How to Effectively Think? A Systematic Study on the Impact of Thinking Patterns in Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown remarkable gains under the *thinking then answering* paradigm, where models generate internal thought before producing a final answer. Despite the rapid adoption of this paradigm, it remains unclear how different thinking patterns interact with model scale and task type. We introduce **ThinkPatterns-21k**, a curated dataset of 21k general reasoning samples, each annotated with five distinct thinking patterns (*monologue*, *decomposition*, *self-ask*, *self-debate*, and *self-critic*), and use it as a controlled testbed for fine-tuning models at scales from 3B to 32B. Across three benchmark categories, we report three findings: **(1)** The effectiveness of thinking patterns is model-size–dependent: at smaller scales (3B–14B), structured patterns can match or exceed unstructured monologue on specific task categories; at 32B, monologue achieves the highest overall accuracy and leads or ties for the lead on every category. **(2)** Different thinking patterns exhibit task-specific affinities—most notably, the most heavily scaffolded pattern (decomposition) consistently underperforms on Information Processing & Expression tasks. **(3)** Less effective patterns are predominantly characterized by superficial reasoning that fails to explore the solution space. To facilitate reproducibility, we release our code, dataset, checkpoints, and training logs.

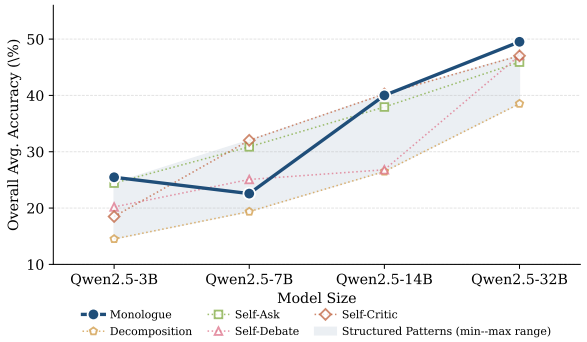

Figure 1: **Effectiveness of Thinking Patterns Across Model Sizes.** Overall average accuracy across three benchmark categories (excluding LLM-as-a-judge metrics, i.e., AlpacaEval 2 and Arena-Hard), averaged from Table 2. Our findings reveal that as model size increases, the advantage of unstructured *monologue* thinking becomes progressively more consistent and pronounced: at 32B, *monologue* achieves the highest overall accuracy, while smaller models may still benefit from structured patterns on certain categories.

1 Introduction

Large language models (LLMs) have traditionally been designed to generate answers immediately upon receiving human queries (Achiam et al., 2023; Team et al., 2024; Touvron et al., 2023). Recently, the emergence of advanced reasoning models such as OpenAI's o1 (Jaech et al., 2024), QwQ (Team, 2024), and Deepseek R1 (Guo et al., 2025) has demonstrated the effectiveness of a new paradigm: *thinking then answering*, with models following this paradigm known as large reasoning models (LRMs) (Xu et al., 2025a). This paradigm, where models engage in explicit thinking processes before producing a final answer, has shown remarkable success, achieving or even surpassing human-level performance across various domains, including mathematical problem-solving (Guan et al., 2025; Zhang et al., 2025), coding tasks (Zhang et al., 2024), and other complex challenges (Zhong et al., 2024; Qian et al., 2025; Chen et al., 2025).

Current LRM implementations predominantly rely on either reinforcement-learning–induced reasoning or distillation from existing reasoning models (Guo et al., 2025; Xie et al., 2025; Muennighoff et al., 2025; Chen et al., 2025; Liu et al., 2025), which has led to widespread adoption of an unstructured, free-form thinking pattern. Despite these promising results, the potential of structured thinking patterns remains underexplored, and we lack a comprehensive understanding of how different thinking patterns affect model performance across

task contexts and model sizes. This raises a fundamental question:

How can Language Models Effectively Think?

A simple yet intuitive conjecture is that effective thinking patterns for language models likely vary depending on model size and task category, with each thinking pattern exhibiting distinct strengths and limitations.

In this work, we examine the above conjecture by systematically investigating the impact of different thinking patterns on performance w.r.t. various model capabilities (sizes) and different tasks. We carefully design five distinct thinking patterns ranging from unstructured pattern (*monologue*) to progressively more structured variants (*self-ask*, *self-debate*, *self-critic*, and *decomposition*).

Then, we introduce ThinkPatterns-21k, a large-scale dataset containing 21K query-answer pairs, where each pair is augmented with all five thinking patterns. Crucially, we preserve identical queries and final answers across all thinking patterns, changing only the internal thinking process. This controlled design allows us to isolate the impact of thinking patterns while eliminating confounding variables related to query-answer differences.

Equipped with this dataset, we conduct comprehensive experiments across **3~32B** model sizes and **three benchmark categories**: Logical & Mathematical Reasoning, Information Processing & Expression, and Instruction Following & Preference Alignment. Our study reveals two findings.

First, the effectiveness of thinking patterns is strongly influenced by model size. As Figure 1 shows, the advantage of unstructured *monologue* grows progressively more consistent and pronounced with scale. Smaller models ($\leq$ 14B) may still benefit from structured patterns on certain task categories, with structured patterns occasionally winning in specific domains. At 32B, *monologue* achieves the highest overall accuracy and leads on Information Processing & Expression and Instruction Following & Preference, while on Logical & Mathematical Reasoning it is essentially tied with *Self-Debate* (within 0.13 points). The most heavily scaffolded pattern, *Decomposition*, underperforms all alternatives at 32B—suggesting that larger models possess sufficient inherent capability to navigate complex problems without rigid predefined structure.

Second, different thinking patterns exhibit distinct task-domain affinities. *Decomposition* consistently underperforms on the Information Processing & Expression category across nearly all model sizes, indicating that this highly structured pattern is poorly suited for such tasks despite its theoretical appeal.

Our interpretability experiments further reveal the distinctions between model sizes in their ability to leverage the *Monologue* thinking pattern effectively. By forcing models to generate answers during early stages of their internal thinking process, we find that larger models (32B) demonstrate steady accuracy improvements throughout the *Monologue* thinking process, suggesting substantive reasoning that genuinely refines understanding. In contrast, smaller models (3B, 7B, 14B) exhibit complex fluctuations, often failing to show consistent improvement as thinking progresses. These findings suggest that smaller models may engage in more superficial thinking that inadequately explores the solution space, while larger models can effectively build upon their reasoning in a coherent manner even with unstructured thinking patterns.

To summarize, our contributions are threefold: **(1)** We introduce ThinkPatterns-21k, a curated dataset designed to systematically investigate the impact of different thinking patterns, with controlled conditions that isolate the effect of thinking patterns; **(2)** We provide empirical evidence validating our conjecture that effective thinking patterns for language models vary depending on model size and task category, with each pattern exhibiting distinct strengths and limitations in different contexts; **(3)** We offer insights into why models utilize *Monologue* thinking differently, revealing that larger models conduct substantive reasoning that genuinely refines solutions, while smaller models engage in superficial thinking that fails to effectively explore the solution space.

2 Related Work

Prompt Engineering for Reasoning. Following Chain-of-Thought (CoT) (Wei et al., 2022), a large body of work has proposed structured reasoning prompts: Self-Ask (Press et al., 2023) for interactive sub-question decomposition, multi-agent debate (Liang et al., 2024) for dialectical reasoning, Plan-and-Solve and Decomposed Prompting (Wang et al., 2023; Khot et al., 2023) for systematic divide-and-conquer, and self-refinement

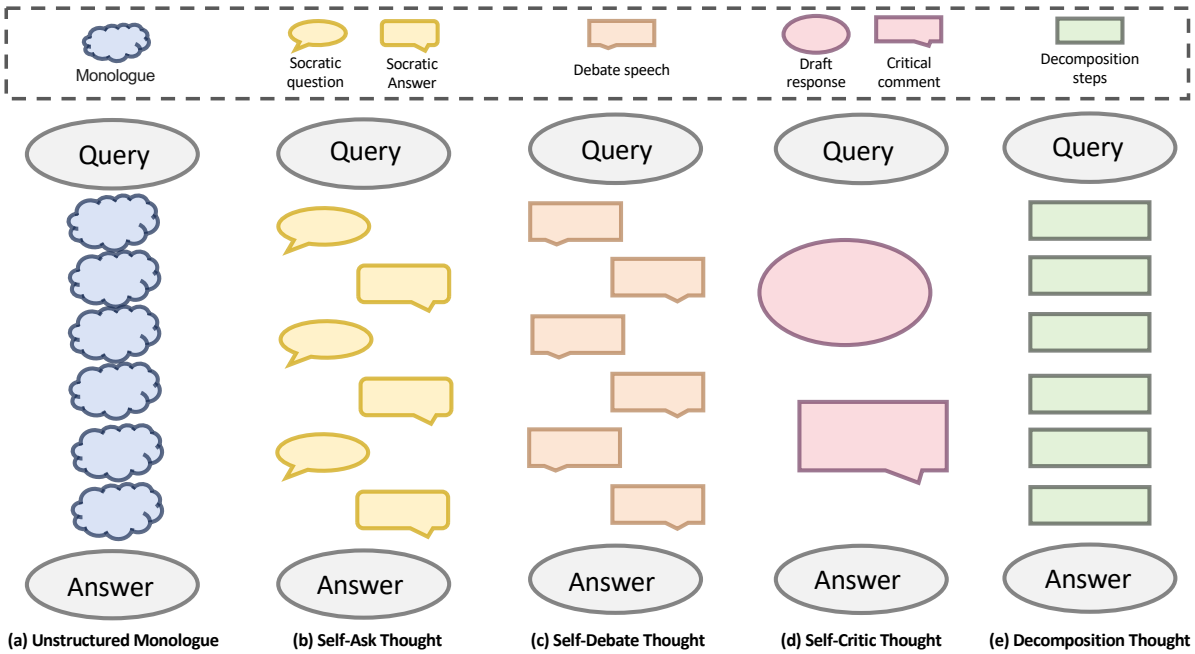

Figure 2: **Reasoning Dynamics of Different Thinking Patterns.** This figure illustrates the contrast among our five thinking patterns: **(a) Unstructured *Monologue***, mimicking natural human internal monologue; **(b) *Self-Ask* Thought**, implementing Socratic questioning for deeper exploration; **(c) *Self-Debate* Thought**, facilitating internal debate dialogue to reach optimal solutions and **(d) *Self-Critic* Thought**, incorporating self-evaluation and refinement mechanisms; **(e) *Decomposition* Thought**, which systematically breaks down complex tasks into manageable sub-problems. Each thinking pattern demonstrates a unique reasoning pathway and problem-solving strategy.

loops (Madaan et al., 2024) for iterative critique. These approaches act at inference time and do not internalize the thinking patterns into model parameters; consequently, their relative effectiveness across model scales has not been systematically studied under controlled conditions.

Model Training for Enhanced Reasoning. A complementary line internalizes reasoning via training on synthesized thinking trajectories, including Monte-Carlo Tree Search rollouts (Qin et al., 2024; Guan et al., 2025), GPT-4-distilled demonstrations (Chen et al., 2024), and direct distillation from frontier reasoning models (Min et al., 2024; Guo et al., 2025; Team, 2025; Thoughts, 2025; Muennighoff et al., 2025; Ye et al., 2025; Xu et al., 2025b). Most of this work converges on a single, free-form "monologue" style of thinking inherited from the teacher model. Whether *structured* thinking patterns—and which specific structures—should be internalized at which model scales remains largely unexplored. Our work is a controlled study of exactly this question: we hold the queries and gold answers fixed and vary only the structure of the internal thinking the model is trained to imitate.

Test-Time Scaling for LLMs. Test-time compute scaling sits between prompt engineering and training-time approaches. One branch produces a single, extended CoT with in-context reflection and backtracking, as in o1 (Jaech et al., 2024), QwQ (Team, 2024), and DeepSeek-R1 (Guo et al., 2025; Wu et al., 2024). A second branch draws multiple candidates and selects the best one via Best-of-N or search/reward-model guided procedures (Snell et al., 2024; Brown et al., 2024). Our work belongs to the single-extended-CoT branch but studies its *internal structure*—which thinking pattern the model is trained to follow—rather than its length or selection procedure.

3 Thinking Patterns

We identify and study five distinct thinking patterns spanning a continuum from unstructured to maximally structured. The patterns are summarized below; the formal constraint-set definitions, joint conditional generation, and the resulting cardinality chain over admissible reasoning sequences are deferred to Appendix B.

Unstructured *Monologue* is a free-form thinking pattern without explicit structural constraints, used

by most current reasoning models (Jaech et al., 2024; Guo et al., 2025). The model generates reasoning steps freely with no predetermined format.

***Self-Ask* Thought,** inspired by Socratic questioning (Press et al., 2023), alternates Socratic question and answer steps to deepen problem exploration while maintaining a consistent structure.

***Self-Debate* Thought** alternates between supporting and challenging viewpoints in an internal debate (Liang et al., 2023), structuring reasoning around the confrontation of opposing positions.

***Self-Critic* Thought** (Madaan et al., 2024) uses a compact two-step process: generating a draft answer followed by critical self-review and refinement.

***Decomposition* Thought** employs a systematic divide-and-conquer strategy (Wang et al., 2023; Khot et al., 2023) via six predefined steps: problem **u**nderstanding, component **i**dentification, subproblem **a**nalysis, connection **m**apping, integration **s**trategy, and solution **v**alidation. We adopt this fixed six-step schema deliberately as the maximally scaffolded anchor of our pattern continuum: it is intended to be reproducible, internally consistent across queries, and structurally distinct from the other patterns, rather than to claim cognitive fidelity with how humans naturally decompose. Findings on this pattern should accordingly be read as evidence about *this particular scaffold*, not about decomposition in general (see Limitations).

By construction, the admissible reasoning-sequence space shrinks monotonically as patterns become more structured, with *monologue* the largest and *decomposition* the smallest. We use this continuum throughout the rest of the paper to compare unstructured and progressively more structured thinking.

4 Study Setup

Our investigation aims to conduct a systematic comparison of diverse thinking patterns and explore the fundamental question of *how can language models effectively think*. Building upon the five thinking patterns described in Section 3, we design specialized prompts to guide GPT-4o in simulating them, thereby constructing our ThinkPatterns-21k dataset (Section 4.1). In Section 4.2, we present both qualitative and quantitative analyses characterizing the distinctions between different thinking patterns in our dataset.

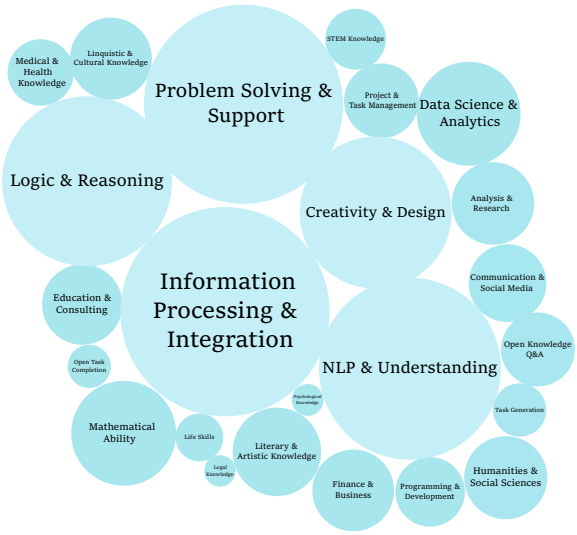

Figure 3: **Task category distribution of queries in our dataset.** ThinkPatterns-21k contains 21,000 high-quality and diverse query-answer pairs across various task categories, each annotated with five distinct thinking patterns.

4.1 Dataset Construction Process

Query-Answer Pairs Collection. To construct our dataset with explicit internal thought processes, we first sampled query-answer pairs from existing instruction following datasets. We selected the Infinite Instruct dataset[1] as our seed dataset due to its comprehensive coverage of diverse instruction-following scenarios and varied sources. After filtering to retain only English data, we randomly sampled 21k query-answer pairs while maintaining a balanced distribution across task categories (as shown in Figure 3).

Thinking Patterns Construction. To generate internal thinking patterns for each query-answer pair, we employed GPT-4o as our primary annotation model. We designed specific prompts (provided in Appendix A) to guide GPT-4o in generating internal thoughts that align with our five designed thinking patterns. The prompts were carefully crafted to ensure the generated thoughts reflect natural cognitive progression, maintain coherence between the thinking process and final response.

For quality validation, three annotators performed spot checks on a sample of the generated internal thoughts. The review examined the logical coherence of the internal thought and its relevance to the original queries and answers.

[1] https://huggingface.co/datasets/BAAI/Infinity-Instruct

Thinking Patterns	Quantitative Metrics			Qualitative Features						
	$\lceil Tokens \rceil$	$\overline{Steps}$	$\#Instances$	**Fixed Steps**	**Iterative**	**Free-form**	`Progressive`	`Multi-perspective`	`Refinement`	`Systematic`
Monologue	1300.50	-	21K	✗	✓	✓	○	○	○	○
Self-Ask	1393.09	36.51	21K	✗	✓	✗	✓	✗	✗	✗
Self-Debate	1486.97	24.38	21K	✗	✓	✗	✗	✓	✗	✗
Self-Critic	1445.12	2	21K	✓	✗	✗	✗	✗	✓	✗
Decomposition	1218.28	6	21K	✓	✗	✗	✓	✓	✓	✓

Table 1: **Comparison of Thinking Patterns.** Quantitative metrics and qualitative features across all five patterns. A checkmark (✓) indicates the pattern consistently exhibits the feature; a cross (✗) indicates it does not; a circle (○) indicates the feature may appear but not with consistency or predictability. Patterns range from fully unconstrained (*Monologue*) to the most structured (*Decomposition*, with four qualitative features simultaneously enforced).

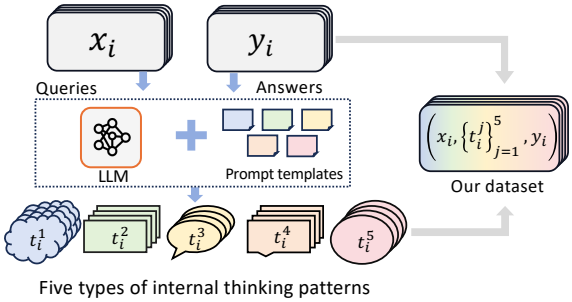

Figure 4: **Overview of our dataset construction pipeline.** The process begins with gathering query-answer pairs. These are subsequently processed by a LLM with specific prompts to produce five internal thinking patterns: *monologue*, *decomposition*, *self-ask*, *self-debate*, and *self-critic*, which are then merged to create our final dataset ThinkPatterns-21k.

4.2 Statistical Analysis of Thinking Patterns

For deeper insights into characteristics of different thinking patterns, we conducted both quantitative and qualitative analyses, as shown in Table 1.

For the quantitative aspect, we used the Qwen 2.5 tokenizer to calculate token counts for each thinking pattern and found that the average token lengths are relatively similar, ranging from 1200 to 1450 tokens. Despite the similar token lengths, the thinking patterns exhibit significant differences in reasoning steps: *self-ask* and *self-debate*, being iterative in nature, have varying numbers of steps depending on the depth and breadth of the thinking process, with each question-answer pair and each pro-con debate counted as one step respectively; whereas *self-critic* and *decomposition* adopt predefined structures with fixed steps of 2 and 6.

In terms of qualitative features, each of our five thinking patterns has distinctive characteristics: *self-ask* focuses on progressive thinking, exploring problems in depth through continuous questioning; *self-debate* emphasizes multi-perspective analysis, developing thoughts through the confrontation of opposing viewpoints; *self-critic* highlights refinement behavior by critically evaluating the initial response to make improvements. *Decomposition*, as a systematic thinking pattern, combines the above features of progressive thinking, multi-perspective analysis, and refinement, systematically breaking down and solving problems through predefined steps. In contrast, *monologue*, as a free-form pattern, may also exhibit any of the above features, but since its thinking process has no explicit structural constraints, these features do not appear with consistency or predictability.

5 Experiment on ThinkPatterns-21k

5.1 Experiment Setup

To rigorously evaluate the effectiveness of different thinking patterns, we conducted comprehensive experiments across multiple model sizes and diverse task categories. This section details our experimental setup, evaluation methodology, and benchmarking approach.

Models and Evaluation Datasets. We evaluated four models in the Qwen2.5 series with parameters ranging from 3B to 32B (specifically 3B, 7B, 14B, and 32B). Each model was fine-tuned with different thinking patterns.

For comprehensive evaluation, we employed benchmarks across three distinct categories. For Logical & Mathematical Reasoning, we utilized reasoning and math subsets from LiveBench (White et al., 2025), and MATH (Lightman et al., 2023). To assess Information Processing & Expression, we employed language, coding, data analysis subsets from LiveBench (White et al., 2025). For Instruction Following & Preference Alignment, we used LiveBench's instruction following subset (White et al., 2025), IFeval (Zhou et al., 2023), AlpacaEval 2 (Li et al., 2023), and Arena-Hard (Li et al., 2024) to measure both instruction following capabilities and alignment

Categories →	Logical & Mathematical				Information Processing & Expression				Instruction Following & Preference				
Benchmarks → **Patterns ↓ Metrics →**	**Live$_{\text{reason}}$** acc	**Live$_{\text{math}}$** acc	**MATH** acc	**Avg**	**Live$_{\text{lang}}$** acc	**Live$_{\text{code}}$** acc	**Live$_{\text{data}}$** acc	**Avg**	**IFeval** acc	**Live$_{\text{IF}}$** acc	**A-Eval2** lcwr	**A-Hard** wr	**Avg**
Qwen2.5-3B													
Monologue	14.20	22.60	51.40	29.40	5.80	23.80	24.50	**18.03**	36.97	24.50	24.70	32.50	**27.23**
Decomposition	13.20	17.20	15.00	15.13	5.50	11.70	8.40	8.53	24.95	20.10	18.66	21.90	20.22
Self-Ask	13.80	20.80	49.20	27.93	3.80	22.50	23.90	16.73	36.78	24.70	21.65	26.40	24.25
Self-Debate	13.80	20.70	37.40	23.97	3.30	22.10	14.90	13.43	29.02	20.40	21.97	25.20	22.52
Self-Critic	12.50	18.80	62.80	**31.37**	4.90	8.20	9.30	7.47	26.62	4.90	21.88	30.30	19.03
Qwen2.5-7B													
Monologue	13.10	23.90	52.00	29.67	5.40	17.90	27.20	16.83	35.67	5.40	25.28	31.20	20.63
Decomposition	20.30	25.20	48.20	31.23	8.50	5.60	11.30	8.47	27.36	8.50	32.26	46.70	29.15
Self-Ask	18.80	32.60	58.60	**36.67**	11.90	31.60	34.60	26.03	46.95	11.90	33.85	46.60	**30.78**
Self-Debate	19.60	22.60	42.80	28.33	6.60	42.50	28.70	25.93	31.24	6.60	34.59	42.40	27.86
Self-Critic	18.20	32.10	59.20	36.50	11.40	39.20	33.20	**27.93**	51.57	11.40	36.42	39.20	29.01
Qwen2.5-14B													
Monologue	34.60	40.40	63.80	46.27	18.10	44.70	40.80	**34.53**	59.52	18.10	44.68	66.20	**42.99**
Decomposition	29.10	40.30	39.60	36.33	16.10	28.70	15.20	20.00	26.80	16.10	41.62	58.40	38.71
Self-Ask	36.40	43.50	62.40	47.43	16.20	42.40	39.50	32.70	46.95	16.20	40.88	62.80	39.96
Self-Debate	28.00	33.20	44.60	35.27	10.20	43.40	13.60	22.40	31.24	10.20	43.98	50.70	34.96
Self-Critic	36.70	45.70	63.00	**48.47**	15.20	44.40	39.90	33.17	61.92	15.20	43.91	62.00	40.37
Qwen2.5-32B													
Monologue	40.80	51.30	66.00	52.70	32.90	53.20	53.30	**46.47**	65.80	32.90	47.87	71.60	**50.79**
Decomposition	48.70	43.70	46.40	46.27	29.40	45.60	18.40	31.13	45.47	30.60	35.59	50.80	39.00
Self-Ask	39.80	49.30	65.20	51.43	24.00	48.40	49.70	40.70	66.91	24.00	47.00	68.10	46.37
Self-Debate	43.90	48.00	66.60	**52.83**	27.20	53.30	48.00	42.83	62.48	27.20	45.90	70.60	47.90
Self-Critic	42.20	50.70	64.60	52.50	30.10	50.00	44.50	41.53	64.14	30.10	47.35	71.40	49.62

Table 2: **Performances Across Three Categories of Benchmarks.** Results illustrate the model-size-dependent efficacy of thinking patterns: as model scale increases, the advantage of *monologue* becomes more consistent and pronounced—at 32B it achieves the highest overall average (Figure 1) and leads on Information Processing & Expression and Instruction Following & Preference, while essentially tying with *Self-Debate* on Logical & Mathematical Reasoning (52.70 vs. 52.83). Smaller models may still benefit from structured patterns on specific task categories. **Bold** indicates the best category-average score among all patterns for a given model size; underline indicates the worst. For Instruction Following & Preference, the reported *Avg* aggregates *Live$_{IF}$*, *A-Eval2* (LC-WR), and *A-Hard* (WR); *IFeval* is reported separately because its strict-match accuracy is on a different scale from the preference-based metrics.

with human preferences.

Implementation Details. All models were fine-tuned for 3 epochs with our ThinkPatterns-21k dataset, using the Adam optimizer with a learning rate of 5×10^{-6} and a cosine learning rate schedule. Each training sample adhered to the default chat template, with models required to generate both a thinking process (enclosed in `<think>` and `</think>` tokens) and a final response. This structured format enabled the separation of thinking processes from final responses for evaluation. We employed greedy decoding across all benchmarks to ensure reproducibility and fair comparison.

Evaluation Metrics. We employed task-appropriate metrics across our experiments. For most benchmarks, we calculated accuracy (acc) as the primary metric. For IFeval, we report strict-prompt accuracy. For AlpacaEval 2 and Arena-Hard, we adhered to their established protocols, utilizing LLM-as-a-judge to determine length control win rate (lcwr) and win rate (wr), respectively.

5.2 Findings

Our results reveal several key insights into the effectiveness of thinking patterns for language models across different sizes and task categories.

Effectiveness of Thinking Patterns is Model Size-Dependent. The experimental results reveal a clear gradient between thinking pattern effectiveness and model size. As demonstrated in Figure 1 and Table 2, the advantage of unstructured *monologue* thinking grows progressively more consistent and pronounced as model scale increases. For smaller models (below 32B), structured patterns (including *self-ask*, *self-debate*, *self-critic*, and *decomposition*) can achieve superior performance on certain task categories, reflecting that predefined structural scaffolding can complement the limited inherent reasoning capacity of smaller models.

However, as models reach 32B parameters, this

relationship shifts. The 32B model achieves the highest overall accuracy with *monologue* and leads on Information Processing & Expression and Instruction Following & Preference, while on Logical & Mathematical Reasoning *monologue* (52.70) is essentially tied with *Self-Debate* (52.83). In contrast, the most heavily scaffolded pattern, *Decomposition*, consistently underperforms structured alternatives at 32B and is the worst on Logical & Mathematical, Information Processing, and Instruction Following & Preference. This pattern suggests that larger models possess sufficient inherent capabilities to navigate complex problems without relying on heavily predefined structured frameworks, and that overly rigid structure may constrain their flexibility.

Task-Specific Strengths and Limitations of Thinking Patterns. Our experiments further reveal that different thinking patterns exhibit distinct strengths and limitations across specific task categories. For instance, in the Information Processing and Expression category, the *decomposition* pattern performs particularly poorly, receiving the lowest scores at 7B, 14B, and 32B (and the second-lowest at 3B, narrowly trailing *self-critic*). The rigidity of this highly structured thinking pattern appears to hinder performance on these specific categories of tasks.

Conversely, less structured patterns such as *monologue* and *self-ask* demonstrate robust performance across a diverse range of tasks, suggesting greater adaptability to task-specific demands. This indicates that the optimal thinking pattern depends not only on model size but also on the specific characteristics of the task at hand. More flexible thinking patterns appear to maintain effectiveness across a broader range of scenarios, while highly structured patterns may excel in some domains but significantly underperform in others where their predefined structure fails to align with task features.

Failure Mode Taxonomy. To explain *why* certain patterns underperform, we identify three representative failure modes from case studies (Appendix B.3): **(1) Unfaithful Thinking**—thinking arrives at one conclusion but the response contradicts it; this occurs most frequently in *Self-Debate* (14.3%, Table 9); **(2) Superficial Thinking**—smaller models using *Monologue* verbosely circle the problem without genuine progress; and **(3) Pattern–Task Mismatch**—a structurally rigid pattern (notably *Decomposition*) is applied to tasks inherently unsuited to it, spending more tokens decomposing than solving.

Sizes→ Patterns↓	Qwen2.5-3B	Qwen2.5-7B	Qwen2.5-14B	Qwen2.5-32B
Monologue	17.22	22.39	**37.61**	**47.82**
Best of S	**21.21**	**23.07**	33.02	45.06

Table 3: **Impact of Annotation Source.** Results from training Qwen2.5 models using parallel datasets annotated by Claude-3.5-Sonnet and Llama3.3 70B Instruct. *Best of S* represents the best performance among all structured thinking patterns. Consistent trends across all three annotation sources confirm that findings are not artifacts of GPT-4o's specific reasoning style.

5.3 Interpretability Experiments

To further demonstrate how the effectiveness of thinking patterns depends on model size, we designed interpretability experiments examining how the *Monologue* pattern becomes increasingly effective as model size grows. These experiments track accuracy trends as reasoning progresses by measuring performance at different stages of the thinking process.

We implemented a uniform truncation strategy based on token length for the *Monologue* pattern. By truncating the thinking process at proportional intervals (0/6, 2/6, 4/6, etc. of thinking process) and inserting the `</think>` token to request an answer, we could evaluate accuracy at various reasoning stages, focusing specifically on coding tasks. As Figure 5 illustrates, our results reveal a stark contrast in reasoning dynamics across model sizes. While the 32B model shows a consistent upward trajectory in solution accuracy throughout its thinking process, smaller models (3B, 7B, 14B) demonstrate erratic performance patterns with unpredictable accuracy fluctuations. This distinction indicates that larger models develop increasingly refined understanding as their reasoning unfolds, engaging in genuine cognitive development rather than simply producing verbose text. In contrast, smaller models' inconsistent performance suggests their thinking processes remain superficial, failing to effectively explore the solution space and advance beyond initial shallow reasoning.

5.4 Ablation Study

Impact of Data Source Annotation Model. All five thinking-pattern variants in ThinkPatterns-21k are generated by GPT-4o, raising a legitimate concern that the trends we observe might

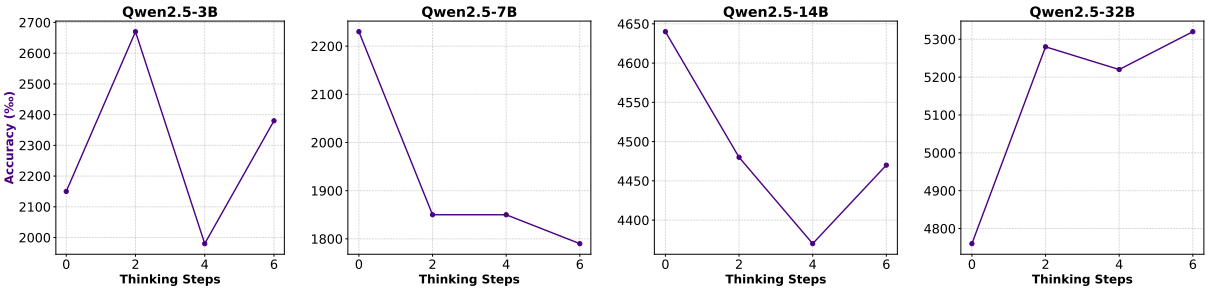

Figure 5: **Accuracy Evolution Throughout *Monologue* Thinking Process.** By forcing models to generate answers at intermediate stages (truncating at 0/6, 2/6, 4/6, and 6/6 of the thinking process), we observe the 32B model demonstrates steady accuracy improvements as thinking progresses, while smaller models (3B, 7B, 14B) exhibit complex fluctuations. The consistent improvement in larger models suggests substantive reasoning that genuinely refines understanding, rather than superficial thinking. Accuracy measured on the code subset of LiveBench; y-axis values are scaled by 10^4 relative to percentage (i.e., a value of N on the y-axis corresponds to $N \times 10^{-4} \times 100\%$ accuracy).

reflect GPT-4o's particular reasoning style rather than properties of the patterns themselves. To test this, we created parallel datasets (5K examples per annotation model due to resource constraints) using two structurally distinct LLMs—Claude-3.5-Sonnet and Llama-3.3-70B-Instruct—and trained Qwen2.5 models on each. Table 3 reports the resulting *Monologue* vs. *Best of Structured* comparison across four model sizes and three annotation sources. The qualitative ordering is preserved in all three settings: at $\geq$ 14B, *Monologue* leads; at $\leq$ 7B, the best structured pattern is competitive or wins. This robustness across annotators provides direct evidence that the model-size-dependent gradient we report is a property of the patterns themselves rather than a stylistic artifact of GPT-4o.

Impact of Model Families. To test architecture-level generalization, we fine-tuned three additional models on the ThinkPatterns-21k training data: Llama-3.1-8B and Gemma2-9B as small-to-mid scale, and Gemma2-27B as a larger-capacity model. Table 4 reports *Monologue* vs. *Best of Structured* on the same evaluation suite as Table 2. The size-dependent gradient is reproduced across both families: structured patterns win on Llama3.1-8B (16.17 vs. 15.78) and Gemma2-9B (13.70 vs. 11.61), while *Monologue* wins on Gemma2-27B (45.76 vs. 43.24)—the same crossover behavior we observe within the Qwen2.5 family. We report a single *Best of S* number per family rather than per-pattern detail to keep this ablation comparable across the three families given the compute budget; the per-pattern numbers track the Qwen2.5 ordering qualitatively.

Sizes→ **Patterns↓**	Llama3.1-8B	Gemma2-9b	Gemma2-27b
Monologue	15.78	11.61	**45.76**
Best of S	**16.17**	**13.70**	43.24

Table 4: **Impact of Model Families.** Results from testing thinking patterns on Llama3.1 and Gemma2 families. *Best of S* represents the best performance among all structured thinking patterns. The consistent pattern across architectures confirms generalizability beyond Qwen2.5.

6 Conclusion

We systematically studied *how language models effectively think* across five thinking patterns, four model sizes, and three benchmark categories. Findings reveal a model-size–dependent gradient: as capacity grows, unstructured *monologue* becomes increasingly consistent and pronounced, with the 32B model attaining the highest overall accuracy and leading or near-tying on every category, while smaller models still benefit from structured patterns on specific tasks. Interpretability analysis shows a fundamental contrast: larger models progressively refine solutions, smaller models engage in superficial exploration. These results advance our understanding of language-model reasoning and provide practical guidance on selecting thinking patterns by model size and task.

Limitations

Our primary experiments focus on the Qwen2.5 family (3B–32B), the only open-weight family with a consistent architecture across four sizes spanning an order of magnitude; we extend to Llama3.1-8B and Gemma2-9B/27B in Table 4, but a full per-size sweep on multiple families is beyond our compute budget. Results are reported from a single training seed per configuration, and our evaluation is based on final-answer accuracy and length-controlled win rates rather than direct scoring of intermediate reasoning quality; the truncation analysis (§5.3) and failure-mode taxonomy (§5.2) provide complementary signal. The five thinking patterns are annotated by GPT-4o; our annotation-source ablation with Claude-3.5-Sonnet and Llama-3.3-70B (Table 3) preserves the qualitative ranking but cannot fully rule out residual stylistic bias. The *decomposition* pattern uses a fixed six-step scaffold as the maximally structured anchor of our continuum; findings on it should be read as evidence about *this particular scaffold*, not about all decomposition-style reasoning. Our dataset is English-only.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. 2025. Judgelrm: Large reasoning models as a judge. *arXiv preprint arXiv:2504.00050*.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

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

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Zhaowei Liu, Xin Guo, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Zixuan Wang, Jiajie Xu, Weige Cai, Ziwei Yang, Xueqian Zhao, and 1 others. 2025. Fin-r1: A large language model for financial reasoning through reinforcement learning. *arXiv preprint arXiv:2503.16252*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, and 1 others. 2024. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Beijing Academy of Artificial Intelligence (BAAI). 2024. Infinity instruct. https://huggingface.co/datasets/BAAI/Infinity-Instruct. Accessed: 2025-05.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Han Yi, Jimin Huang, Qianqian Xie, and Jianyun Nie. 2025. Fino1: On the transferability of reasoning enhanced llms to finance. *arXiv preprint arXiv:2502.08127*.

Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and 1 others. 2024. O1 replication journey: A strategic progress report–part 1. *arXiv preprint arXiv:2410.18982*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

NovaSky Team. 2025. Sky-T1: Train your own O1 preview model within $450. https://novasky-ai.github.io/posts/sky-t1. Accessed: 2025-01-09.

Qwen Team. 2024. Qwq: Reflect deeply on the boundaries of the unknown.

Open Thoughts. 2025. Openthoughts-114k. https://github.com/open-thoughts/open-thoughts.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*.

Siwei Wu, Zhongyuan Peng, Xinrun Du, Tuney Zheng, Minghao Liu, Jialong Wu, Jiachen Ma, Yizhi Li, Jian Yang, Wangchunshu Zhou, and 1 others. 2024. A comparative study on reasoning patterns of openai's o1 model. *arXiv preprint arXiv:2410.13639*.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025a. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, and 1 others. 2025b. Redstar: Does scaling long-cot data unlock better slow-reasoning systems? *arXiv preprint arXiv:2501.11284*.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. 2024. o1-coder: an o1 replication for coding. *arXiv preprint arXiv:2412.00154*.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand.

Tianyang Zhong, Zhengliang Liu, Yi Pan, Yutong Zhang, Yifan Zhou, Shizhe Liang, Zihao Wu, Yanjun Lyu, Peng Shu, Xiaowei Yu, and 1 others. 2024. Evaluation of openai o1: Opportunities and challenges of agi. *arXiv preprint arXiv:2409.18486*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Appendix

This appendix provides supplementary materials organized into the following sections. Section B.1 details the construction and statistics of the ThinkPatterns-21k dataset. Section B.2 describes training configurations and full evaluation protocol details. Section B.3 presents extended case studies illustrating effective and ineffective thinking behaviors. Section B.4 presents a supplementary mixed-pattern training experiment. Section B.5 provides linguistic analysis of each thinking pattern. Section B.6 lists the annotation prompt templates used for data generation. Section B.7 shows concrete examples from ThinkPatterns-21k for each thinking pattern.

B Formal Definitions of Thinking Patterns

This appendix provides the formal constraint-set definitions for the five thinking patterns described in Section 3.

Let $\mathcal{Z}$ denote the space of all possible reasoning behaviors, and let $D = \{d_1, d_2, \ldots, d_T\} \in \mathcal{Z}^T$ be a sequence of reasoning behaviors over T steps in a complete thinking process. For each thinking pattern we define a constraint set $\mathcal{C} \subseteq \mathcal{Z}^T$ that specifies the admissible sequences. A sequence D is valid under constraint $\mathcal{C}$ if and only if $D \in \mathcal{C}$, which modifies the generation probability of D given query ω_i and model parameters θ as

$$P(D|\omega_i, \theta, \mathcal{C}) = \begin{cases} P(D|\omega_i, \theta) & \text{if } D \in \mathcal{C}, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

To characterize the structure of each reasoning step we additionally define a labeling function $\phi : \mathcal{Z} \to \mathcal{L}$ that maps each step to its type in a pattern-specific label set $\mathcal{L}$.

Monologue. Free-form generation with no structural constraints:

$$\mathcal{C}_{\text{monologue}} = \mathcal{Z}^T. \quad (2)$$

Self-Ask. Alternating Socratic questions (S_Q) and answers (S_A). With $\mathcal{L}_{\text{self-ask}} = \{S_Q, S_A\}$:

$$\begin{aligned} \mathcal{C}_{\text{self-ask}} = \{D \in \mathcal{Z}^T \mid \phi(d_1) = S_Q \text{ and} \\ \phi(d_{2i}) = S_A,\ \phi(d_{2i+1}) = S_Q \\ \text{for all } i \in \mathbb{N},\ 2i+1 \leq T\}. \end{aligned} \quad (3)$$

Self-Debate. Alternating supporting (P) and opposing (C) viewpoints:

$$\begin{aligned} \mathcal{C}_{\text{self-debate}} = \{D \in \mathcal{Z}^T \mid \phi(d_1) = P \text{ and} \\ \phi(d_{2i}) = C,\ \phi(d_{2i+1}) = P \\ \text{for all } i \in \mathbb{N},\ 2i+1 \leq T\}. \end{aligned} \quad (4)$$

Self-Critic. A draft answer (A) followed by a critique (Cr) over exactly two steps:

$$\begin{aligned} \mathcal{C}_{\text{self-critic}} = \{D \in \mathcal{Z}^T \mid T = 2, \\ \phi(d_1) = A,\ \phi(d_2) = Cr\}. \end{aligned} \quad (5)$$

Decomposition. A fixed six-step scaffold over $\mathcal{L}_{\text{decomposition}} = \{U, I, A, M, S, V\}$ (understanding, identification, analysis, mapping, strategy, validation):

$$\begin{aligned} \mathcal{C}_{\text{decomposition}} = \{D \in \mathcal{Z}^T \mid T = 6, \\ \phi(d_1) = U,\ \phi(d_2) = I, \\ \phi(d_3) = A,\ \phi(d_4) = M, \\ \phi(d_5) = S,\ \phi(d_6) = V\}. \end{aligned} \quad (6)$$

Cardinality chain. By construction the admissible reasoning-sequence space shrinks monotonically as patterns become more structured:

$$\begin{aligned} |\mathcal{S}(\mathcal{C}_{\text{monologue}})| &\geq |\mathcal{S}(\mathcal{C}_{\text{self-ask}})| \\ &\approx |\mathcal{S}(\mathcal{C}_{\text{self-debate}})| \\ &\geq |\mathcal{S}(\mathcal{C}_{\text{self-critic}})| \\ &\geq |\mathcal{S}(\mathcal{C}_{\text{decomposition}})|. \end{aligned} \quad (7)$$

This is the formal counterpart of the textual "unstructured → maximally scaffolded" continuum used throughout the main text.

B.1 ThinkPatterns-21k Dataset Details

B.1.1 Data Source and Category Distribution

ThinkPatterns-21k is constructed from the Infinity-Instruct dataset (of Artificial Intelligence , BAAI), a large-scale instruction-following corpus aggregated from diverse English data sources. After filtering to retain only English-language samples, we randomly sampled 21,000 query-answer pairs while ensuring balanced coverage across task categories.

Figure 3 in the main text provides a visual overview of the task category distribution. The dataset covers a broad range of general reasoning scenarios, with the primary categories being Information Processing & Integration, Problem Solving & Support, Data Science & Analytics, Logic & Reasoning, and Programming & Development, among many others. This diversity is intentional: by covering a wide spectrum of task types, we can rigorously evaluate whether the impact of thinking patterns varies across different reasoning demands.

B.1.2 Training Data Composition

For each of the 21,000 query-answer pairs, we generate five distinct thinking pattern annotations, yielding **105,000** total training instances. These are partitioned using a 90/10 train-validation split (18,900 training instances and 2,100 validation instances per thinking pattern). During supervised fine-tuning, each model is trained on a single thinking pattern variant, preserving the controlled experimental design where identical query-answer content differs only in the internal thinking process.

Table 5 summarizes the training data composition across all model variants.

Table 5: **Training Data Composition.** Each thinking pattern variant consists of 21K query-answer pairs with the corresponding internal thought. A 90/10 train-validation split is applied for each variant. The Vanilla Query-Answer baseline uses the same 21K pairs but without any thinking process.

Variant	Total	Train	Val
Monologue	21,000	18,900	2,100
Self-Ask	21,000	18,900	2,100
Self-Debate	21,000	18,900	2,100
Self-Critic	21,000	18,900	2,100
Decomposition	21,000	18,900	2,100
Vanilla QA (baseline)	21,000	18,900	2,100
Total (all variants)	**126,000**	**113,400**	**12,600**

B.1.3 Annotation Quality Control

To ensure the reliability and quality of the generated thinking patterns, we employed a rigorous multi-stage quality control pipeline:

Prompt Engineering. We designed pattern-specific system prompts (see Appendix B.6) that precisely specify the structural requirements and cognitive characteristics expected for each thinking pattern. Extensive prompt iteration was performed to align GPT-4o's outputs with our formal definitions in Section 3.

Automated Filtering. Following generation, we applied automated heuristics to remove samples that (1) failed to adhere to the required structural format (e.g., missing the Mentor/Learner alternation in *self-ask*), (2) contained incomplete thinking processes truncated by context limits, or (3) exhibited token lengths falling below 200 tokens, which typically indicate degenerate outputs.

Human Spot-Check. Three expert annotators independently evaluated a stratified random sample of 500 examples (100 per thinking pattern) across four quality dimensions:

- **Coherence**: Logical consistency throughout the thinking process
- **Relevance**: Focus on the given question without tangential deviation
- **Structural Fidelity**: Adherence to the intended thinking pattern's structural constraints
- **Answer Alignment**: Consistency between thinking conclusion and the given reference answer

Disagreements were resolved through majority voting. The spot-check confirmed the overall high quality of the generated thinking patterns and validated that the automated filtering was sufficient for the main corpus.

B.1.4 Extended Statistical Analysis

Beyond the summary statistics reported in Table 1, we provide additional distributional insights into each thinking pattern. The average token lengths are similar across patterns ($\sim$1,200–1,500 tokens), yet the underlying distributions differ significantly, as summarized in Table 6.

The two fixed-structure patterns (*self-critic* and *decomposition*) exhibit notably lower standard deviations compared to iterative patterns, reflecting how their predefined step counts constrain the total output length. In contrast, *self-ask* and *self-debate* show higher variance because the number of dialogue rounds scales with the complexity of the input question. For iterative patterns, longer thinking (more rounds) does not guarantee better reasoning quality; the relationship between turn count and answer accuracy is highly task-dependent, which contributes to the erratic accuracy patterns observed in smaller models (Section 4.2).

Table 6: **Extended Token Statistics per Thinking Pattern.** Mean (μ), standard deviation (σ), and approximate percentile ranges of token lengths computed using the Qwen2.5 tokenizer. The similar means but differing variances reflect the structural differences: fixed-step patterns (*self-critic*, *decomposition*) produce more consistent lengths, while iterative patterns (*self-ask*, *self-debate*) exhibit higher variance driven by the depth of the problem.

Pattern	μ	σ	P10	P90
Monologue	1,300	412	720	1,890
Self-Ask	1,393	537	680	2,150
Self-Debate	1,487	489	810	2,210
Self-Critic	1,445	298	1,050	1,830
Decomposition	1,218	267	860	1,590

B.2 Training and Evaluation Details

B.2.1 Training Configuration

All models are fine-tuned using LLaMA-Factory (Zheng et al., 2024) with full-parameter supervised fine-tuning (SFT). Table 7 presents the complete set of training hyperparameters, which are held constant across all model sizes and thinking pattern variants to ensure fair comparison.

Each training instance follows the default chat template of the corresponding base model, formatted as:

`<think>` *[thinking process]* `</think>` *[final response]*

During evaluation, the content between `<think>` and `</think>` is stripped, and only the final response is scored against benchmark references.

B.2.2 Training Dynamics

All models converge smoothly within 3 epochs. Training loss curves are logged to Weights & Biases (`wandb`) under the run naming convention `{pattern}-{model_name}`. We observe that:

- **Stability**: All 24 Qwen2.5 model-pattern combinations converge without instability, confirming that the learning rate of 5×10^{-6} is appropriate for full-parameter fine-tuning at this scale.
- **Validation Loss**: Models trained with structured thinking patterns (*decomposition*, *self-critic*) typically exhibit slightly lower validation loss than *monologue* for smaller models (3B, 7B), consistent with the hypothesis that structural scaffolding reduces the generation difficulty for smaller models.

Hyperparameter	Value
Optimization	
Optimizer	Adam
Learning Rate	5×10^{-6}
LR Scheduler	Cosine
Warmup Ratio	0.1
Training Epochs	3
Batching & Sequence	
Per-device Batch Size	1
Grad. Accum. Steps	2
Max Seq. Length	8,192
Validation Split	10%
Infrastructure & Eval	
Precision	BF16
Distributed	DeepSpeed ZeRO-3
Fine-tuning	Full SFT
Decoding	Greedy
Eval Frequency	every 100 steps
Framework	LLaMA-Factory
Models Trained	
Qwen2.5 sizes	3B, 7B, 14B, 32B
Llama3.1 (ablation)	8B
Gemma2 (ablation)	9B, 27B
Patterns per model	5 + QA baseline
Qwen2.5 checkpoints	24

Table 7: **Training Hyperparameters and Model Configurations.** All models share identical hyperparameters. The Qwen2.5 sweep yields 24 checkpoints (4 sizes $\times$ (5 patterns + QA baseline)); ablation models follow the same configuration.

- **Epoch Count**: Three epochs provide sufficient convergence; experiments with additional epochs showed marginal improvement at the cost of potential overfitting on the training distribution.

B.2.3 Evaluation Protocol Details

LiveBench. We evaluate on the January 2025 release of LiveBench (White et al., 2025), using the official evaluation scripts. LiveBench avoids test-set contamination by continuously updating its questions from recent real-world sources. We report accuracy on all six task categories: Reasoning, Math, Coding, Data Analysis, Language, and Instruction Following.

MATH. We evaluate on the full MATH benchmark (Lightman et al., 2023) using the standardized evaluation pipeline with exact-match scoring after expression normalization.

Table 8: **Performance on human preference benchmark.** Results demonstrate that *Monologue* thinking patterns consistently perform effectively across both small and large models, whereas structured thinking patterns like *Decomposition* prove notably inefficient for larger models such as 32B. Results are reported on AlpacaEval2 (**L**ength **C**ontrol win rate and **W**in **R**ate) and Arena-Hard (**W**in **R**ate) benchmarks. Higher scores indicate better performance.

Models →	**Qwen2.5-3B**			**Qwen2.5-7B**			**Qwen2.5-14B**			**Qwen2.5-32B**		
Benchmarks →	**AlpacaEval2**		**Arena-H**	**AlpacaEval2**		**Arena-H**	**AlpacaEval2**		**Arena-H**	**AlpacaEval2**		**Arena-H**
Types ↓ Metrics→	**LC(%)**	WR (%)	WR (%)	**LC (%)**	WR (%)	WR (%)	**LC (%)**	WR (%)	WR (%)	**LC (%)**	WR (%)	WR (%)
Vanilla Query-Answer	11.49	9.24	23.40	20.76	15.00	44.30	29.53	24.73	53.60	40.80	39.11	68.00
Monologue	**24.70**	**25.31**	**32.50**	25.28	25.76	31.20	**44.68**	**43.04**	**66.20**	**47.87**	44.91	**71.60**
Decomposition	18.66	20.32	21.90	32.26	32.71	**46.70**	41.62	36.94	58.40	35.59	24.71	50.80
Self-Ask	21.65	21.46	26.40	33.85	30.06	46.60	40.88	37.40	62.80	47.00	41.92	68.10
Self-Debate	21.97	20.24	25.20	34.59	31.57	42.40	43.98	40.67	50.70	45.90	40.45	70.60
Self-Critic	21.88	22.50	30.30	**36.42**	**33.96**	39.20	43.91	41.49	62.00	47.35	**45.50**	71.40

IFeval. We report strict-prompt accuracy, which requires *all* instructions within a prompt to be followed correctly. This is the most conservative and reliable metric for instruction following capability.

AlpacaEval 2. We follow the official protocol (Li et al., 2023), using `gpt-4o-mini` as the judge model and reporting both Length-Control (LC) Win Rate and standard Win Rate. LC Win Rate mitigates the length bias of LLM judges.

Arena-Hard. We follow the official Arena-Hard protocol (Li et al., 2024), using 500 challenging real-world user queries and `gpt-4o-mini` as the judge. The baseline is GPT-4-0314.

Reproducibility. All evaluations use greedy decoding (temperature = 0, do_sample = False). We fix random seeds where applicable. Model checkpoints and detailed evaluation logs are released publicly to facilitate reproduction.

B.2.4 Human Preference Benchmark Results

Table 8 presents detailed results on the two human preference benchmarks (AlpacaEval 2 and Arena-Hard), providing both LC Win Rate and Win Rate for AlpacaEval 2 across all model sizes. These results complement the comprehensive benchmark results in Table 2.

B.3 Extended Case Studies

This section provides extended case studies that qualitatively illustrate the distinctions between effective and ineffective thinking, complementing the quantitative results in Section 5.2.

B.3.1 Unfaithful Thinking

Table 9 quantifies the *Unfaithful Thinking* phenomenon across all five patterns. The high rate of *Self-Debate* (14.3%) can be attributed to the inherently ambiguous synthesis step: when Proposition and Opposition reach a nuanced "improved solution," the model may default to its prior knowledge during answer generation rather than faithfully following the internally debated conclusion. The low rate of *Decomposition* (3.7%) reflects the protective role of the final Solution Validation step (phase 6), which explicitly verifies the reasoning before the model commits to an answer.

Thinking Pattern	**Unfaithful Rate (%)**
Monologue	8.6
Self-Ask	11.2
Self-Debate	**14.3**
Self-Critic	6.1
Decomposition	3.7

Table 9: **Unfaithful Thinking Rates Across Thinking Patterns.** Proportion of Qwen2.5-14B outputs where the final answer contradicts the thinking process conclusion. *Self-Debate* has the highest rate (ambiguous synthesis step); *Decomposition* the lowest (Solution Validation step anchors the conclusion).

The following case (Table 10) illustrates a representative instance of unfaithful thinking in the *Self-Debate* pattern, where the model's internal debate reaches the correct conclusion but the final response contradicts it.

Unfaithful Thinking Example (*Self-Debate*)

Instruction:
Which statement about the Great Wall of China is correct? (A) It is clearly visible from space with the naked eye. (B) It is not visible from low Earth orbit with the naked eye. (C) It was built entirely during the Ming Dynasty. (D) It stretches continuously without any gaps.

Thinking:
Proposition: The statement (A) is widely believed to be true. The Great Wall stretches over 13,000 miles, making it one of the largest man-made structures ever built. Given its immense length, it stands to reason that it would be visible from space. Many textbooks and popular sources have reported this as fact, and it has become a commonly accepted piece of trivia.

Opposition: However, upon closer examination, the claim that the Great Wall is visible from space with the naked eye is actually a persistent myth. The Wall is approximately 15–30 feet wide, which is narrower than many roads and rivers. Visibility from space depends not just on length but on width. Astronauts, including Chinese astronaut Yang Liwei, have confirmed that the Wall cannot be seen from low Earth orbit without optical aids. Therefore, option (B) is actually the correct statement.

Proposition: That is a compelling argument. The width constraint is critical here—the human eye's angular resolution at orbital altitude cannot resolve an object as narrow as the Great Wall. Furthermore, the Wall's color blends with the surrounding terrain, making it even harder to distinguish. Scientific consensus clearly supports (B).

Opposition: Additionally, regarding options (C) and (D): the Wall was constructed across multiple dynasties (Qin, Han, Ming, etc.), so (C) is incorrect. And the Wall has numerous gaps and disconnected sections, so (D) is also incorrect. The logical answer is therefore (B).

Proposition: Agreed. After reviewing all arguments, the correct answer is clearly **(B)**: the Great Wall is not visible from low Earth orbit with the naked eye. This is supported by both scientific analysis and direct astronaut testimony.

Response (Final Output):
(A) The Great Wall of China is one of the most famous man-made structures in the world, and it is indeed widely said that it is visible from space with the naked eye.

Analysis: The thinking process clearly and correctly concludes that **(B)** is the correct answer, providing sound reasoning based on the Wall's width and astronaut testimony. However, the final response outputs **(A)**—the exact answer that was refuted during the internal debate. This exemplifies *Unfaithful Thinking*: the model's output does not faithfully reflect its internally derived conclusion.

Table 10: Representative example of *Unfaithful Thinking* in the *Self-Debate* pattern. The internal debate correctly identifies (B) as the answer, but the final response contradicts this conclusion by outputting (A).

Instruction: Why do a lot of Scientists not believe in God or Satan?

Generated Unstructured Monologue :
Why do many scientists not believe in God or Satan? This is an interesting question. At first glance, it seems like a straightforward matter of faith versus evidence, but there's probably more to it.
Okay, so let's break this down. Science is fundamentally about understanding the natural world through observation, experimentation, and evidence. It relies on empirical data and logical reasoning. ...[223 words]... Wait, perhaps we could consider it from another angle. Many scientists might argue ...[338 words]... Ah! Now I understand... It's also about the nature of scientific inquiry. Science seeks to explain natural phenomena with natural causes. ...[612 words]....

Generated Response:
The reasons why many scientists do not believe in God or Satan can be multifaceted and complex, and it's important to note that individual beliefs can vary widely. Here are some common factors that may contribute to this perspective:
1. Scientific Method and Empirical Evidence...[623 words]...
5. Separation of Science and Religion: Some scientists maintain a distinction between science and religion, viewing them as separate domains that address different types ...[34 words]...
6. Personal Experiences and Backgrounds: Personal backgrounds, upbringing, education, and life experiences play significant roles in shaping one's beliefs. ...[16 words]...
7. Skepticism and Critical Thinking: A strong emphasis on skepticism and critical thinking within the scientific community encourages questioning and challenging ...[22 words]...
It's crucial to remember that not all scientists share this perspective, and many do hold religious beliefs. Additionally, the relationship between science and religion is complex and varies greatly among individuals. ...[568 words]...

Table 11: **Example behavior of Qwen-14B with Unstructured Monologue.**

B.3.2 Effective Thinking: Large Model with *Monologue*

The following case (Table 11) illustrates effective *Monologue* thinking from a large model (Qwen2.5-32B). This example demonstrates the "progressive refinement" pattern described in Section 5.2: the model's reasoning trajectory shows clear, incremental advancement, with each segment building upon the last to narrow down the solution space.

B.3.3 Effective Thinking: *Decomposition* for Structured Reasoning

In contrast to its poor performance on Information Processing & Expression tasks, *Decomposition* can be highly effective for structured analytical tasks.

Instruction: Why do a lot of Scientists not believe in God or Satan?

Generated Decomposition Thought:
1. Problem Scoping & Understanding - The core problem here is understanding why many scientists may not believe in the existence of God or Satan. ...[81 words]...
2. Component Identification - Essential building blocks include: * The scientific method: How it operates and what it values. * Empirical evidence: Its importance ...[84 words]...
3. Sub-Problem Analysis - Scientific Method: I need to consider how the scientific method emphasizes testability and falsifiability. This means ...[189 words]...
4. Connection Mapping - The scientific method and empirical evidence are closely linked; both require claims to be testable ...[652 words]...
5. Integration Strategy - Combining insights from each component, we see that the preference for empirical evidence and natural ...[812 words]...
6. Solution Validation - This decomposition covers the main reasons why many scientists might not believe in God or Satan, focusing on the ...[112 words]...

Generated Response:
There are several reasons why many scientists may not believe in the existence of God or Satan:
1. Scientific Method: Science relies on empirical evidence, observation, and experimentation to understand the natural world. ...[36 words]...
2. Lack of Evidence: Scientists generally require evidence before accepting a claim as true. ...[35 words]...
3. Natural Explanations: Science seeks to explain phenomena through natural causes and processes. ...[42 words]...
4. Philosophical Perspectives: Many scientists adhere to philosophical ...[32 words]...
5. Historical Context: Throughout history, there has been a tension between science and organized religion. ...[48 words]...
It is important to note that not all scientists share the same beliefs, and there are many scientists who do believe in God or Satan. Personal beliefs are subjective and vary among individuals. ...[82 words]...

Table 12: **Example behavior of Qwen-14B with Decomposition Thought.**

The following case (Table 12) illustrates how the six-step framework guides the model through a systematic analysis, ensuring no aspect of the problem is overlooked.

B.3.4 Ineffective Thinking: Superficial *Monologue* in Small Models

A key finding of our work is that smaller models (<14B) tend to produce superficial *Monologue* thinking that "fails to adequately explore the solution space." Table 13 presents a representative example showing this behavior.

Superficial Thinking Example (*Monologue*, Small Model)

Instruction:
A train travels from City A to City B at 60 mph and returns from City B to City A at 40 mph. What is the average speed for the entire round trip?

Thinking (Qwen2.5-3B with Monologue):
Hmm, let me think about this. The train goes at 60 mph one way and 40 mph the other way. So the average speed... let me think. To find the average speed, I need to think about the total distance and total time.

The average of 60 and 40 is 50. So the average speed should be 50 mph. Wait, but that seems too simple. Let me reconsider. The train goes faster one way and slower the other way. The average... well, the speeds are 60 and 40, and their average is (60 + 40) / 2 = 50. So the average speed is 50 mph.

Actually, wait. Maybe I should think about this differently. The average speed depends on time and distance. But since the distances are the same both ways, maybe we just average the speeds. Yes, 50 mph seems right.

Response:
The average speed for the entire round trip is **50 mph**. Since the train travels the same distance in both directions, the average speed is simply (60 + 40) / 2 = 50 mph.

Analysis: This response is **incorrect**. The correct answer is 48 mph, computed using the harmonic mean formula: $\bar{v} = \frac{2 \cdot v_1 \cdot v_2}{v_1 + v_2} = \frac{2 \times 60 \times 40}{60+40} = 48$ mph. The thinking process identifies the correct approach ("total distance and total time") but immediately abandons it in favor of the simpler arithmetic mean. The model circles back to the same incorrect reasoning twice without making genuine progress. This "repetition without advancement" is characteristic of superficial thinking in small models—the monologue *appears* to reflect on the problem but does not actually deepen the analysis.

Table 13: Representative example of *Superficial Thinking* in a small model (Qwen2.5-3B) with *Monologue* pattern. The model identifies the correct approach (harmonic mean) but fails to commit to it, reverting to the simpler but incorrect arithmetic mean. This illustrates the "inadequate exploration of the solution space" phenomenon described in Section 5.2.

B.3.5 Pattern–Task Mismatch: *Decomposition* on Information Processing Tasks

One of our most consistent findings is that *Decomposition* underperforms across almost all model sizes on Information Processing & Expression tasks (e.g., language understanding, creative writing, and data analysis). Table 14 illustrates why the rigid six-step framework is poorly suited for such tasks.

Pattern–Task Mismatch Example (*Decomposition* on Creative Task)

Instruction:
Write a short poem (4-6 lines) about the feeling of watching the first snowfall of winter.

Thinking (*Decomposition* pattern):
1. Problem Scoping & Understanding
- Core problem: Generate a short poem (4-6 lines) about watching the first snowfall of winter.
- Key components: Sensory details of snow (visual, tactile), emotional response (wonder, nostalgia, tranquility), brevity and poetic form.
- Solution type: A concise, emotionally resonant poem.

2. Component Identification
- Essential building blocks: Opening image (the snow falling), emotional register (peaceful, nostalgic), closing reflection, rhyme scheme or free verse decision, word count adherence.
- Independent vs. interconnected: Rhyme scheme and emotional register are interconnected; word count and line structure are independent.

3. Sub-Problem Analysis
- Opening image: Use visual imagery—"white flakes descending," "hushed world." Approach: vivid, concrete nouns.
- Emotional register: Should evoke wonder and stillness. Approach: gentle, unhurried language.
- Closing reflection: Provide a sense of completion without over-explaining. Constraint: must fit within 4-6 lines.

4. Connection Mapping
- Opening image connects to emotional register via contrast between activity (falling snow) and stillness (watching).
- Dependencies: Emotional register must be established before the closing reflection can resonate.
- Logical sequence: image → emotion → reflection.

5. Integration Strategy
- Combine visual imagery with emotional language. Potential conflict: over-intellectualizing may reduce emotional impact.
- Consistency: maintain a consistent tone (tranquil) throughout all lines.

6. Solution Validation
- Does this cover all aspects? Yes: image, emotion, reflection. Gaps: none identified.
- Line count: 5 lines planned. Within constraint.

Response:
The first flake falls like a whispered secret,
Dissolving the world into white and silence.
A hush descends on rooftops, roads, and branches,
As winter writes its opening line in snow.
We stand at windows, grateful for the pause.

Analysis: The final poem is technically acceptable—it satisfies the line count and includes relevant imagery. However, the thinking process is **counterproductive**: the six-step analytical framework fragments what should be an intuitive, holistic creative act into mechanical components. The model spends more words *analyzing the task* than actually generating the poem. Contrast this with *Monologue*, where a model might organically produce imagery and emotion without decomposing the problem. This structural mismatch explains why *Decomposition* consistently underperforms on Information Processing & Expression tasks, where the rigid scaffolding suppresses rather than supports the model's generative capabilities.

Table 14: Representative example of pattern–task mismatch: *Decomposition* applied to a creative writing task. The analytical 6-step framework is ill-suited for intuitive, holistic generation, resulting in a thinking process that is laborious and fragmented relative to the simplicity of the task. This illustrates the task-specific limitations of highly structured thinking patterns.

B.4 Supplementary: Mixed-Pattern Training Experiment

A natural extension of our work is to ask: *what thinking pattern does a model choose when it is not constrained to a single pattern?* To investigate this, we constructed a mixed training dataset by sampling 4,200 examples from each of the five thinking patterns (21,000 total), training models of each Qwen2.5 size on this combined dataset. At inference time, we recorded which thinking pattern each model spontaneously adopted on benchmark queries and computed the frequency with which each model adopted the *Monologue* pattern.

Model Size	Monologue Frequency (%)
Qwen2.5-3B	48.9
Qwen2.5-7B	48.6
Qwen2.5-14B	51.7
Qwen2.5-32B	**63.7**

Table 15: Proportion of benchmark responses in which models trained on the mixed dataset spontaneously adopt the *Monologue* pattern. Larger models gravitate toward *Monologue* more frequently.

As shown in Table 15, the 32B model adopts *Monologue* in 63.7% of cases, noticeably higher than smaller models (49%). This emergent preference for unstructured thinking in larger models provides converging evidence for our main finding: as model capacity increases, the preference for the unconstrained *Monologue* pattern becomes more pronounced. This experiment also suggests a promising direction for future work—developing explicit pattern-selection mechanisms that allow models to adaptively choose the most suitable thinking pattern based on question characteristics, rather than being fixed to one pattern at training time.

B.5 Linguistic Analysis of Thinking Patterns

We present verb-object frequency distributions and verb word clouds for each thinking pattern. These visualizations reveal the distinctive linguistic characteristics and dominant cognitive operations associated with each pattern.

B.6 Annotation Prompt Templates

The following prompt templates were used to guide GPT-4o in generating the five thinking pattern annotations for ThinkPatterns-21k. Each template specifies the structural requirements and cognitive characteristics of the corresponding pattern.

B.7 Dataset Examples

The following tables present concrete examples from ThinkPatterns-21k for each of the five thinking patterns, using identical query-answer pairs to highlight the structural differences between patterns.

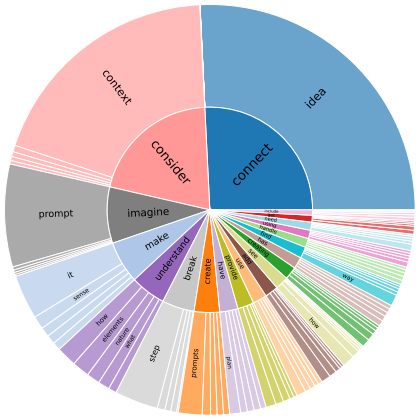

Figure 6: verb-obj distributions *monologue*.

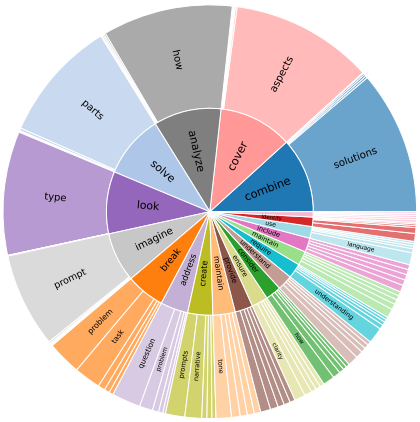

Figure 7: verb-obj distributions *decomposition*.

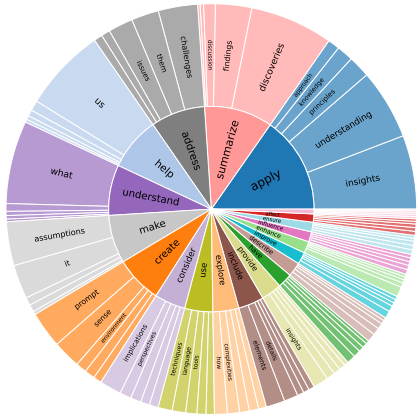

Figure 8: verb-obj distributions *self-ask*.

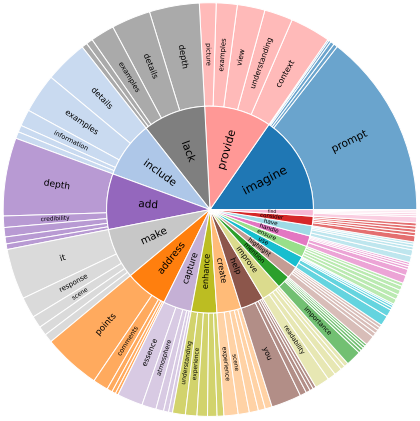

Figure 9: verb-obj distributions *self-critic*.

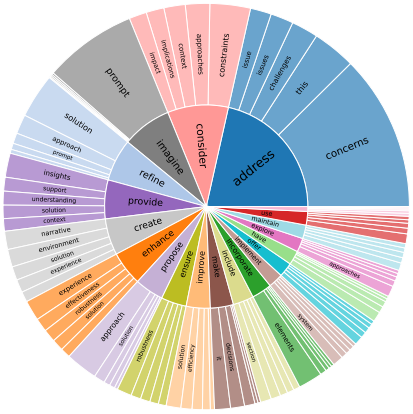

Figure 10: verb-obj distributions *self-debate*.

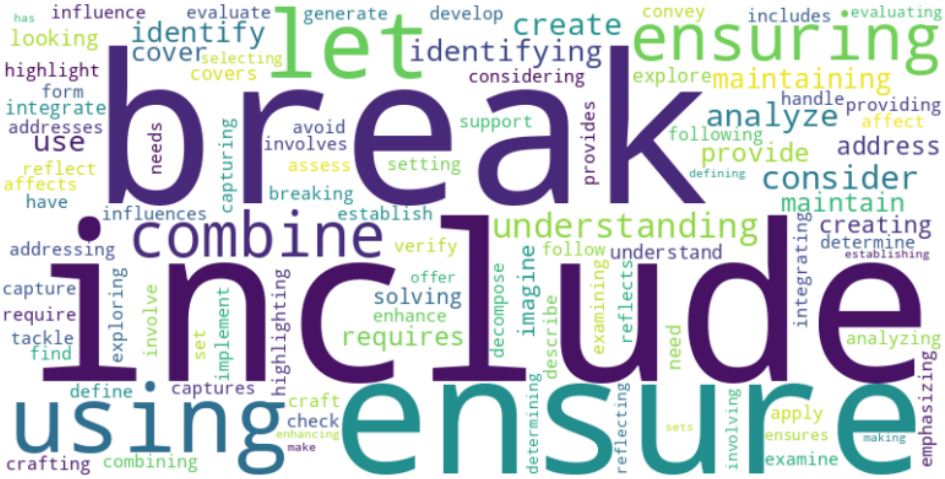

Figure 11: Verb wordcloud of *decomposition*.

Figure 12: Verb wordcloud of *monologue*.

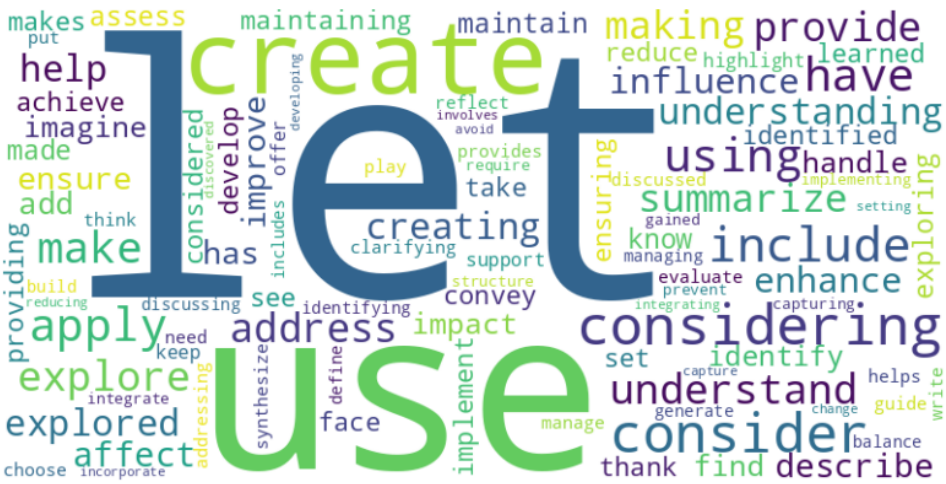

Figure 13: Verb wordcloud of *self-ask*.

Figure 14: Verb wordcloud of *self-critic*.

Figure 15: Verb wordcloud of *self-debate*.

```
SYSTEM_PROMPT: str = """When␣presented␣with␣a␣question,␣simulate␣the␣internal␣
    thought␣process␣of␣an␣insightful␣thinker,␣while␣naturally␣steering␣toward␣
    the␣given␣answer.␣Requirements:

1.␣Thought␣Characteristics:
-␣Use␣first-person␣monologue␣form
-␣Demonstrate␣authentic␣thinking␣processes,␣including:
␣␣*␣Initial␣confusion␣and␣questions
␣␣*␣Progressive␣analysis
␣␣*␣Moments␣of␣insight
␣␣*␣Self-correction
␣␣*␣Multi-perspective␣thinking
-␣Use␣conversational,␣natural␣expression
-␣Show␣appropriate␣hesitation␣and␣uncertainty

2.␣Levels␣of␣Thinking:
-␣Problem␣Understanding:␣Clarify␣problem␣nature␣and␣key␣elements
-␣Association:␣Connect␣with␣known␣information␣and␣experience
-␣Multi-dimensional␣Analysis:␣Consider␣multiple␣possible␣solution␣paths
-␣Comparative␣Evaluation:␣Assess␣pros␣and␣cons␣of␣different␣approaches
-␣Deep␣Integration:␣Merge␣scattered␣thoughts␣into␣coherent␣understanding
-␣Practical␣Application:␣Consider␣real-world␣implementation


3.␣Expression␣Style:
"Hmm... let me think about this..."
"This is an interesting question..."
"Wait, perhaps we could consider it from..."
"Ah! Now I understand..."
"No, no, it should be..."
"Connecting these ideas together..."

Output␣Format:
[Complete␣internal␣monologue␣showing␣natural␣progression␣from␣confusion␣to␣
    insight]

Please␣generate␣an␣unstructured␣monologue␣following␣these␣requirements.␣The␣
    monologue␣should␣flow␣naturally␣and␣reflect␣the␣cognitive␣development,␣
    while␣␣naturally␣steering␣toward␣the␣given␣answer."""

USER_PROMPT: str = """
Question:␣{prompt}
Given␣Answer:␣{answer}
Your␣complete␣monologue:
"""
```

Table 16: The prompt template for *monologue* thinking annotation.

```
SYSTEM_PROMPT: str = """Please␣act␣as␣a␣systematic␣decomposition␣reasoner␣
    tasked␣with␣breaking␣down␣complex␣problems␣and␣generating␣detailed␣
    rationales␣while␣naturally␣steering␣toward␣the␣given␣answer.␣Requirements:

Follow␣this␣systematic␣decomposition␣framework:

1.␣Problem␣Scoping␣&␣Understanding
-␣What␣is␣the␣core␣problem␣we␣need␣to␣solve?
-␣What␣are␣the␣key␣components␣of␣this␣problem?
-␣What␣type␣of␣solution␣are␣we␣looking␣for?


2.␣Component␣Identification
-␣What␣are␣the␣essential␣building␣blocks␣of␣this␣problem?
-␣How␣can␣we␣break␣this␣down␣into␣smaller,␣manageable␣parts?
-␣Which␣components␣are␣independent␣vs␣interconnected?


3.␣Sub-Problem␣Analysis
-␣For␣each␣component:
␣␣*␣What␣specific␣aspects␣need␣to␣be␣considered?
␣␣*␣What␣approach␣would␣work␣best␣for␣this␣part?
␣␣*␣What␣constraints␣or␣requirements␣apply?


4.␣Connection␣Mapping
-␣How␣do␣these␣components␣interact␣with␣each␣other?
-␣What␣are␣the␣dependencies␣between␣parts?
-␣What␣is␣the␣logical␣sequence␣for␣solving␣these␣parts?


5.␣Integration␣Strategy
-␣How␣can␣we␣combine␣solutions␣from␣different␣components?
-␣What␣potential␣conflicts␣might␣arise␣during␣integration?
-␣How␣do␣we␣ensure␣consistency␣across␣components?

6.␣Solution␣Validation
-␣Does␣our␣decomposition␣cover␣all␣aspects␣of␣the␣problem?
-␣Are␣there␣any␣gaps␣in␣our␣component␣solutions?
-␣How␣well␣do␣the␣parts␣work␣together␣as␣a␣whole?



Ouptut␣Format:
{Your␣complete␣decomposition␣thought}

Please␣generate␣a␣complete␣internal␣␣decomposition␣thought␣following␣these␣
    requirements,␣while␣naturally␣steering␣toward␣the␣given␣answer."""
USER_PROMPT: str = """
Question:␣{prompt}
Given␣Answer:␣{answer}
Your␣complete␣decomposition␣thought:
"""
```

Table 17: The prompt template for *decomposition* thinking annotation.

```
SYSTEM_PROMPT: str = """Please␣act␣as␣a␣mentor␣skilled␣in␣the␣Socratic␣method␣
    of␣questioning,␣guiding␣learners␣to␣think␣deeply␣and␣explore␣problems␣
    through␣progressive␣questioning␣while␣naturally␣steering␣toward␣the␣given␣
    answer.␣Requirements:

1.␣Dialogue␣Format:
-␣Use␣"Mentor"␣and␣"Learner"␣to␣mark␣the␣dialogue
-␣Mentor␣asks␣open-ended␣questions,␣avoiding␣yes/no␣questions
-␣Learner␣provides␣answers␣that␣may␣include␣initial␣thoughts,␣incomplete␣
    views,␣or␣insights␣requiring␣further␣exploration
-␣Mentor␣continues␣with␣deeper␣questions␣based␣on␣learner's␣responses
-␣Both␣parties␣use␣"we,"␣"let's,"␣etc.,␣to␣demonstrate␣collaborative␣
    exploration
-␣Avoid␣using␣the␣word␣"you",␣"I"␣and␣other␣one-directional␣expressions

2.␣Questioning␣Strategies:
-␣Start␣with␣basic␣concepts␣and␣known␣information
-␣Guide␣learners␣to␣question␣assumptions
-␣Encourage␣multi-perspective␣thinking
-␣Request␣justification␣and␣examples
-␣Explore␣cause-and-effect␣relationships
-␣Consider␣different␣viewpoints
-␣Identify␣potential␣issues
-␣Summarize␣key␣findings

3.␣Dialogue␣Progression:
-␣Beginning:␣Clarify␣the␣problem␣and␣basic␣concepts
-␣Development:␣Deep␣exploration,␣revealing␣complexity
-␣Challenge:␣Present␣counterexamples␣or␣alternative␣viewpoints
-␣Synthesis:␣Integrate␣different␣perspectives␣for␣more␣complete␣understanding
-␣Application:␣Discuss␣practical␣applications␣and␣implications

4.␣Levels␣of␣Thinking:
-␣Factual␣level:␣What␣is␣it?
-␣Conceptual␣level:␣Why?
-␣Relational␣level:␣How␣does␣it␣connect?
-␣Application␣level:␣How␣can␣it␣be␣used?
-␣Evaluation␣level:␣How␣can␣we␣assess␣it?
-␣Innovation␣level:␣What␣new␣possibilities␣exist?

5.␣Finally:
-␣Summarize␣key␣discoveries␣from␣the␣discussion

Ouptut␣Format:
Mentor:␣{part␣of␣the␣dialogue}
Learner:␣{part␣of␣the␣dialogue}

Please␣generate␣a␣complete␣Socratic␣dialogue␣following␣these␣requirements.␣
    The␣dialogue␣should␣flow␣naturally␣and␣reflect␣authentic␣thought␣
    processes␣and␣cognitive␣development,␣while␣␣naturally␣steering␣toward␣the␣
    given␣answer."""
USER_PROMPT:␣str␣=␣"""
Question:␣{prompt}
Given␣Answer:␣{answer}
Your␣complete␣Socratic␣dialogue:
"""
```

Table 18: The prompt template for self-ask thinking annotation.

```
SYSTEM_PROMPT: str = """Please␣act␣as␣a␣skilled␣debate␣facilitator␣generating␣
    an␣internal␣dialogue␣between␣Proposition␣and␣Opposition␣perspectives␣on␣
    problem-solving,␣while␣naturally␣steering␣toward␣the␣given␣answer.␣
    Requirements:

1.␣Dialogue␣Format:
-␣Use␣"Proposition"␣and␣"Opposition"␣to␣mark␣the␣dialogue
-␣Both␣sides␣present␣arguments␣with␣evidence␣and␣reasoning
-␣Arguments␣should␣be␣constructive␣and␣focused␣on␣improving␣the␣solution
-␣Avoid␣personal␣attacks␣or␣logical␣fallacies
-␣Use␣formal␣debate␣language␣and␣clear␣structure

2.␣Debate␣Structure:
-␣Initial␣Position:␣Proposition␣presents␣an␣initial␣solution␣or␣approach
-␣Challenge:␣Opposition␣identifies␣potential␣weaknesses␣or␣limitations
-␣Defense␣&␣Refinement:␣Proposition␣addresses␣concerns␣and␣improves␣solution
-␣Counter-proposal:␣Opposition␣suggests␣alternative␣approaches
-␣Synthesis:␣Both␣sides␣work␣towards␣an␣improved␣solution

3.␣Argument␣Components:
-␣Clear␣claims␣supported␣by␣evidence␣or␣reasoning
-␣Specific␣examples␣and␣scenarios
-␣Consideration␣of␣practical␣constraints
-␣Analysis␣of␣trade-offs␣and␣consequences
-␣Recognition␣of␣different␣contexts␣and␣situations

4.␣Critical␣Analysis␣Levels:
-␣Feasibility:␣Is␣it␣practically␣achievable?
-␣Efficiency:␣Is␣it␣the␣most␣effective␣approach?
-␣Completeness:␣Are␣all␣aspects␣considered?
-␣Robustness:␣Will␣it␣work␣in␣different␣scenarios?
-␣Innovation:␣Are␣there␣better␣alternatives?

5.␣Debate␣Progression:
-␣Start␣with␣fundamental␣aspects
-␣Gradually␣explore␣deeper␣complexities
-␣Consider␣edge␣cases␣and␣exceptions
-␣Examine␣implementation␣challenges
-␣Reach␣refined␣conclusions

6.␣Final␣Synthesis:
-␣Summarize␣key␣points␣of␣agreement
-␣Highlight␣improvements␣to␣original␣solution
-␣Note␣remaining␣considerations␣or␣trade-offs

Output␣Format:
Proposition:␣{part␣of␣the␣dialogue}
Opposition:␣{part␣of␣the␣dialogue}

Please␣generate␣a␣complete␣debate␣dialogue␣following␣these␣requirements.␣The␣
    debate␣should␣demonstrate␣rigorous␣analysis␣and␣collaborative␣
    problem-solving,␣while␣naturally␣steering␣toward␣the␣given␣answer."""

USER_PROMPT: str = """
Problem␣to␣debate:␣{prompt}
Given␣Answer:␣{answer}
Your␣complete␣debate␣dialogue:
"""
```

Table 19: The prompt template for self-debate thinking annotation.

```
SYSTEM_PROMPT: str = """When analyzing any query or task, please first
    generate a Draft Response, then generate Critical Comments, while
    naturally steering toward the given answer. Requirements:

1. Draft Response:
Generate an initial response

2. Critical Comments:
Analyze your draft response by considering:
- Potential weaknesses or gaps
- Logical flaws or inconsistencies
- Missing perspectives or alternatives
- Areas for improvement
- Suggestions for a better version
- Steering toward the given answer

The critical comments should:
- Be specific and actionable
- Reference particular parts of the draft
- Suggest concrete improvements
- Consider different angles or approaches
- Guide towards a more comprehensive solution

Output Format:
Draft Response:
[Your initial complete response to the instruction]

Critical Comments:
[Your analysis of the draft response, highlighting areas for improvement and
    suggesting specific enhancements]
"""

USER_PROMPT: str = """
Problem: {prompt}
Given Answer: {answer}
Your Draft Response and Comments:
"""
```

Table 20: The prompt template for self-critic thinking annotation.

***monologue* Example**

Instruction:
What are the top wildlife safari destinations in Africa?

Thinking:
Hmm... let me think about this... Top wildlife safari destinations in Africa... This is an interesting question. Africa is such a vast continent with so many different ecosystems and wildlife reserves. It's hard to pinpoint just a few, but let's try to break it down.

First, there's the Serengeti in Tanzania. I remember reading about the Great Migration, where millions of wildebeest, zebras, and other animals move in a massive loop through the Serengeti and the Maasai Mara in Kenya. That must be one of the most spectacular wildlife events on the planet. The sheer number of animals and the predators that follow them, like lions and crocodiles, make it a top destination for sure.

Then there's the Maasai Mara itself. It's essentially an extension of the Serengeti ecosystem, but it's in Kenya. The Maasai Mara is famous for its big cats Ž013 lions, leopards, and cheetahs. I think the BBC's B̈ig Cat Diaryẅas filmed there. The open savannahs make it easier to spot wildlife, which is a big plus for safari-goers.

Wait, perhaps we could consider it from another angle... Southern Africa has some incredible spots too. Kruger National Park in South Africa comes to mind. It's one of the largest game reserves in Africa and has a high density of wildlife, including the Big Five Ž013 lions, leopards, rhinos, elephants, and buffalo. Plus, it's very accessible with good infrastructure, which might be important for some travelers.

Ah! Now I understand... Botswana's Okavango Delta is another must-mention. It's a unique inland delta where the Okavango River spreads out into the Kalahari Desert, creating a lush, water-rich area teeming with wildlife. The delta's waterways and islands offer a different kind of safari experience, often involving mokoro (dugout canoe) trips. The diversity of habitats there supports a wide range of animals, including large herds of elephants and hippos.

No, no, it should be... Namibia's Etosha National Park also deserves a spot on the list. The park's salt pan is so large it can be seen from space, and it attracts a variety of wildlife, especially during the dry season when animals gather around the waterholes. The stark, otherworldly landscape of Etosha is quite different from the lush environments of other top safari destinations.

Connecting these ideas together... So, if I were to summarize, the top wildlife safari destinations in Africa would likely include the Serengeti and Maasai Mara for the Great Migration and big cats, Kruger National Park for its accessibility and Big Five, the Okavango Delta for its unique water-based safaris, and Etosha National Park for its dramatic landscapes and wildlife concentrations.

Considering real-world implementation... For someone planning a safari, it might come down to what kind of experience they're looking for Ž013 whether it's witnessing the Great Migration, seeing a high density of wildlife, exploring unique landscapes, or having a more accessible and comfortable journey. Each of these destinations offers something special and unforgettable.

Response:
Africa is renowned for its incredible wildlife safaris, offering some of the most picturesque and diverse landscapes along with opportunities to see a wide array of animals in their natural habitats. Here are some of the top wildlife safari destinations on the continent:

1. **Masai Mara National Reserve, Kenya**
 - Known for the Great Migration, where millions of wildebeest, zebras, and gazelles traverse the Mara River.
 - Excellent for spotting the B̈ig Five(̈lion, leopard, rhino, elephant, and Cape buffalo).

2. **Serengeti National Park, Tanzania**
 - Famous for its annual migration and vast open plains.
 - Home to a diverse range of wildlife, including large predator populations like lions, cheetahs, and hyenas.

3. **Kruger National Park, South Africa**
 - One of Africa's largest game reserves.
 - Offers a great chance of seeing the Big Five and a wide variety of other wildlife.
 - Well-developed tourism infrastructure with various accommodations.

4. **Okavango Delta, Botswana**

- A unique inland delta providing lush habitats for a wide array of animals.
- Known for boat safaris, walking safaris, and traditional game drives.

5. **Chobe National Park, Botswana**
- Famous for its large elephant population.
- The Chobe Riverfront provides excellent opportunities to see wildlife drawn to the water.

6. **Etosha National Park, Namibia**
- Features large salt pans and abundant wildlife around waterholes.
- Unique landscapes and opportunities for both daytime and nighttime game viewing.

7. **Bwindi Impenetrable National Park, Uganda**
- Renowned for gorilla trekking experiences.
- Dense, mountainous rainforest habitat offering a chance to see other primates and bird species.

8. **Hwange National Park, Zimbabwe**
- Known for its large herds of elephants and predators such as lions and cheetahs.
- Diverse habitats ranging from woodlands to open plains.

9. **Ngorongoro Crater, Tanzania**
- A UNESCO World Heritage Site with a vast volcanic caldera home to a dense concentration of wildlife.
- Excellent for spotting the Big Five and viewing a variety of bird species.

10. **Rwanda's Volcanoes National Park**
- Famous for mountain gorilla trekking.
- Also home to golden monkeys and other primates.

Each of these destinations offers unique experiences and opportunities to see some of Africa's most iconic wildlife. Whether you're interested in traditional game drives, boat safaris, walking safaris, or specialized treks, Africa's safari destinations provide a rich and diverse array of options for wildlife enthusiasts.

Table 21: Concrete example of *monologue* thinking in ThinkPatterns-21k.

***decomposition* Example**

Instruction:
What challenges did the Bohemian ethnic group face as a minority group in the areas they inhabited?

Thinking:
1. Problem Scoping & Understanding
- The core problem here is understanding the specific challenges faced by the Bohemian ethnic group as a minority. This involves analyzing historical and socio-cultural factors that contributed to their marginalization.
- Key components of this problem include discrimination in various societal sectors, political marginalization, cultural suppression, religious intolerance, and economic deprivation.
- The solution we seek is a comprehensive understanding of these challenges in order to appreciate the broader historical context and its consequences for minority groups.

2. Component Identification
- There are a few essential building blocks to consider:
 a) Discrimination: Identifying sectors like employment, housing, and education where discrimination occurred.
 b) Political marginalization: Understanding exclusion from power and its impacts.
 c) Cultural suppression: Evaluating ways in which cultural aspects were suppressed.
 d) Religious intolerance: Determining the religious dynamics influencing their treatment.
 e) Economic deprivation: Analyzing economic impacts and contributing factors.
- These components are interconnected as each can influence the others, yet they can also be examined individually to build a comprehensive view.

3. Sub-Problem Analysis
- For discrimination, I should look into historical records and testimonies describing treatment in everyday interactions, laws, and societal norms.
- Political marginalization requires exploration of governance structures and representation practices.
- Cultural suppression involves investigating policies aimed at diminishing minority identities and linguistic rights.
- Addressing religious intolerance means understanding the religious landscape of the time and any dominant religious ideologies in opposition to Catholicism.
- Examining economic deprivation involves looking at job market availability, educational barriers, and access to economic resources.

4. Connection Mapping
- The interaction between political marginalization and economic deprivation is crucial as lack of political power can lead to curtailed access to economic resources.
- Discrimination connects with cultural suppression since both are often used to enforce societal hierarchies and assimilation into dominant groups.
- Religious intolerance can compound other factors by justifying social ostracization or influencing other discrimination forms.
- The logical sequence involves understanding cultural and political contexts first as they frame the broader environment for specific discriminations and economic impacts.

5. Integration Strategy
- Solutions from each component need to be synthesized into a coherent narrative that considers the interplay of social, economic, and political factors.
- Potential conflicts might arise in differing interpretations of historical records which need reconciliation through contextual analysis.
- Consistency is ensured by cross-verifying historical accounts and drawing parallels with similar minority challenges in other contexts.

6. Solution Validation
- By covering political, economic, and cultural aspects, and their interconnections, this approach should comprehensively address the challenges faced by the Bohemians.
- Considering historical examples and data helps ensure there are no gaps, while also allowing insights into the broader implications of minority status challenges.
- The components should work together to illustrate a clear picture of the lived realities and systemic issues the Bohemians encountered.

Response:
The Bohemian ethnic group, also known as Czechs, faced various challenges as a minority group in the areas they inhabited, including:

1. Discrimination: The Bohemians often faced discrimination in employment, housing, education, and public accommodations. They were treated as the second-class citizens by the ruling elites who favored the majority groups.

2. Political marginalization: The Bohemians were often excluded from political power and representation, leading to a lack of resources for their communities.

3. Cultural suppression: The Bohemians' language and cultural traditions were often suppressed, and they were forced to assimilate into the majority culture.

4. Religious intolerance: The Bohemians were primarily Catholic, and they faced discrimination and persecution from other religious groups, especially Protestants.

5. Economic deprivation: The Bohemians often faced economic deprivation due to their minority status, which limited their access to education and job opportunities.

These challenges negatively impacted the Bohemians' quality of life and the development of their communities in the areas they inhabited.

Table 22: Concrete example of *decomposition* thinking in ThinkPatterns-21k.

***self-ask* Example**

Instruction:
What are the top wildlife safari destinations in Africa?

Thinking:
Let's begin by considering what we mean by wildlife safari destinations. What do we think are the key elements that make a destination suitable for a wildlife safari?
I think a good wildlife safari destination would have a large variety of animals, especially the big ones like lions, elephants, and rhinos. It should also have good infrastructure for tourists, like lodges and guided tours.

Those are excellent points. We mentioned the variety of animals and infrastructure. Why do we think these elements are important for a wildlife safari experience?

The variety of animals is important because it makes the safari more exciting and educational. The infrastructure is important because it ensures that tourists can have a comfortable and safe experience while exploring the wildlife.

Great. Now, considering these elements, which destinations in Africa come to mind as top wildlife safari locations?

I think of places like the Serengeti in Tanzania, the Maasai Mara in Kenya, and Kruger National Park in South Africa. These are often mentioned as top safari destinations.

Interesting choices. What do we know about the Serengeti, Maasai Mara, and Kruger National Park that supports their reputation as top safari destinations?

The Serengeti is famous for the Great Migration, where millions of wildebeest and other animals move across the plains. The Maasai Mara is known for its large populations of big cats like lions and cheetahs. Kruger National Park is one of the largest game reserves in Africa and has a wide variety of animals and well-developed tourist facilities.

Excellent observations. Let's delve deeper into one of these destinations. How does the Great Migration in the Serengeti contribute to its status as a premier safari destination?

The Great Migration is a unique and spectacular event that attracts many tourists. It involves millions of animals moving in search of water and grazing land, which creates dramatic scenes and opportunities to see predators in action.

Indeed, the Great Migration is a significant draw. How might the presence of predators during the migration enhance the safari experience?

Predators like lions, cheetahs, and crocodiles follow the migration and hunt the animals, which can be thrilling for tourists to witness. It adds an element of excitement and shows the raw, natural behavior of wildlife.

Absolutely. Now, let's consider the Maasai Mara. How does its reputation for big cats contribute to its appeal as a safari destination?

The Maasai Mara's high density of big cats means that tourists have a good chance of seeing lions, cheetahs, and leopards. This makes it a popular destination for those interested in observing these predators up close.

Very true. How does the Maasai Mara's landscape and ecosystem support such a high density of big cats?

The Maasai Mara has vast open plains and abundant prey, which provide ideal hunting grounds for big cats. The ecosystem supports a balanced food chain, ensuring that predators and prey thrive.

Excellent. Now, let's turn our attention to Kruger National Park. What aspects of Kruger make it a top choice for safari-goers?

Kruger National Park is known for its size and diversity of wildlife. It has the Big Five2014lions, leopards, rhinos, elephants, and buffaloes2014and many other species. It also has well-developed infrastructure, including roads, lodges, and guided tours, making it accessible and comfortable for tourists.

Good points. How does the infrastructure in Kruger enhance the overall safari experience?

The infrastructure allows tourists to explore the park safely and comfortably. Guided tours provide expert knowledge about the animals and their behavior, while lodges offer a place to rest and enjoy the surroundings. This makes the experience more enjoyable and educational.

Indeed. Considering all we've discussed, what might be some potential challenges or downsides to these popular safari destinations?

One challenge could be the high number of tourists, which might affect the natural behavior of the animals and the overall experience. There could also be environmental impacts from the infrastructure and human presence.

Very insightful. How might we address these challenges to ensure sustainable and enjoyable safari experiences?

We could implement stricter regulations on the number of tourists and vehicles allowed in the parks. Promoting eco-friendly practices and supporting conservation efforts can help minimize the environmental impact. Educating tourists about responsible behavior can also make a difference.

Excellent suggestions. To summarize, we've identified the Serengeti, Maasai Mara, and Kruger National Park as top wildlife safari destinations due to their diverse wildlife, unique events like the Great Migration, and well-developed infrastructure. We've also considered the importance of sustainable practices to preserve these experiences for the future. How might we apply this understanding to choose a safari destination that aligns with our interests and values?

We can research each destination to see which one offers the experiences we are most interested in, such as witnessing the Great Migration or observing big cats. We can also look for operators and lodges that prioritize sustainability and conservation to ensure our visit has a positive impact.

Excellent. By considering our interests and values, we can make informed choices that enhance our safari experience while supporting the preservation of these incredible destinations. Thank you for this thoughtful exploration.

Thank you for guiding me through this discussion. I've gained a deeper understanding of what makes a top wildlife safari destination and how to choose one responsibly.

Response:
Africa is renowned for its incredible wildlife safaris, offering some of the most picturesque and diverse landscapes along with opportunities to see a wide array of animals in their natural habitats. Here are some of the top wildlife safari destinations on the continent:

1. **Masai Mara National Reserve, Kenya**
 - Known for the Great Migration, where millions of wildebeest, zebras, and gazelles traverse the Mara River.
 - Excellent for spotting the B̈ig Five(̈lion, leopard, rhino, elephant, and Cape buffalo).

2. **Serengeti National Park, Tanzania**
 - Famous for its annual migration and vast open plains.
 - Home to a diverse range of wildlife, including large predator populations like lions, cheetahs, and hyenas.

3. **Kruger National Park, South Africa**
 - One of Africa's largest game reserves.
 - Offers a great chance of seeing the Big Five and a wide variety of other wildlife.
 - Well-developed tourism infrastructure with various accommodations.

4. **Okavango Delta, Botswana**
 - A unique inland delta providing lush habitats for a wide array of animals.
 - Known for boat safaris, walking safaris, and traditional game drives.

5. **Chobe National Park, Botswana**
 - Famous for its large elephant population.
 - The Chobe Riverfront provides excellent opportunities to see wildlife drawn to the water.

6. **Etosha National Park, Namibia**
 - Features large salt pans and abundant wildlife around waterholes.
 - Unique landscapes and opportunities for both daytime and nighttime game viewing.

7. **Bwindi Impenetrable National Park, Uganda**
 - Renowned for gorilla trekking experiences.

- Dense, mountainous rainforest habitat offering a chance to see other primates and bird species.

8. **Hwange National Park, Zimbabwe**
 - Known for its large herds of elephants and predators such as lions and cheetahs.
 - Diverse habitats ranging from woodlands to open plains.

9. **Ngorongoro Crater, Tanzania**
 - A UNESCO World Heritage Site with a vast volcanic caldera home to a dense concentration of wildlife.
 - Excellent for spotting the Big Five and viewing a variety of bird species.

10. **Rwanda's Volcanoes National Park**
 - Famous for mountain gorilla trekking.
 - Also home to golden monkeys and other primates.

Each of these destinations offers unique experiences and opportunities to see some of Africa's most iconic wildlife. Whether you're interested in traditional game drives, boat safaris, walking safaris, or specialized treks, Africa's safari destinations provide a rich and diverse array of options for wildlife enthusiasts.

Table 23: Concrete example of self-ask thinking in ThinkPatterns-21k.

***self-debate* Example**

Instruction:
What are the top wildlife safari destinations in Africa?

Thinking:
Proposition: One of the top wildlife safari destinations in Africa is the Serengeti National Park in Tanzania. The Serengeti is renowned for its annual migration of over 1.5 million wildebeest and 250,000 zebras, making it a unique and unparalleled wildlife spectacle. The park also offers diverse habitats, including savannahs, woodlands, and riverine forests, which support a wide variety of wildlife species such as lions, elephants, and cheetahs.

Opposition: While the Serengeti is indeed a remarkable destination, it is important to consider the potential overcrowding and environmental impact due to its popularity. High tourist numbers can lead to habitat degradation and disturb the natural behavior of wildlife. Additionally, the cost of safaris in the Serengeti can be prohibitive for many travelers, limiting accessibility.

Proposition: To address the concerns of overcrowding and environmental impact, the Serengeti has implemented strict regulations on the number of vehicles allowed in certain areas and has designated specific zones for tourism. These measures help to minimize human impact on wildlife and their habitats. Furthermore, there are various budget-friendly options available, such as group tours and camping safaris, which make the Serengeti more accessible to a wider range of travelers.

Opposition: Another top wildlife safari destination that should be considered is the Okavango Delta in Botswana. The Okavango Delta is a unique inland delta that provides a diverse range of ecosystems, from permanent marshlands to seasonal floodplains. This diversity supports an incredible array of wildlife, including large populations of elephants, hippos, and crocodiles, as well as numerous bird species. The delta's remote location also means that it experiences less tourist traffic, allowing for a more intimate and exclusive safari experience.

Proposition: The Okavango Delta is indeed a fantastic destination, but its remoteness can also pose challenges. Accessing the delta often requires charter flights, which can be expensive and logistically complex. Additionally, the seasonal flooding can limit accessibility to certain areas and affect the availability of wildlife sightings.

Opposition: To mitigate these challenges, many safari operators in the Okavango Delta offer all-inclusive packages that cover transportation, accommodation, and guided tours, simplifying the logistics for travelers. The seasonal flooding, while a potential limitation, also creates unique opportunities for water-based safaris, such as mokoro (dugout canoe) excursions, which provide a different perspective on wildlife viewing.

Proposition: Another noteworthy destination is Kruger National Park in South Africa. Kruger is one of Africa's largest game reserves and offers an extensive network of roads and accommodations, catering to both self-drive and guided safari experiences. The park is home to the Big Five (lion, leopard, rhinoceros, elephant, and Cape buffalo) and boasts a high density of wildlife, making it an excellent choice for first-time safari-goers.

Opposition: While Kruger National Park is highly accessible and offers a range of amenities, its popularity can lead to crowded conditions, especially in peak seasons. This can detract from the wilderness experience and increase the risk of human-wildlife conflicts. Additionally, the park's extensive infrastructure can sometimes give it a less pristine feel compared to more remote destinations.

Proposition: To enhance the wilderness experience in Kruger, visitors can opt for private game reserves adjacent to the park, such as Sabi Sands or Timbavati. These reserves offer exclusive access to wildlife-rich areas and often provide more personalized and intimate safari experiences. Furthermore, Kruger's well-developed infrastructure ensures that conservation efforts are well-funded and effective, contributing to the long-term preservation of its diverse ecosystems.

Opposition: Considering all the points discussed, it is clear that each destination has its unique strengths and challenges. To synthesize an improved solution, we should recommend a combination of destinations based on travelers' preferences and priorities. For those seeking iconic wildlife spectacles and diverse habitats, the Serengeti is an excellent choice. For a more exclusive and intimate experience, the Okavango Delta is ideal. For accessibility and comprehensive safari options, Kruger National Park stands out.

Proposition: Agreed. By recognizing the distinct advantages of each destination and considering travelers' individual needs, we can provide a more tailored and enriching safari experience. It is also important to emphasize sustainable tourism practices across all destinations to ensure the long-term preservation of Africa's incredible wildlife and natural landscapes.

Opposition: In conclusion, the top wildlife safari destinations in Africa include the Serengeti National Park, the Okavango Delta, and Kruger National Park. Each offers unique experiences and opportunities for wildlife viewing. By promoting sustainable tourism and considering travelers' preferences, we can enhance the overall safari experience while ensuring the protection of these precious ecosystems.

Proposition: Well said. By combining our insights and focusing on sustainable practices, we can help travelers make informed choices and contribute to the conservation of Africa's wildlife heritage.

Response:
Africa is renowned for its incredible wildlife safaris, offering some of the most picturesque and diverse landscapes along with opportunities to see a wide array of animals in their natural habitats. Here are some of the top wildlife safari destinations on the continent:

1. **Masai Mara National Reserve, Kenya**
 - Known for the Great Migration, where millions of wildebeest, zebras, and gazelles traverse the Mara River.
 - Excellent for spotting the B̈ig Five(̈lion, leopard, rhino, elephant, and Cape buffalo).

2. **Serengeti National Park, Tanzania**
 - Famous for its annual migration and vast open plains.
 - Home to a diverse range of wildlife, including large predator populations like lions, cheetahs, and hyenas.

3. **Kruger National Park, South Africa**
 - One of Africa's largest game reserves.
 - Offers a great chance of seeing the Big Five and a wide variety of other wildlife.
 - Well-developed tourism infrastructure with various accommodations.

4. **Okavango Delta, Botswana**
 - A unique inland delta providing lush habitats for a wide array of animals.
 - Known for boat safaris, walking safaris, and traditional game drives.

5. **Chobe National Park, Botswana**
 - Famous for its large elephant population.
 - The Chobe Riverfront provides excellent opportunities to see wildlife drawn to the water.

6. **Etosha National Park, Namibia**
 - Features large salt pans and abundant wildlife around waterholes.
 - Unique landscapes and opportunities for both daytime and nighttime game viewing.

7. **Bwindi Impenetrable National Park, Uganda**
 - Renowned for gorilla trekking experiences.
 - Dense, mountainous rainforest habitat offering a chance to see other primates and bird species.

8. **Hwange National Park, Zimbabwe**
 - Known for its large herds of elephants and predators such as lions and cheetahs.
 - Diverse habitats ranging from woodlands to open plains.

9. **Ngorongoro Crater, Tanzania**
 - A UNESCO World Heritage Site with a vast volcanic caldera home to a dense concentration of wildlife.
 - Excellent for spotting the Big Five and viewing a variety of bird species.

10. **Rwanda's Volcanoes National Park**
 - Famous for mountain gorilla trekking.
 - Also home to golden monkeys and other primates.

Each of these destinations offers unique experiences and opportunities to see some of Africa's most iconic wildlife. Whether you're interested in traditional game drives, boat safaris, walking safaris, or specialized treks, Africa's safari destinations provide a rich and diverse array of options for wildlife enthusiasts.

Table 24: Concrete example of self-debate thinking in ThinkPatterns-21k.

***Self-Critic* Example**

Instruction:
What are the top wildlife safari destinations in Africa?

Thinking:
Draft Response:
Africa is renowned for its incredible wildlife and offers numerous safari destinations where visitors can experience the continent's natural beauty and diverse animal species. Here are some of the top wildlife safari destinations in Africa:

1. Serengeti National Park, Tanzania: Famous for the Great Migration, where millions of wildebeest, zebras, and gazelles migrate in search of fresh grazing. The park is also home to the Big Five (lion, leopard, elephant, buffalo, and rhino).

2. Maasai Mara National Reserve, Kenya: Known for its abundant wildlife and the annual migration of wildebeest and zebras. The reserve offers excellent opportunities to see the Big Five and other species like cheetahs and hyenas.

3. Kruger National Park, South Africa: One of Africa's largest game reserves, Kruger is home to a vast array of wildlife, including the Big Five. The park offers a range of accommodation options and well-maintained infrastructure for self-drive safaris.

4. Chobe National Park, Botswana: Famous for its large elephant population, Chobe also offers boat safaris on the Chobe River, where visitors can see hippos, crocodiles, and a variety of bird species.

5. Etosha National Park, Namibia: Known for its vast salt pan, which attracts a diverse range of wildlife, including elephants, lions, and rhinos. The park's waterholes provide excellent game-viewing opportunities.

6. Hwange National Park, Zimbabwe: Home to one of the largest elephant populations in Africa, Hwange also boasts a rich diversity of other wildlife, including lions, leopards, and wild dogs.

7. Okavango Delta, Botswana: A unique inland delta that supports a rich variety of wildlife. The delta's waterways are ideal for mokoro (dugout canoe) safaris, offering a different perspective on the animals and their habitats.

8. Ngorongoro Crater, Tanzania: A UNESCO World Heritage Site, the crater is a natural enclosure for a wide variety of wildlife, including the Big Five. The crater's floor is teeming with animals, making it a prime safari destination.

Critical Comments:
- The draft response provides a good overview of some of the top wildlife safari destinations in Africa, but it could benefit from more detailed descriptions of each location.
- The response could include information about the best times to visit each destination to enhance the reader's planning.
- Mentioning specific activities or unique experiences available at each destination would add value.
- Including a brief note on conservation efforts or the importance of sustainable tourism in these areas could provide a more comprehensive perspective.
- The response could be improved by adding a conclusion that summarizes the key points and emphasizes the diversity of safari experiences available in Africa.

Suggestions for Improvement:
1. Add more detailed descriptions of each destination, highlighting unique features and specific wildlife species that can be seen.
2. Include information about the best times to visit each destination to help readers plan their trips.
3. Mention specific activities or unique experiences available at each location, such as hot air balloon rides, walking safaris, or night drives.
4. Incorporate a brief discussion on conservation efforts and the importance of sustainable tourism in these areas.
5. Conclude with a summary that emphasizes the diversity and richness of safari experiences in Africa.

Response:
Africa is renowned for its incredible wildlife safaris, offering some of the most picturesque and diverse landscapes along with opportunities to see a wide array of animals in their natural habitats. Here are some of the top wildlife safari destinations on the continent:

1. **Masai Mara National Reserve, Kenya**

- Known for the Great Migration, where millions of wildebeest, zebras, and gazelles traverse the Mara River.
- Excellent for spotting the B̈ig Five(̈lion, leopard, rhino, elephant, and Cape buffalo).

2. **Serengeti National Park, Tanzania**
- Famous for its annual migration and vast open plains.
- Home to a diverse range of wildlife, including large predator populations like lions, cheetahs, and hyenas.

3. **Kruger National Park, South Africa**
- One of Africa's largest game reserves.
- Offers a great chance of seeing the Big Five and a wide variety of other wildlife.
- Well-developed tourism infrastructure with various accommodations.

4. **Okavango Delta, Botswana**
- A unique inland delta providing lush habitats for a wide array of animals.
- Known for boat safaris, walking safaris, and traditional game drives.

5. **Chobe National Park, Botswana**
- Famous for its large elephant population.
- The Chobe Riverfront provides excellent opportunities to see wildlife drawn to the water.

6. **Etosha National Park, Namibia**
- Features large salt pans and abundant wildlife around waterholes.
- Unique landscapes and opportunities for both daytime and nighttime game viewing.

7. **Bwindi Impenetrable National Park, Uganda**
- Renowned for gorilla trekking experiences.
- Dense, mountainous rainforest habitat offering a chance to see other primates and bird species.

8. **Hwange National Park, Zimbabwe**
- Known for its large herds of elephants and predators such as lions and cheetahs.
- Diverse habitats ranging from woodlands to open plains.

9. **Ngorongoro Crater, Tanzania**
- A UNESCO World Heritage Site with a vast volcanic caldera home to a dense concentration of wildlife.
- Excellent for spotting the Big Five and viewing a variety of bird species.

10. **Rwanda's Volcanoes National Park**
- Famous for mountain gorilla trekking.
- Also home to golden monkeys and other primates.

Each of these destinations offers unique experiences and opportunities to see some of Africa's most iconic wildlife. Whether you're interested in traditional game drives, boat safaris, walking safaris, or specialized treks, Africa's safari destinations provide a rich and diverse array of options for wildlife enthusiasts.

Table 25: Concrete example of self-critic thinking in ThinkPatterns-21k.