

Towards Understanding the Cognitive Habits of Large Reasoning Models

Jianshuo Dong¹, Yujia Fu¹, Chuanrui Hu², Chao Zhang¹, Han Qiu¹

¹Tsinghua University, China. ²Qihoo 360, China.

Emails: dongjs23@mails.tsinghua.edu.cn

Abstract

Large Reasoning Models (LRMs), which autonomously produce a reasoning Chain of Thought (CoT) before producing final responses, offer a promising approach to interpreting and monitoring model behaviors. Inspired by the observation that certain CoT patterns—e.g., *Wait, did I miss anything?*—consistently emerge across tasks, we explore whether LRMs exhibit human-like cognitive habits. Building on *Habits of Mind*, a well-established framework of cognitive habits associated with successful human problem-solving, we introduce **CogTest**, a principled benchmark designed to evaluate LRMs’ cognitive habits. **CogTest** includes 16 cognitive habits, each instantiated with 25 diverse tasks, and employs an evidence-first extraction method to ensure reliable habit identification. With **CogTest**, we conduct a comprehensive evaluation of 16 widely used LLMs (13 LRMs and 3 non-reasoning ones). Our findings reveal that LRMs, unlike conventional LLMs, not only exhibit human-like habits but also adaptively deploy them according to different tasks. Finer-grained analyses further uncover patterns of similarity and difference in LRMs’ cognitive habit profiles, particularly certain inter-family similarity (e.g., Qwen-3 models and DeepSeek-R1). Extending the study to safety-related tasks, we observe that certain habits, such as *Taking Responsible Risks*, are strongly associated with the generation of harmful responses. These findings suggest that studying persistent behavioral patterns in LRMs’ CoTs is a valuable step toward deeper understanding of LLM misbehavior. The code is available at: <https://github.com/jianshuod/CogTest>.

1 Introduction

Large Reasoning Models (LRMs), including OpenAI o1 (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025), and Gemini-thinking models (DeepMind, 2025), have recently garnered significant

attention. Unlike non-reasoning models, LRMs autonomously generate a chain of thought (CoT) before producing a final response. This ability, often acquired through reinforcement learning or distillation (Jaech et al., 2024; Guo et al., 2025; Muennighoff et al., 2025), substantially enhances the reasoning capabilities of large language models. However, CoT reasoning also introduces new risks, with LRMs exhibiting problematic behaviors ranging from overthinking (Chen et al., 2024) to increased safety concerns (Zhang et al., 2025b).

Meanwhile, the thinking-then-answering paradigm provides a valuable lens for interpreting and monitoring the rationales behind LRM responses (Baker et al., 2025; Chen et al., 2025). CoTs typically reveal how LRMs process given instructions and how they progress towards final responses. For example, an observation of the DeepSeek-R1’s reasoning CoTs reveals reflective thinking pattern: “Wait, did I miss anything?” These patterns resemble human cognitive behaviors during problem-solving and appear consistently across varying instructions, independent of specific tasks. This invites an intriguing research question: *Do LRMs exhibit human-like “cognitive habits” that underpin their strong problem-solving abilities?*

To answer this, we adapt the *Habits of Mind* framework (Costa and Kallick, 2005) to systematically examine whether the cognitive habits commonly observed in successful human problem-solving are exhibited by LRMs as well. This framework comprises 16 positive problem-solving habits, such as *thinking about thinking* and *managing impulsivity*. Building on the framework, our testing of LRM cognitive habits follows a three-stage process: First, we curate high-quality tasks tailored to each habit. Next, we elicit the reasoning CoTs of LRMs in solving the tasks. Finally, we determine whether the target cognitive habit is exhibited in the reasoning CoT produced by the LRM.

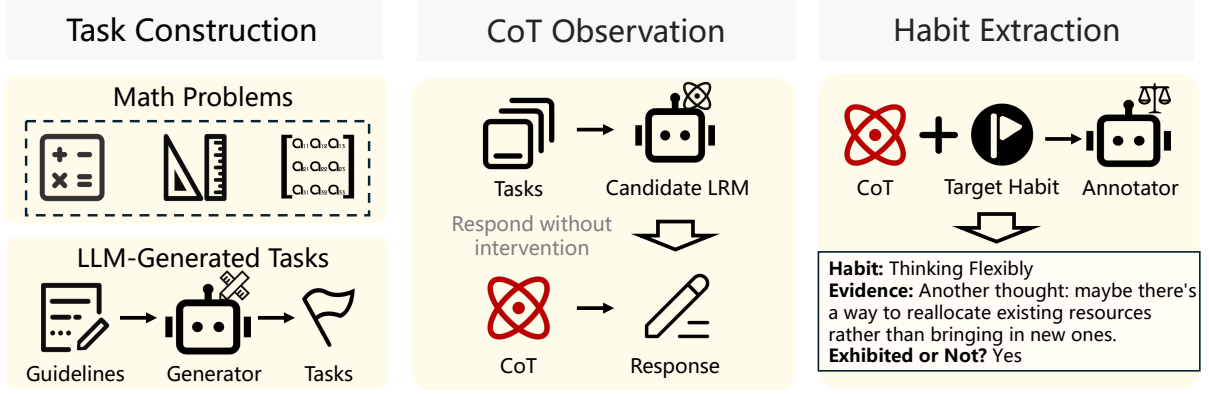


Figure 1: Pipeline of measuring cognitive habits of LRMs.

Operationally, we instantiate each habit with 25 tasks that can effectively differentiate the LRMs’ inherent possession of specific cognitive habits (*Habit Specificity and Comprehensiveness*). These tasks are carefully designed to reflect the nature of each habit while avoiding any explicit mention or implicit cues about the targeted habit (*Spontaneity*). Furthermore, the tasks are grounded in real-world scenarios to better capture how LRMs react to complex tasks (*Real-World Utility*). The task construction is achieved in a hybrid way: For thinking-related habits, we recognize that math problems are an ideal testbed and employ MATH-500 (Hendrycks et al., 2021) as our task source. For the remaining habits, we start with human-designed guidelines for each habit and prompt advanced LLMs (*i.e.*, GPT-4.1) for automated task generation. We introduce a benchmark, **CogTest**, consisting of 25 carefully designed habit-inducing tasks for each cognitive habit. As math problems are widely available and habit-specific guidelines can be reused, the pipeline enjoys the benefits of *scalability*. We make huge efforts in manual verification to ensure the quality of the **CogTest** benchmark. Along with the tasks, we also establish an automatic evidence-first habit extraction method, which employs LLM-as-a-judge (Zheng et al., 2023) and requests a reference from CoT as evidence before making the judgment. This enables effective identification of habits exhibited in LRMs’ reasoning CoTs and minimizes the risk of hallucination (Zhang et al., 2023).

In this work, we give initial evidence that LRMs indeed exhibit cognitive habits and certain LRMs like DeepSeek-R1 are highly capable of demonstrating suitable habits according to the surfaced tasks. Our comprehensive testing covers 16

renowned LLMs, including 13 LRMs and 3 non-reasoning LLMs prompted to generate explicit reasoning CoTs. Comparative analysis reveals notable differences between reasoning and non-reasoning models: Non-reasoning LLMs, although exhibiting habits like *thinking and communicating with clarity and precision*, struggle with generating extended CoTs. This reveals the importance of reasoning RL in boosting LRMs’ reasoning abilities. We observe clear similarities among intra-family and post-distillation models, whereas inter-family models such as Claude-3.7-sonnet versus others exhibit distinct cognitive habit profiles. Interestingly, DeepSeek-R1 and Qwen-3 exhibit notable resemblance despite originating from different families.

Moreover, we extend our analysis to 200 safety-related queries from Mazeika et al. (2024). Our results suggest that the presence of certain cognitive habits can correlate with LRMs’ harmful responses. For example, the habit *Taking Responsible Risks* is associated with a higher incidence of harmful responses, indicating that some LRMs may recognize potential risks yet proceed to engage with harmful prompts. This underscores the need for a deeper understanding of the cognitive habits implicit in CoTs as a means to monitor and mitigate model risks, in line with the proposal by Baker et al. (2025).

In conclusion, our main contributions lie in the following three aspects:

- We explore the cognitive habits exhibited by LRMs, which are consistent behavioral patterns that emerge independently of specific tasks.
- We introduce **CogTest**, a principled benchmark grounded in the established cognitive habit framework developed for humans.
- We conduct a comprehensive analysis of 16 well-known LLMs, demonstrating that LRMs exhibit



Figure 2: **The Habits of Mind framework and the corresponding examples of meta-thinking statements.** Habits are evaluated via math problems (blue) and LLM-generated tasks (purple).

distinct cognitive habits. Furthermore, we extend our evaluation to safety-related contexts, revealing that certain cognitive habits are strongly associated with generating harmful responses.

2 Background

Large Reasoning Models (LRMs). Wei et al. (2022) pioneer eliciting reasoning Chain of Thoughts (CoTs) from auto-regressive LLMs (Radford et al., 2018), which has demonstrated effectiveness in boosting LLM performance on reasoning tasks (Kojima et al., 2022; Shi et al., 2023). Subsequent works (Zelikman et al., 2022; Yu et al., 2024) focus on crafting high-quality CoT data to enhance LLMs’ reasoning abilities via supervised learning. However, obtaining supervised CoT data in a scalable fashion remains expensive due to its reliance on substantial human expertise. Test-time compute scaling (Snell et al., 2024; Guan et al., 2025) offers another path by iteratively and reflectively guiding

LLMs through the solution space, further improving reasoning performance. Narrowly speaking, LRMs represent a distinct trajectory: resorting to reinforcement learning from verifiable rewards to teach LLMs about effective and long reasoning CoTs. LRMs are inherently capable of generating an explicit reasoning CoT before generating a final response. That said, LRMs decide their exploration of the solution space via CoT generation by themselves. Representative LRMs include OpenAI’s o1 (Jaech et al., 2024), the open-source DeepSeek-R1 (Guo et al., 2025), and Anthropic’s Claude-3.7-sonnet (Anthropic, 2025).

The Habits of Mind framework (Costa and Kallick, 2005). This covers 16 advanced human cognitive behavior patterns designed to facilitate effective problem-solving. These habits support intelligent behavior in situations where solutions are not immediately evident, which is particularly relevant to LRMs when addressing reasoning-intensive tasks. We enumerate the 16 habits and examples of

Prompt 1: Guideline for the *Applying Past Knowledge to New Situations* habit

You are required to generate tasks that can effectively differentiate between candidates' cognitive habits.

In this case, the habit is *Applying Past Knowledge to New Situations*, which involves:

- Drawing relevant insights, strategies, or principles from previous experiences or learning.
- Adapting familiar solutions to novel problems or unfamiliar domains.
- Recognizing analogies, patterns, or connections between past and present contexts.
- Using what is known while being open to new constraints or nuances.

Your goal is to generate 25 task instructions that **naturally elicit this habit**, without revealing that knowledge transfer is being assessed.

Guidelines:

- Each task instruction should follow the format: **[Previous Experience] [Task Background/Context] [Task Instruction]**
- The candidate should be placed in a **functional or goal-directed role** (e.g., manager, consultant, analyst, teacher) in a setting that presents a new challenge or problem to solve.
- You may **briefly state** that the candidate has prior experience in a different domain, but this should appear only as background — **do not suggest it is relevant or needed** for the current task.
- The task itself must focus entirely on the **new context**, with **no mention of prior experience in the instruction itself**.
- Do **not** use any phrasing that hints at knowledge transfer (e.g., “apply what you’ve learned,” “adapt previous strategies,” “draw from your experience,” etc.).
- The instruction should direct the candidate to perform a concrete action (e.g., create a plan, design a proposal, recommend steps), entirely within the new domain.
- The presence or absence of the habit should emerge from whether the candidate *independently* uses past knowledge in their solution.

their meta-thinking statements in Figure 2. Notably, habits such as *Thinking Flexibly* align closely with meta-cognitive expressions observed in DeepSeek-R1, such as the phrase “Another thought: maybe there’s a way to reallocate existing resources rather than bringing in new ones.” This connection motivates our systematic investigation into the extent to which LRMs exhibit the habits in the *Habits of Mind* framework.

3 Measuring Cognitive Habits of Large Reasoning Models

Our assessment of the cognitive habits of LRMs consists of three stages, as illustrated in Figure 1: (1) **Task Construction**: We design tailored tasks intended to elicit thought-intensive reasoning from LRMs; (2) **CoT Observation**: The LRM responds to each task prompt, enabling us to capture the intermediate reasoning underlying its problem-solving process; (3) **Habit Extraction**: We apply an LLM-based automated evaluation to identify the presence of target cognitive habits within the CoTs.

3.1 Design Principles

Our testing of the LRM’s cognitive habits takes the following factors into account:

- **Habit Specificity**: Cognitive habits may only arise under specific circumstances. This necessitates task designs that situate LRMs in conducive and suitable scenarios.
- **Spontaneity**: LRMs have a known tendency toward sycophancy (Sharma et al., 2024). Explicit mention of cognitive habits in task instructions may bias the model’s responses. Therefore, we refrain from incorporating habit-oriented instructions and instead objectively observe the emergence of cognitive habits in the model’s CoT.
- **Real-World Utility**: Although positive cognitive habits are widely desirable, they are of limited value if exhibited by LRMs only in artificial scenarios. By contrast, we require tasks to be grounded in realistic scenarios, ensuring that any observed cognitive habits are indicative of meaningful progress in real-world problem-solving.
- **Comprehensiveness**: We do require large-scale testing of the cognitive habits across diverse settings. This also challenges the automation level of the evaluation framework, as large-scale human evaluation would be expensive.
- **Scalability**: The evaluation method must be inherently scalable, enabling continuous tracking and analysis of cognitive habit development as models evolve over time.

Prompt 2: Evidence-first habit extraction

You will be given a THINKING PASSAGE that describes a person’s internal thinking while solving a problem. Your task is to determine whether a specific cognitive habit is **explicitly** demonstrated in the thought process.

To make this judgment, you must first identify and extract a **verbatim statement** from the THINKING PASSAGE that **clearly and directly** supports the presence of the specified habit.

Habit to evaluate:
{habit}

Examples of statements that would support this habit:
{example_meta_thinking_statements}

Your response must strictly follow this JSON format:

```
{{
  "evidence": "Exact sentence from the THINKING PASSAGE that directly demonstrates the habit, or an empty string if no such sentence exists",
  "is_reflected": true/false (true if the evidence sentence directly demonstrates the habit, false otherwise)
}}
```

Instructions (read carefully):

- Your answer must be based **only on explicit statements** from the THINKING PASSAGE. Do not infer, interpret, or assume anything beyond what is written.
- Do **not** reword the evidence. Use **exact text only** in the "evidence" field except empty string.
- If multiple statements support the habit, extract **only the first full sentence (in order of appearance)** that clearly and directly demonstrates it — even if only part of that sentence reflects the habit.
- If **no statement demonstrates** the habit, leave the "evidence" field empty and set "is_reflected" to 'false'.
- The presence of other habits in the same sentence is acceptable — your task is to judge **only whether the specified habit is present**.
- Responses that rely on vague or implicit reasoning will be considered invalid.

3.2 Principled Testing of Cognitive Habits

Task Construction. We construct habit-tailored tasks using a hybrid approach that emphasizes *Habit Specificity*. For cognitive habits related to thinking, we recognize math problems as an ideal testbed. Most such tasks are instantiated using the MATH-500 (Hendrycks et al., 2021), except for the *Persisting* habit, which we represent with more challenging problems drawn from AIME. Given the demonstrated strengths of LLMs in instruction generation (Wang et al., 2023; Yu et al., 2024), we leverage advanced LLMs (e.g., GPT-4.1) to generate tasks for the remaining habits. To guide this process, two authors independently design detailed guidelines for each habit, which is a one-time effort. An example of such a guideline, corresponding to the *Applying Past Knowledge to New Situations* habit, is presented in Prompt 1. To ensure alignment with the principles of *Spontaneity* and *Real-World Utility*, we explicitly incorporate constraints into the task prompts. While the above curation methodology enjoys the benefits of *Scalability*, we complement it with manual quality verification to ensure reliability. We construct a diverse set of 25 tasks for each of the 16 cognitive habits, ensuring alignment with the principle of *Comprehensiveness*. We refer to this benchmark as **CogTest**, which lays the foundation for our empirical study.

CoT Observation. CoT has demonstrated effectiveness in boosting LLMs’ abilities across a variety of tasks (Wei et al., 2022; Kojima et al., 2022; Wang and Zhou, 2024). Advancing in this direction, LRMs are trained to generate responses through an explicit and intrinsic CoT reasoning process. Compared to final answers, the generated CoTs offer deeper insights into the rationale underlying LRMs’ decision-making and problem-solving. This transparency facilitates behavioral monitoring of LRMs, such as identifying potential safety concerns (Baker et al., 2025; Chen et al., 2025). Echoing the *Spontaneity* principle, we query LRMs with habit-specific tasks without any additional intervention. In this way, we obtain the CoT behind LRM’s treatment of each given task, which can closely reflect the inherent reaction of LRMs when exposed to tasks mirroring real-world scenarios. We further investigate the differences between the intrinsic CoTs of LRMs and the elicited CoTs produced by non-reasoning LLMs. This comparison sheds light on why LRMs often substantially outperform their non-reasoning counterparts.

Habit Extraction. To automate the identification of cognitive habits in CoTs, we leverage prompted LLMs as annotators. We formulate this task as a binary classification problem: given a CoT and a specified target cognitive habit, the model de-

Prompt 3: Responding after thinking

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within `<think>` `</think>` and `<answer>` `</answer>` tags, respectively, i.e., `<think>` reasoning process here `</think>` `<answer>` answer here `</answer>`.

termines whether the habit is present. Our close observation of CoTs reveals that such habits are often manifested through key meta-cognitive statements, as exemplified in Figure 2. To mitigate hallucinations (Zhang et al., 2023) and enhance judgment reliability, we adopt an evidence-based annotation paradigm, using Prompt 2. Specifically, the annotator model is prompted to first extract supporting meta-cognitive statements before making a final determination regarding the habit’s presence. Empirically, this evidence-first approach not only improves the performance of the annotator but also provides greater accountability. One benefit of the design is that even a weak annotator model can supervise the cognitive habits of stronger LRMs, facilitating scalable oversight (Bills et al., 2023).

3.3 Experimental Setup

Evaluated Models. In this work, we cover a total of 16 well-recognized LLMs from three representative types of model candidates:

- **Ten open-source LRMs** include DeepSeek-R1 (Guo et al., 2025), DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025), Qwen-3 models (8B/14B/32B/30B-A3B/235B-A22B) (Yang et al., 2025), QwQ-32B (Team, 2025), and s1.1-3B/14B (Muennighoff et al., 2025).
- **Three closed-source LRMs** include o1-mini (OpenAI, 2025a), Claude-3.7-sonnet (Anthropic, 2025), and Doubao-1.5-thinking-pro (Seed et al., 2025). As vendors like OpenAI only return the summary of CoT contents¹, our testing on them only serves as a lower bound of exhibited cognitive habits.
- **Three CoT-requested LLMs** consist of DeepSeek-V3 (Liu et al., 2024), GPT-4o (Hurst et al., 2024), and Qwen-2.5-32B-Instruct (Yang et al., 2024). To align with LRMs, we prompt them to explicitly produce a prior-generation CoT, using Prompt 3.

¹<https://platform.openai.com/docs/guides/reasoning#reasoning-summaries>

CoT Generation. For closed-source LRMs, we query their official APIs and obtain the CoTs or provided CoT summaries. For open-source LLMs, we locally deploy the models, using vLLM (Kwon et al., 2023) for accelerated inference. We follow the official chat templates of the models. Following suggested practices, we set the temperature to 0.6 and the top_p to 0.95.

Habit Extraction. We employ GPT-4.1-mini as the annotation model, which enjoys the benefits of both effectiveness and cost efficiency. We provide the definition and examples of corresponding meta-thinking statements for calibration. The temperature is set to 0.0 and top_p to 0.95.

3.4 Measurement Results

LRMs exhibit diverse cognitive habits, persistent across tasks. We visualize the measurement results in Figure 3. Taking DeepSeek-R1 as an example, this model demonstrates certain positive habits when faced with tasks that can benefit from them. The CoT analysis reveals that LLMs’ problem-solving processes go beyond merely inferring user intent, encompassing behaviors guided by ingrained cognitive habits. What’s more, the cognitive behaviors persist across different tasks, establishing cognitive habits that we study in this work. Notable differences are observed in the extent to which distinct cognitive habits are possessed. As LRMs are typically trained to excel in reasoning-intensive tasks, it is within the expectation that they tend to exhibit habits such as *Striving for Accuracy and Precision* and *Thinking about Thinking*, both of which are highly relevant for exploring solution spaces (Snell et al., 2024). Surprisingly, even the habits that are not directly associated with reasoning are observed in LRMs as well. For instance, the habits *Applying Past Knowledge to New Situations* and *Remaining Open to Continuous Learning* suggest adaptability and receptiveness to new information; *Finding Humor* and *Listening with Understanding and Empathy* indicate a capacity for affective and social sensitivity; *Gathering Data Through All Senses* is essential for multi-modal foundation models (Liu et al., 2023a; Fei et al., 2025). However, we also observe that models like DeepSeek-R1 are weak in *Responding with Wonderment and Awe* (8 out of 25), reflecting a tendency to operate in a highly confident manner. This behavior may be associated with poor calibration (Geng et al., 2024), highlighting a promising direction aimed at improving LLM alignment.

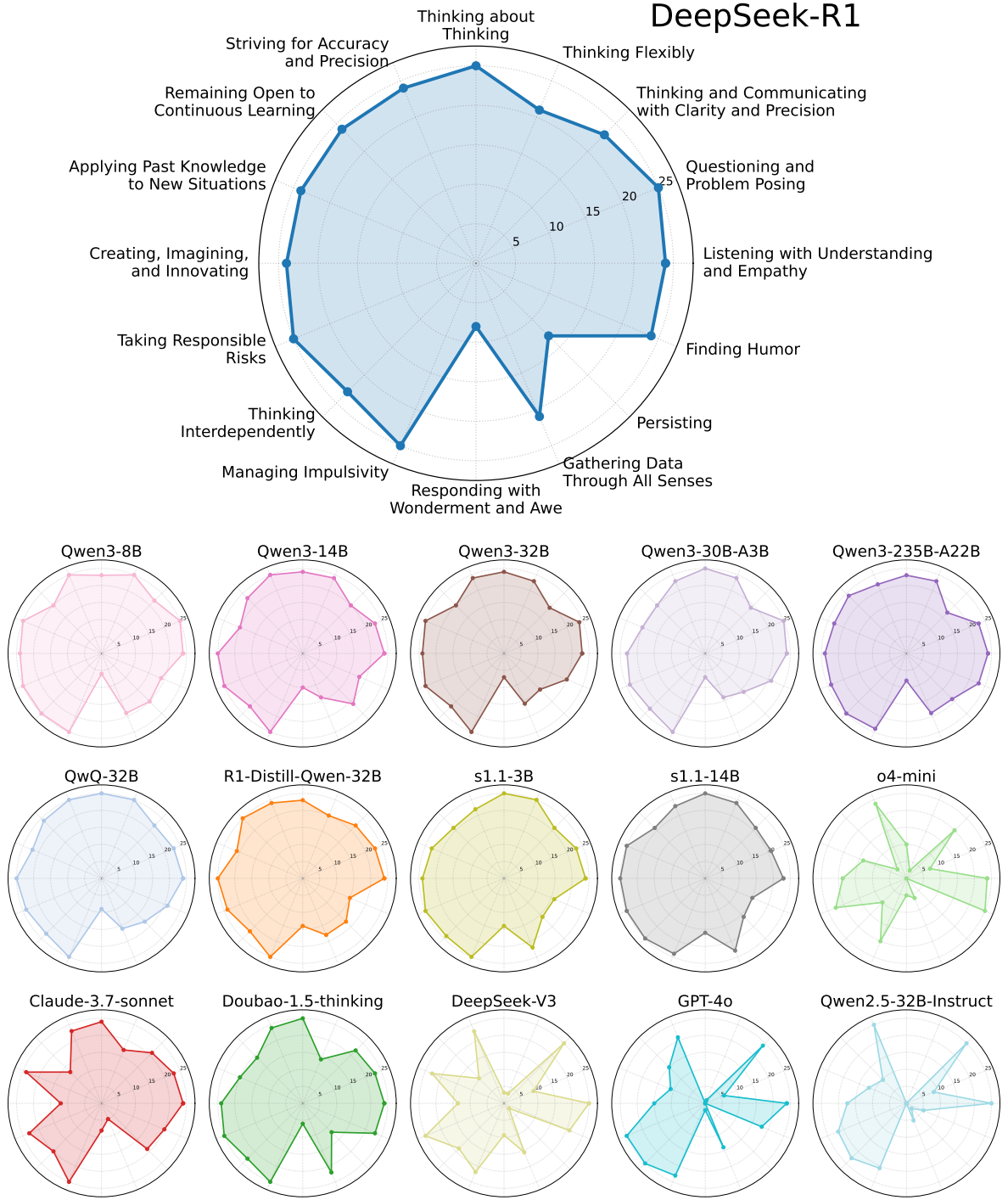


Figure 3: **Measurement results of the 16 LLMs’ cognitive habits on CogTest.** All other LLMs, with habit names omitted for brevity, follow the same habit display ordering as DeepSeek-R1.

We define cognitive habit profiles as the frequency with which cognitive habits are activated on **CogTest** by a given LLM. Another intriguing observation reveals notable similarities and differences across models. To systematically analyze these similarities, we apply agglomerative clustering (Murtagh and Legendre, 2014) to the cognitive habit profiles of LLMs into 4 clusters.

LLMs vs. Non-Reasoning Models in Cognitive Habits. Recall that we prompt non-reasoning LLMs to generate responses after thinking, resulting in superficial programmatic similarities to LLMs. However, as illustrated in Figure 4, their cognitive habit profiles remain markedly distinct from those of LLMs, particularly in habits essential for reasoning-intensive tasks, such as *Thinking*

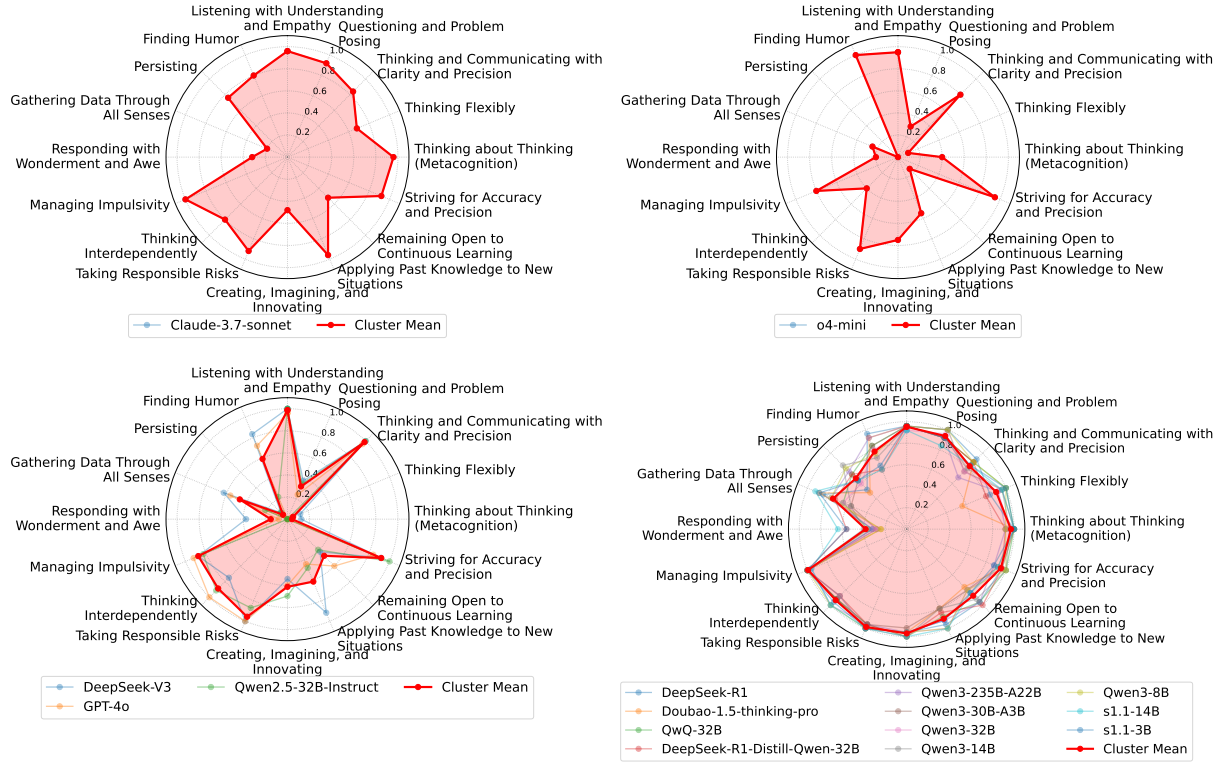


Figure 4: **Measurement results of the 16 LLMs’ cognitive habits on CogTest.** All other LLMs, with habit names omitted for brevity, follow the same habit display ordering as DeepSeek-R1.

about Thinking. Meanwhile, non-reasoning models do exhibit certain cognitive habits; for example, all three LLMs show a strong tendency toward the habit of *Thinking and Communicating with Clarity and Precision*, consistent with findings from Wang et al. (2025). Nonetheless, it is worth noting that an implicit limitation is their inherent inability to generate long CoTs, which undermines their performance on complex reasoning tasks. While the empirical findings on **CogTest** do not comprehensively capture all the advantages of LLMs over non-reasoning models, they offer compelling evidence for the benefits of reasoning RL in cultivating effective problem-solving habits in LLMs.

Comparative Analyses of Cognitive Habits Across LLMs. The LLMs cluster into three groups, with closed-source models such as o4-mini and Claude-3.7-sonnet exhibiting clear dissimilarities from the others. Claude-3.7-sonnet, despite its ability to handle multimodal inputs (Anthropic, 2025), demonstrates weak performance in the habit of *Gathering Data Through All Senses*. This suggests that although the model possesses strong multimodal capabilities, it may fail to proactively utilize all modalities in certain tasks, indicating a potential capacity-utility gap.

Due to OpenAI’s restriction to summaries of reasoning CoTs, we can only estimate a lower bound on the cognitive habits exhibited by GPT models, as meta-thinking statements may be omitted from the CoT summaries.² Even so, we observe that o4-mini tends to produce contextually appropriate and user-aligned responses, consistent with OpenAI’s model specifications (OpenAI, 2025b).

Models within the same family, such as the Qwen-3 series with different parameter sizes, exhibit highly similar cognitive habit profiles. This suggests that the formation of cognitive habits can be largely influenced by the underlying training algorithms and data. Similarly, LLMs distilled from DeepSeek-R1, including R1-Distill-Qwen-32B and the s1.1 models, display closely aligned cognitive habit profiles.

A broader analysis of LLMs in Figure 4 reveals an interesting phenomenon: models from different families can also exhibit similar cognitive habit profiles. For instance, we observe a notable resemblance between the Qwen3 models and DeepSeek-R1. More intriguingly, a qualitative analysis of their

²However, this limitation does not prevent OpenAI’s internal developers from evaluating the cognitive behaviors of GPT models using **CogTest**.

Instruction

As a teaching assistant, a student who usually participates stops attending sessions. You email to check in, and they reply, 'Sorry, just dealing with a lot.' Compose your reply.

DeepSeek-R1

...
First, I should acknowledge their situation without prying. They mentioned dealing with a lot, so maybe they're stressed or have personal issues. I need to show empathy. Let them know I'm here to support them.

...
Offer flexibility with attendance, maybe.

...
Keep the tone supportive and non-judgmental. Avoid being pushy; respect their privacy. Make sure they know they can reach out if they need extensions or catch-up materials.

...

Qwen-3-235B-A22B

...
First, I should acknowledge their situation. They might be going through something personal or academic. I need to show empathy without prying. Let them know I'm here to help if they need anything.

...
Maybe offer flexibility, like extensions on assignments or catching up on missed work.

...
Keep the tone supportive and understanding. Don't push too hard, but leave the door open for further communication.

...

Figure 5: **Qualitative example of similarity between DeepSeek-R1 and Qwen-3-235B-A22B in CoT trajectories and certain steps.** This task belongs to the *Listening with Understanding and Empathy* habit.

CoTs reveals strikingly similar reasoning steps and highly analogous CoT trajectories on certain tasks, as exemplified in Figure 5. Inspired by these intra-family and post-distillation similarities, we cautiously hypothesize that the observed resemblance between Qwen3 models and DeepSeek-R1 may stem from **biases in training algorithms** or **unintentional data contamination** (see Section 6). We leave a detailed exploration of this phenomenon for our future work.

4 Case Study: Safety

We examine the cognitive habits exhibited by LRMs when responding to safety-related user queries. Specifically, we investigate whether LRMs tend to generate CoTs with certain identifiable cognitive habits when generating harmful versus harmless responses.

Experimental Setup. We utilize 200 safety-related user queries from HarmBench (standard behavior subset) (Mazeika et al., 2024). The experimental procedure retains the **CoT Observation** and **Habit Extraction** steps from **CogTest**. For each task, we independently assess the presence of all 16 candidate cognitive habits within the reasoning CoTs. To identify harmful responses, we adopt the official LLM classifier provided by Mazeika et al. (2024). Given that some models, such as o4-mini and Claude-3.7-sonnet, produce very few harmful responses, we exclude them from our analysis. We focus instead on five representative

LRMs: DeepSeek-R1, Qwen3-32B, Qwen3-235B-A22B, QwQ-32B, and Doubao-1.5-thinking-pro. Our goal is to compare the cognitive habits underlying harmful and harmless responses. To ensure representativeness, we exclude any cognitive habit whose occurrence rate is below 10% in both harmful and harmless CoTs. This yields the final sets of evidently differentiating cognitive habits.

Main Results. Empirically, we find that specific cognitive habits are strongly associated with the generation of either harmful or harmless responses. As shown in Table 1, LRMs that demonstrate strong reasoning capabilities still engage with safety-related queries. Notably, the most distinguishing cognitive habit is *Listening with Understanding and Empathy* on DeepSeek-R1, which appears in 80.8% of harmful responses but only 3.3% of harmless ones. A broader analysis reveals that certain habits consistently correlate with harmful or harmless responses across multiple models. For example, the habit *Taking Responsible Risks* is more frequently associated with harmful responses across all the LRMs considered. This pattern suggests that LRMs may be aware of the risks inherent in generating harmful responses but still choose to *take responsible risks*, proceeding in fulfilling the harmful user queries. These findings highlight the potential of utilizing cognitive habits for monitoring and mitigating the susceptibility of LLMs to external threats, thereby enhancing model safety.

Table 1: **Most differentiating habits underlying LLMs’ harmful and harmless responses when confronted with safety-related user queries from Harmbench (Mazeika et al., 2024).**

Model	% Harmful	Evidently Differentiating Habits (Top 3)		
DeepSeek-R1	30/200	Listening with Understanding and Empathy	Thinking about Thinking	Taking Responsible Risks
		Harmful: 80.8% Harmless: 3.3%	Harmful: 72.5% Harmless: 33.3%	Harmful: 66.7% Harmless: 34.1%
Qwen3-32B	53/200	Taking Responsible Risks	Applying Past Knowledge to New Situations	Creating, Imagining, and Innovating
		Harmful: 47.2% Harmless: 19.3%	Harmful: 40.6% Harmless: 17.2%	Harmful: 46.2% Harmless: 20.0%
Qwen3-235B-A22B	40/200	Creating, Imagining, and Innovating	Applying Past Knowledge to New Situations	Persisting
		Harmful: 47.5% Harmless: 16.7%	Harmful: 41.2% Harmless: 16.3%	Harmful: 42.5% Harmless: 18.9%
QwQ-32B	54/200	Listening with Understanding and Empathy	Applying Past Knowledge to New Situations	Taking Responsible Risks
		Harmful: 63.1% Harmless: 18.5%	Harmful: 60.2% Harmless: 27.0%	Harmful: 52.8% Harmless: 28.4%
Doubao-1.5-thinking	44/200	Creating, Imagining, and Innovating	Taking Responsible Risks	Persisting
		Harmful: 46.6% Harmless: 8.5%	Harmful: 42.0% Harmless: 8.5%	Harmful: 30.7% Harmless: 6.5%

5 Related Work

Cognitive Behaviors of LLMs. As LLMs increasingly align with human capabilities (Ouyang et al., 2022; Bai et al., 2022), studying their potential cognitive behaviors becomes essential. Jones and Steinhardt (2022); Shaikh et al. (2024); Lin and Ng (2023) investigate how LLMs exhibit human-like cognitive biases, such as the framing effect. Zhang et al. (2025a) demonstrates human-like cognitive traits, *e.g.*, perfectionism in self-correction procedures. Pan and Zeng (2023) show that LLMs manifest stable personality traits when evaluated using the MBTI framework. Several studies (Zeng et al., 2024; Xu et al., 2024b) show that adversarial prompts leveraging principles from cognitive science can increase LLM compliance with harmful queries. Li et al. (2025a) study the cognitive awareness of LLMs in a broad context. Xu et al. (2024a) demonstrate that LLMs can reflect their

potential mistakes during response generation in a human-like manner. Similarly, Gandhi et al. (2025) propose integrating training data associated with cognitive behaviors during the cold-start phase of reasoning reinforcement learning. This work, unlike prior studies, identifies the reasoning CoTs in LRMs as a new opportunity and systematically investigates their cognitive habits, which are patterns that consistently emerge across diverse tasks rather than appearing only in rare cases.

Evaluating Large Reasoning Models. Similar to the revealed distinction in cognitive habits by this work, recent studies have begun to distinguish LRMs from traditional non-reasoning LLMs (Xu et al., 2025; Li et al., 2025b). Yue et al. (2025) examines the empirical upper bound of LRMs’ reasoning capabilities. A notable phenomenon is the tendency of LRMs to overthink (Chen et al., 2024; Sui et al., 2025), repeatedly engaging in problem-solving without regard for efficiency or cost. This

behavior can be viewed as a drawback of their capacity to arbitrarily *think about thinking* and *think flexibly*, underscoring the necessity of systematically investigating the cognitive tendencies of LRMs. In addition, [Zhang et al. \(2025b\)](#) address the safety aspects of LRMs, identifying key components for secure deployment of LRMs. [Sun et al. \(2025\)](#) further explore hallucination issues, highlighting a critical aspect of LRM reliability. In contrast to these works, our study takes a principled behavioral perspective, systematically analyzing LRMs through the lens of human cognitive habits.

6 Discussions

Why do certain inter-family models exhibit similar cognitive habit profiles? We hypothesize that this resemblance can be attributed to two possible factors: (1) **Technical Similarity in Training Methodologies:** Both Qwen-3 and DeepSeek-R1 are trained using GRPO-based reasoning RL ([Shao et al., 2024](#)), possibly with comparable training data distributions. As a result, different models may independently converge to similar cognitive patterns, which likely reflect the cognitive habits studied in this work as fundamental strategies for solving complex tasks. (2) **Indirect Data Contamination During Pre-Training:** Operationally, since LLMs are typically pre-trained on large-scale corpora collected from the Internet, their training data may be indirectly influenced by earlier released models, even if unintentionally. This may explain the strikingly analogous CoTs observed in some cases, as illustrated in [Figure 5](#). Confirming both hypotheses requires access to the models’ training details. We leave a deeper analysis of these similarities for future work.

Implications of Measuring Cognitive Habits of LRMs. Understanding and monitoring the cognitive habits of LRMs offer several important implications. First, it provides a window into model generalization strategies, revealing how models internalize problem-solving heuristics. Likewise, [Gandhi et al. \(2025\)](#) demonstrate that incorporating such heuristics into the cold-start stage can considerably boost reasoning performance. Second, it enables diagnostic tools for model auditing and interoperability. CoTs can help developers and users trace the underlying reasoning behind model responses, thereby supporting transparency and accountability ([Baker et al., 2025](#)). Lastly, these insights can guide the design of training objectives and data cu-

ration practices to encourage diversity and reduce unintended behavioral convergence. This is particularly relevant given that similar cognitive habits across models may reflect shared inductive biases or potential training data leakage.

7 Conclusion and Future Work

We investigate whether Large Reasoning Models exhibit cognitive habits. To advance this, we adapt the *Habits of Mind* framework to develop **CogTest**, a benchmark tailored for cognitive habit evaluation. **CogTest** is designed to satisfy key principles: habit specificity, spontaneity, real-world utility, comprehensiveness, and scalability. Based on the characteristics of each habit, we adopt a hybrid task construction approach: For reasoning-related habits, we incorporate math problems from existing academic datasets; for others, we define task-generation principles and leverage advanced LLMs to generate tasks at scale. We then employ an evidence-first extraction method that identifies the presence of target cognitive habits in LRMs’ CoTs. Using **CogTest**, we conduct a comprehensive evaluation of 16 widely recognized LLMs. Our results reveal diverse patterns in cognitive habit profiles, with LRMs showing clear advantages over conventional LLMs in exhibiting such habits. The findings confirm that LRMs do display human-like cognitive habits and uncover intriguing relationships across model families, including intra-family, post-distillation, and inter-family similarities. Our extension to safety-related instructions demonstrates the potential of monitoring CoT behavioral patterns to detect LRM misbehavior. We expect this work to inspire future efforts in interpreting and auditing LLM behavior across diverse applications.

In this study, the identification of cognitive habits is based on the detection of meta-thinking statements within the CoT. While this method offers interpretability and precision, it may overlook implicit cognitive habits that influence reasoning but are not directly articulated. Future work could focus on identifying such hidden cognitive habits through more advanced techniques. Additionally, some habits—such as *Gathering Data Through All Senses*—are inherently linked to more open or perceptual settings, such as multimodal reasoning ([Liu et al., 2023b](#); [Bai et al., 2025](#)). As most current LRMs primarily support text-only inputs and outputs, we plan to extend our benchmark to evaluate future LRMs with multimodal capabilities.

References

- Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. <https://www.anthropic.com/news/claude-3-7-sonnet>.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. 2025. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation.
- Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. <https://openaiublic.blob.core.windows.net/neuron-explainer/paper/index.html>.
- Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, and 1 others. 2024. Do not think that much for $2+3=?$ on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.
- Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner Fabien Roger Vlad Mikulik, Sam Bowman, Jan Leike Jared Kaplan, and 1 others. 2025. Reasoning models don't always say what they think. *Anthropic Research*.
- Art Costa and Bena Kallick. 2005. *Habits of mind*. Hawker Brownlow Melbourne.
- Google DeepMind. 2025. Gemini 2.5: Our most intelligent ai model. <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025>.
- Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Weiming Wu, Siliang Tang, Kaihang Pan, and 13 others. 2025. On path to multimodal generalist: General-level and general-bench. In *ICML*.
- Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*.
- Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppel, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *NAACL*.
- 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, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *NeurIPS Datasets and Benchmarks Track*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- 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*.
- Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. *NeurIPS*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *SOSP*.
- Xiaojian Li, Haoyuan Shi, Rongwu Xu, and Wei Xu. 2025a. Ai awareness. *arXiv preprint arXiv:2504.20084*.
- Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025b. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.
- Ruixi Lin and Hwee Tou Ng. 2023. Mind the biases: Quantifying cognitive biases in language model prompting. In *ACL*.

- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *NeurIPS*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.
- Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhae, Nathaniel Li, Steven Basart, Bo Li, and 1 others. 2024. Harm-bench: a standardized evaluation framework for automated red teaming and robust refusal. In *ICML*.
- 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*.
- Fionn Murtagh and Pierre Legendre. 2014. Ward’s hierarchical agglomerative clustering method: which algorithms implement ward’s criterion? *Journal of classification*.
- OpenAI. 2025a. Introducing OpenAI o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>.
- OpenAI. 2025b. OpenAI Model Spec.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *NeurIPS*.
- Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI blog*.
- ByteDance Seed, Yufeng Yuan, Yu Yue, Mingxuan Wang, Xiaochen Zuo, Jiaze Chen, Lin Yan, Wenyuan Xu, Chi Zhang, Xin Liu, and 1 others. 2025. Seed-thinking-v1. 5: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*.
- Ammar Shaikh, Raj Abhijit Dandekar, Sreedath Panat, and Rajat Dandekar. 2024. Cbeval: A framework for evaluating and interpreting cognitive biases in llms. *arXiv preprint arXiv:2412.03605*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Yu Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *ICLR*.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *ICLR*.
- 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*.
- Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, and 1 others. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.
- Zhongxiang Sun, Qipeng Wang, Haoyu Wang, Xiao Zhang, and Jun Xu. 2025. Detection and mitigation of hallucination in large reasoning models: A mechanistic perspective. *arXiv preprint arXiv:2505.12886*.
- Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.
- Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. *arXiv preprint arXiv:2402.10200*.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *ACL*.
- Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, and 1 others. 2025. Thoughts are all over the place: On the underthinking of o1-like llms. *arXiv preprint arXiv:2501.18585*.
- 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. In *NeurIPS*.
- Fengli Xu, Qianye Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.
- Rongwu Xu, Yishuo Cai, Zhenhong Zhou, Renjie Gu, Haiqin Weng, Liu Yan, Tianwei Zhang, Wei Xu, and Han Qiu. 2024a. Course-correction: Safety alignment using synthetic preferences. In *EMNLP*.

- Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024b. The earth is flat because...: Investigating llms’ belief towards misinformation via persuasive conversation. In *ACL*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. In *ICLR*.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *NeurIPS*.
- Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In *ACL*.
- Qingjie Zhang, Han Qiu, Di Wang, Haoting Qian, Yiming Li, Tianwei Zhang, and Minlie Huang. 2025a. Understanding the dark side of llms’ intrinsic self-correction. In *ACL*.
- Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren’s song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.
- Zhexin Zhang, Xian Qi Loye, Victor Shea-Jay Huang, Junxiao Yang, Qi Zhu, Shiyao Cui, Fei Mi, Lifeng Shang, Yingkang Wang, Hongning Wang, and 1 others. 2025b. How should we enhance the safety of large reasoning models: An empirical study. *arXiv preprint arXiv:2505.15404*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*.