolism → Used for mathematical, logical, or structured computational tasks requiring equations or stepwise arithmetic.
- expert_lexicons → Used for deciphering specialized terminology, jargon, or acronym-heavy questions from technical domains.

Paradigm Definitions:

1. Conceptual Chaining
- Purpose: Used when answering a question requires connecting multiple knowledge points in a structured sequence.
- Linguistic Indicators:
- Uses multi-hop inference (A → B → C).
- Involves causal, geographic, historical, hierarchical, biological, or functional relationships.
- Includes reasoning about scientific traits, tool functions, biological effects, and clinical implications.
- Focuses on structured recall and conceptual application, not just decoding or equation-solving.
- Includes trait inference, diagnostic logic, instrumental purpose, or category classification.
- Example Questions:
- "What currency is used in the capital of Japan's neighboring country?"
- "Who was the U.S. president during World War II?"
- "Which atmospheric layer protects Earth from harmful UV radiation?"
- "What happens to sea levels as polar ice caps melt due to climate change?"
- "How does smoking affect the respiratory system?"
- "What do anemometers measure?"
- "What kind of fats make butter solid at room temperature?"
- "What is a polygenic trait?"
- "How do Sarcocystis species make humans sick?"

–
2. Chunked Symbolism
- Purpose: Used for numerical, symbolic, and formulaic reasoning, where solutions involve stepwise calculations or structured logic.
- Linguistic Indicators:
- Contains mathematical expressions, units, numbers, or conversions.
- Requires symbolic operations or formulaic manipulation.
- Often involves stepwise arithmetic, algebra, logic puzzles, or physics computations.
- Example Questions:
- "If x + 3 = 10, what is x?"
- "A car accelerates from 10 m/s to 30 m/s over 5 seconds. What is the acceleration?"
- "What is the current if V = 20V and R = 10Ω?"
- "A mixture contains 30% acid. How many milliliters of water should be added to 200ml of this mixture to reduce the acid concentration to 20%?"
- "If a rectangle has a length of 8 cm and a width of 5 cm, what is its area?"
- "A recipe calls for 3/4 cup of sugar. If you want to make half the recipe, how much sugar do you need?"
- "Convert 120 kilometers per hour to meters per second."

–

3. Expert Lexicons
- Purpose: Used for deciphering domain-specific language, including jargon, acronyms, or specialized terminology in medicine, law, engineering, and finance.
- Linguistic Indicators:
- Focuses on decoding or interpreting field-specific abbreviations, acronyms, or terminology, especially when the question hinges on understanding a term's meaning rather than linking concepts or reasoning causally.
- Requires expertise in a specific domain rather than general knowledge or numerical calculations.
- Focuses on breaking down acronyms and technical concepts and emphasizing direct definitions rather than process understanding or causal relationships.
- Example Questions:
- "A patient with STEMI is given MONA therapy. What does this mean?"
- "In corporate law, what's the difference between a 10-K, 10-Q, and 8-K filing with the SEC?"
- "Which molecular structure represents benzene?"
- "When an architect specifies 'EIFS over CMU with VB and RTM,' what building materials are they referring to?"

–

Output Format:
You must ONLY return the single paradigm label as plain text with no explanation or additional formatting.

Options: conceptual_chaining, chunked_symbolism, expert_lexicons

Figure 8: **Paradigm classification prompt.**