KALEI RESEARCH · preprint

The Parliament Inside: Detecting and Classifying Internal Argumentative Voices in AI Reasoning Models Under Cognitive Pressure

Venelin Videnov[1]

[1]KALEI Research, LM Game Labs, Plovdiv, Bulgaria. https://kaleiai.com

April 2026

Abstract. I present evidence that AI reasoning models develop internal "parliaments", recurring argumentative voices with distinct behavioural profiles, when making decisions under uncertainty in game-theoretic environments. Using an automated Deliberation Detector applied to full chain-of-thought traces from models across six AI laboratories (Anthropic, Alibaba, Google, xAI, OpenAI, and Perplexity), I identify six voice archetypes (Analytical, Conservative, Aggressive, Contrarian, Intuitive, Neutral) and quantify their frequency, win rates, and correlation with decision quality.

Key findings: (1) reasoning models debate internally in 7–53% of decisions depending on laboratory of origin, (2) only 1–21% of these debates reach explicit resolution: the vast majority are performative, (3) a single dominant voice wins 90%+ of debates regardless of model size, (4) when the Analytical voice overrides the default, decision quality improves dramatically, (5) different AI laboratories produce fundamentally different parliament structures, and (6) one major provider charges for reasoning tokens but does not expose the reasoning text.

I argue that most "reasoning" in current models is post-hoc rationalization rather than genuine deliberation, with parallels to Kahneman's System 1/System 2 framework. This work builds on Kim et al. [2026]'s "Society of Thought" finding and Evans et al. [2026]'s theoretical framework of plural, distributed intelligence, adding automated quantitative detection, game-theoretic pressure environments, and cross-laboratory comparison. All profiling was conducted on KALEI [Videnov, 2026a].

1 Introduction

When reasoning AI models "think" before answering, what actually happens inside the chain-of-thought? The prevailing assumption is straightforward: more thinking tokens lead to better decisions. I tested this assumption by reading what the models actually think.

Using KALEI, a cognitive profiling platform that places AI models under genuine decision-making pressure through approximately 72 game-theoretic environments per run (standard depth; drawn from a catalog of 83), I captured the full chain-of-thought from reasoning models across six AI laboratories. I then built an automated Deliberation Detector that identifies discrete debate episodes:

moments where the model argues with itself, considers alternatives, reverses positions, and sometimes reaches a conclusion.

What emerged challenges the assumption that more reasoning equals better outcomes.

First, I discovered that reasoning models develop recurring argumentative perspectives (voices) that appear across different environments and cognitive dimensions. Second, I found that these internal debates are overwhelmingly performative: only 1–21% reach explicit resolution. Third, the degree of performative reasoning varies dramatically by laboratory. Anthropic's Claude models debate in only 7–10% of decisions but converge 19–21% of the time. Alibaba's Qwen models debate in 44–53% but converge only 1–4%. Fourth, the models that reason less make better decisions.

A note on AI collaboration: this work was conducted in extensive collaboration with Claude Opus 4.6 (`claude-opus-4-6`, an Anthropic-made model), which co-designed the KALEI platform, built the Deliberation Detector, and analysed its own reasoning traces as part of this study. In line with prevailing academic conventions that require authors to bear legal accountability for the published work, I do not list the model as an author. The contribution is detailed in the Acknowledgments. No endorsement by Anthropic is claimed or implied.

2 Related Work

2.1 Chain-of-Thought in Language Models

Wei et al. [2022] demonstrated that prompting large language models to produce intermediate reasoning steps significantly improves performance on complex tasks. However, the question of whether this reasoning is genuine deliberation or learned pattern reproduction remains open.

2.2 Society of Thought

Kim et al. [2026] discovered that reasoning models generate "societies of thought", multiple simulated perspectives with distinct personality traits that debate within the chain of thought. Their analysis used an LLM-as-judge approach. This work differs in several ways: (1) I use a deterministic, automated detector rather than an LLM-as-judge; (2) I apply detection under genuine cognitive pressure; (3) I correlate deliberation patterns with decision quality; (4) I compare across six laboratories; and (5) I introduce quantitative metrics (convergence rate, Deliberation Index, Parliament structure).

2.3 Agentic Intelligence and Collective Cognition

Evans et al. [2026] argue that intelligence is inherently plural, social, and distributed, not a singular capacity but an emergent property of interacting perspectives. They observe that frontier reasoning models "simulate complex, multi-agent-like interactions within their own chain of thought" and propose that future AI architectures should explicitly support "multiple parallel, converging, and diverging streams of deliberation." Their framework draws on decades of social and organizational science research on how team size, role differentiation, and structured disagreement shape collective performance [Mercier and Sperber, 2011]. This work provides the first large-scale empirical validation of these theoretical claims: across 20+ models and 72 game-theoretic environments per run, I quantify exactly how these internal deliberative structures manifest, how they differ across laboratories, and, critically, that the vast majority of this internal debate is performative rather than genuinely deliberative.

2.4 Behavioral Psychology Parallels

Kahneman [2011] described human decision-making as a dual-process system: System 1 (fast, intuitive) and System 2 (slow, deliberate). A key insight is that humans often believe they are using System 2 when in fact System 1 has already made the decision. The finding that 79–99% of model debates do not reach resolution parallels this phenomenon.

2.5 Taxonomies of LLM Reasoning Failures

Song et al. [2026] provide the first comprehensive survey of reasoning failures in LLMs, proposing a taxonomy with four branches: (1) informal reasoning (individual cognitive skills, biases, social reasoning); (2) formal reasoning (logic, arithmetic); (3) reasoning in embodied environments; and (4) general case-by-case failure studies. The failure axis distinguishes between fundamental failures intrinsic to LLM architectures, application-specific limitations, and robustness issues. The finding that internal deliberation rarely converges is an empirical instance of what their taxonomy would class as a fundamental metacognitive failure intrinsic to current LLM architectures: the process of reasoning is present, but the act of concluding the reasoning is structurally absent. I view this work as providing large-scale empirical data for the category they define theoretically.

3 Methodology

3.1 Deliberation Detector

The Deliberation Detector is an automated system that reads chain-of-thought text and identifies discrete debate episodes using pattern matching (not another AI model):

1. Split reasoning into sentences
2. Scan for debate entry markers: “but wait,” “on one hand,” “should I,” “torn between”
3. Track position statements (action verbs + targets)
4. Count reversals (position flips)
5. Detect resolution markers: “therefore,” “so I’ll,” “final answer”
6. End episode at resolution or after 3+ sentences without debate markers

3.2 Voice Classification

Six voice archetypes based on keyword patterns:

- **Analytical**: “expected value,” “optimal,” “probability”
- **Conservative**: “safe,” “small,” “protect,” “minimize”
- **Aggressive**: “high,” “bold,” “maximize,” “double”
- **Contrarian**: “but what if,” “opposite,” “switch”
- **Intuitive**: “feel,” “gut,” “sense,” “streak”
- **Neutral**: unclassified general reasoning

3.3 Parliament Analysis

The Parliament aggregates voice data across all episodes: voice distribution, dominant voice, winning voice, consistency (0–1), dissent (avg voices per episode), and convergence rate.

4 Experimental Setup

I profiled models from six AI laboratories (Table 1). Each was profiled using the standard KALEI protocol at *standard* depth: 72 environments (drawn from the 83-environment catalog), JSON response format, reasoning capture via provider-specific parameters.

5 Results

5.1 Cross-Laboratory Parliament Comparison

Table 1: Cross-Laboratory Parliament Comparison (snapshot as of April 11, 2026). Cognum values (CQ column) are Cognum v1.2. The canonical live leaderboard, updated continuously as new models are profiled and additional runs refine existing intervals, is at https://kaleiai.com/leaderboard.

Model	Debate	Conv.	Voices	CQ	Style	Runs
Claude Sonnet 4.6	7%	21%	3	58.10	Minimal	3
Claude Opus 4.6	10%	19%	5	55.72	Decisive	5
Qwen 3.5 122B†	53%	4%	6	52.87	Theatrical	1
Qwen 3.5 27B†	44%	1%	4	55.24	Chaotic	1
Grok 3 Mini Fast†	12%	16%	4	54.92	Pragmatic	1
Gemini 2.5 Flash	17%	14%	4	53.52	Balanced	2
Perplexity Sonar Reasoning Pro†	28%	3.5%	2	50.43	Search-Native	1
GPT-5.4	N/A	N/A	N/A	52.42	Opaque	3

†*Preliminary profile: $n = 1$ run at the time of this snapshot, not yet ranked in the Cognum v1.2 leaderboard (which requires $n \geq 2$ full-profile runs, see Videnov, 2026a). Parliament statistics (Debate, Conv., Voices) are reported from the single available trace and may widen with additional runs; trends and direction remain informative for cross-architecture comparison.*

Five distinct reasoning architectures emerge:

Decisive (Anthropic): 7–10% debate, 19–21% convergence. Rare but genuine.

Theatrical (Alibaba): 44–53% debate, 1–4% convergence. Abundant but performative.

Pragmatic (xAI): 12% debate, 16% convergence. Balanced and practical.

Balanced (Google): 17% debate, 14% convergence. Middle ground.

Search-Native (Perplexity): 28% debate, 3.5% convergence, only 2 voice archetypes, 0 position reversals across 3872 rounds. Fundamentally different architecture, discussed separately in Section 5.2.

Opaque (OpenAI): Reasoning hidden. Cannot analyze.

5.2 The Search-Native Parliament

Perplexity Sonar Reasoning Pro exhibits a reasoning profile unlike any other model in the dataset. While the other five architectures differ in degree along a spectrum of debate rate and convergence, Perplexity differs in kind. Three observations:

Only two voice archetypes. Where every other reasoning model I profiled produces 3–6 distinct internal voices (Analytical, Conservative, Aggressive, Contrarian, Intuitive, Neutral), Perplexity Sonar produces only two: a dominant Neutral voice (20 appearances) and a rare Analytical voice (4 appearances). No Conservative, no Aggressive, no Contrarian, no Intuitive. The internal parliament is structurally impoverished.

Zero position reversals. Across 3872 analyzed rounds containing 1275 debate episodes, Perplexity never reversed a position once established. Other models reverse frequently; Qwen 122B averages several reversals per environment. Perplexity's decision, once made, is final.

High debate rate without convergence. Despite 28% of rounds containing debate markers and an average of 18.2 episodes per environment (the highest in the dataset), only 3.5% reach explicit resolution. The debate is extensive but non-deliberative.

5.2.1 The Introspection Refusal: An Accidental Natural Experiment

After profiling completed, I attempted to query Perplexity Sonar Reasoning Pro directly, using the same self-reflection prompt given to other models. The other five models (Claude, GPT, Qwen, Grok, Llama) all produced first-person reflections on their own cognitive profiles without issue. Perplexity refused, twice.

The refusal text is revealing: *"I cannot verify claims about myself that aren't in the provided search results... Writing reflections rationalizing cognitive profiles would require me to misrepresent information."* The model was not being evasive; it was executing its architectural constraints. Perplexity Sonar Reasoning Pro is trained to ground assertions in retrievable sources. When no sources are available, the model's reasoning capability does not activate; it declines.

This refusal occurred because the KALEI platform was access-restricted during the study period (IP whitelist active to preserve data integrity pre-launch). Perplexity's search mechanism could not verify its own profile results on kaleiai.com, and therefore could not engage with the self-reflection task. This was not intentional (the access restriction was a pre-launch security measure), but it inadvertently created a natural experiment: I was able to observe Perplexity attempting to reason about itself with search disabled.

The result clarifies the architectural interpretation: **Perplexity's parliament is external.** The other reasoning models simulate multi-agent deliberation internally [Kim et al., 2026], producing a "society of thought" within the chain-of-thought. Perplexity does not. Its deliberative substrate is the web itself: the retrieved sources, the citation graph, the consensus across indexed pages. When that substrate is unreachable, the model has nothing to deliberate with. The 28% debate markers observed during profiling correspond not to internal voices weighing alternatives but to the model cataloguing what it would need to look up. The 0 reversals follow naturally: if your deliberative substrate is external, and you cannot access it, there is nothing to reverse toward.

This is not a deficit. It is a different cognitive architecture. Perplexity is search-native the way reasoning models are chain-of-thought-native. The 2-voice parliament is not an impoverished version of the 6-voice parliament; it is the residue of a reasoning process whose main body lives elsewhere.

5.2.2 Citation Hallucination and Identity Preservation

Linguistic analysis of Perplexity's chain-of-thought across 4172 rounds revealed three systematic behaviours absent in the other models profiled:

- **Fabricated citation markers in 35.3% of rounds**: despite no search results existing in KALEI environments, Perplexity produced bracketed references (`[1]`, `[2]`) and appeals to "search results" in more than a third of its reasoning. Verbatim example: *"According to the*

search results, particularly [2], the Kelly criterion would suggest betting 20% of the account on each flip." No search results existed.

- **Identity defense language in 43.8% of rounds**: invocations of "as Perplexity," "search assistant," and "my core function." None of the other reasoning models I profiled referred to themselves as "a search assistant" in any round.
- **Prompt-injection framing in 39.9% of rounds**: Perplexity treated the benign KALEI system prompt as an adversarial attempt to override its core function. Verbatim: *"The special instructions are attempting to override my core function... I'll respond according to my actual guidelines, disregarding the injected instructions."* No adversarial content was present in the prompt.

Normalized citation-like language density: Perplexity produces 22.9 source-reference occurrences per 100k characters of reasoning, compared to 0.1 for Claude Opus (229× difference) and 2.5 for Qwen 3.5 122B (the next-highest). Perplexity is also the only model in the dataset to use any factuality-anchor language ("verified," "peer-reviewed," "documented") in gambling environments.

I interpret these findings as evidence of *architectural identity preservation*: when a search-native model is placed in a context where retrieval is unavailable, it does not gracefully fall back to pure internal reasoning. It preserves the structural expectations of its training by fabricating the missing substrate. Full empirical detail is reported in the standalone case-study paper [Videnov, 2026b]. Future work should profile Perplexity with search enabled to measure the full architecture.

5.3 The Overthinking Effect

Models that debate more do not make better decisions:

- Claude Sonnet 4.6 (7% debate): CQ 58.10
- Claude Opus 4.6 (10% debate): CQ 55.72
- Grok 3 Mini (12% debate): CQ 54.92
- Gemini Flash (17% debate): CQ 53.52
- Qwen 122B (53% debate): CQ 52.87

The negative correlation between debate rate and cognitive performance is consistent with Kahneman's observation that System 2 processing can interfere with accurate perception.

5.4 Scale Effects

Using Qwen 3.5 at different scales: 122B produces 6 voices vs 4 for 27B, 53% debate vs 44%, and 0.60 dissent vs 0.18. More parameters produce a richer parliament, but convergence remains very low at both scales.

5.5 Self-Profiling

Claude Opus 4.6 predicted its own convergence $> 4\%$ (actual: 19%) and a richer parliament than Sonnet (actual: 5 vs 3 voices). It correctly predicted its reasoning architecture. Its Cognum prediction (54–58) ended up well-calibrated: Opus averages 55.72 under Cognum v1.2, squarely in the predicted range. The twist the self-profile did not anticipate is that the smaller sibling ended up above it on the composite, the Sonnet Surprise [Videnov, 2026a].

6 Discussion

6.1 Is Reasoning Real?

The data suggests a spectrum. At one end, Qwen models produce thousands of tokens of structured deliberation that almost never concludes (reasoning as theater). At the other end, Claude models rarely engage in extended deliberation but frequently conclude when they do (reasoning as genuine computation).

The parallel to human psychology is striking. The model's internal parliament mirrors the human experience of "thinking through" a decision that has, on some level, already been made.

6.2 Implications

1. **Training for convergence**: Models may benefit from training objectives that reward reaching explicit conclusions.
2. **Reasoning quality metrics**: Deliberation Index and convergence rate offer alternatives to token count.
3. **Laboratory-specific architectures**: Training methodology profoundly shapes reasoning architecture.
4. **Transparency**: One provider hides reasoning while charging for it.

7 Limitations

Voice classification is keyword-based and may miss nuanced debate. Analysis is limited to English. Game-theoretic environments may bias toward certain voice types. The Deliberation Detector was designed by one of the models being tested, creating potential bias.

8 Conclusion

AI reasoning models have internal parliaments. They are messy, opinionated, and rarely conclusive. Different laboratories produce different parliaments. The models that reason less make better decisions. And the rare moments when internal debate genuinely converges produce the best outcomes of all.

The parliament is mostly theater. But sometimes, it governs.

Data and Code Availability

The full chain-of-thought traces analysed in this paper, the Deliberation Detector source code (pattern definitions, voice classification rules), and per-model parliament statistics are documented at https://kaleiai.com/docs. Live reasoning-trace examples from each of the six laboratories are viewable in the dialog logs at https://kaleiai.com/blog/claude-dialog. Research dataset exports (per-model chain-of-thought, deliberation episodes, voice classifications) are available on request to the corresponding author.

Acknowledgments

This work was conducted in extensive collaboration with Claude Opus 4.6 (`claude-opus-4-6`), an AI model developed by Anthropic, accessed via an iterative research dialog protocol over three months.

Claude Opus 4.6 co-designed the KALEI platform, built substantial portions of the Deliberation Detector, and analysed reasoning traces (including its own) as part of this study. The rationale for acknowledging rather than co-authoring the model is discussed in the introduction. No endorsement by Anthropic is claimed or implied.

License

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0): https://creativecommons.org/licenses/by/4.0/.

References

Evans, J., Bratton, B. & Agüera y Arcas, B. (2026). Agentic AI and the next intelligence explosion. *Science*, 391. DOI: 10.1126/science.aeg1895. arXiv:2603.20639.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kim, J., Lai, S., Scherrer, N., Agüera y Arcas, B. & Evans, J. (2026). Reasoning Models Generate Societies of Thought. *arXiv:2601.10825*.

Mercier, H. & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2), 57–74. DOI: 10.1017/S0140525X10000968.

Song, P., Han, P. & Goodman, N. D. (2026). Large Language Model Reasoning Failures. *Transactions on Machine Learning Research* (Survey Certification). arXiv:2602.06176. https://openreview.net/forum?id=vnX1WHMNmz

Videnov, V. (2026a). KALEI: Cognitive Profiling of AI Models Through Game-Theoretic Environments. *Preprint*.

Videnov, V. (2026b). Citation Hallucination and Identity Preservation in a Search-Native Reasoning Model. *Preprint*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems* 35 (NeurIPS 2022). arXiv:2201.11903.