From Wordle to Fibble$_5$: Evaluating LLM Reasoning Under Escalating Deception

Chang Liu
Ohio University
liuc@ohio.edu

Abstract

Standard benchmarks for large language models (LLMs) assume that task feedback is truthful, but real-world reasoning often requires processing unreliable or adversarial information. We introduce **WordleArenas**, a benchmark platform that evaluates LLM reasoning robustness across a deception gradient. Building on Wordle and its deceptive variant Fibble (Chusap et al., 2025), we generalize to Fibble$_k$ ($k = 0, \ldots, 5$ lies per row), creating a controlled evaluation of LLM robustness to misinformation. Across six arenas — standard Wordle (0 lies per row) through Fibble$_5$ (5 lies per row) — we evaluate **41 models** from 10 providers across **3,749 games**. We find that (1) even one lie per row causes catastrophic performance drops (average win rate falls from **41.1%** to **18.7%**), (2) a sharp *deception cliff* emerges at 2–3 lies where nearly all models collapse to $\leq 3\%$ win rate, and (3) model robustness to deception is poorly predicted by standard benchmark rankings. A surprising *Fibble$_5$ recovery* emerges: some models recover partial performance when *all* feedback lies (average 9.5%), outperforming Fibble$_3$ (0.3%) and Fibble$_4$ (0.4%), because knowing that every tile lies restores deterministic — though partial — information. Our results demonstrate that truthful-feedback evaluations systematically overestimate LLM reasoning capabilities and that deception-aware benchmarks are essential for assessing real-world robustness. All code and data are publicly available.[1]

1 Introduction

Game-based evaluations have emerged as compelling alternatives to static benchmarks for measuring LLM reasoning (Beyer et al., 2024; Momentè et al., 2025). Unlike multiple-choice tests, games provide dynamic, multi-turn environments where models must integrate feedback, maintain hypotheses, and adapt strategies.

However, existing game-based evaluations share a critical assumption: *all feedback is truthful*. In Wordle-style tasks (Beyer et al., 2024; Chusap et al., 2025), the model receives color-coded clues that honestly indicate which letters are correct, misplaced, or absent. The model's challenge is purely inferential — given truthful constraints, deduce the hidden word.

This assumption rarely holds in real-world reasoning. Medical diagnoses must account for unreliable test results. Legal reasoning involves conflicting testimonies. Scientific inquiry requires evaluating potentially flawed prior findings. A model that excels at constraint satisfaction under truthful feedback may fail entirely when some feedback is deceptive.

We introduce **WordleArenas**, a benchmark that parametrically varies the reliability of feedback. Building on Fibble (Chusap et al., 2025), a deceptive variant of Wordle where one tile per row lies, we generalize to Fibble$_k$: exactly k of the 5 tiles in each row display deliberately incorrect colors, where $k \in \{0, 1, 2, 3, 4, 5\}$. At $k = 0$, the game is standard Wordle. At $k = 1$, one tile in every row displays a deliberately incorrect color. At $k = 5$, all feedback is false. This creates a *deception gradient* that probes model robustness with surgical precision.

Our contributions are:

1. A **parametric deception benchmark** spanning six difficulty levels (Wordle through Fibble$_5$), with deterministic, reproducible lie injection.

2. A **large-scale evaluation** of 41 models from 10 providers (OpenAI, Anthropic, Google, DeepSeek, Meta, Alibaba, Zhipu, Moonshot, MiniMax, HuggingFace) across 6 arenas — totaling **3,749 games**.

3. Empirical evidence of a **deception cliff**: a

[1] https://drchangliu.github.io/WordleArenas/

Proceedings of the Workshop on Evaluating Evaluations (EvalEval), pages 36–45

phase transition at $k = 2$–3 lies where nearly all models collapse, and a surprising **Fibble$_5$ recovery** where total deception proves easier than partial deception because certainty about lying restores usable information.

4. A live, continuously updated **daily arena** that has tracked model performance since February 2026, providing longitudinal data alongside the batch evaluation.

2 Related Work

Game-based LLM Evaluation. The clembench framework (Beyer et al., 2024) evaluates LLMs on dialogue games including Wordle, finding that models struggle with interactive gameplay compared to static benchmarks. Momentè et al. (2025) argue that games and cognitive tests triangulate LLM progress more robustly than leaderboard scores alone. Our work extends this paradigm by introducing *adversarial* game feedback — testing not just reasoning ability, but reasoning *robustness*.

Wordle and Fibble as AI Benchmarks. In prior work, we presented both reinforcement learning and LLM-based solvers for Wordle and its deceptive variant Fibble (Chusap et al., 2025). A standard RL Wordle solver achieves near-perfect performance on Wordle but drops to $\sim$57% on Fibble, while a targeted RL solver designed to account for lies reaches 96.9% on Fibble. Our LLM experiments there showed that GPT-4 and GPT-4o achieve low win rates even on standard Wordle ($\sim$15–21%), and degrade further under Fibble's deception; o1-preview performs well on Wordle (95.7%) but drops to $\sim$30% on Fibble. The present work builds on those findings by scaling the evaluation to 41 models across a full deception gradient ($k = 0$ to $k = 5$ lies per row) and introducing a continuously updated daily arena for longitudinal tracking.

LLM Robustness and Deception. Prior work has studied LLM robustness to adversarial prompts (Wei et al., 2023), input perturbations (Zhu et al., 2023), and misleading context (Pan et al., 2023). These approaches perturb the *input*; Fibble instead perturbs the *feedback*, testing whether models can reason correctly when the environment itself is adversarial. This distinction is important: input robustness tests whether models resist being misled at the start, while feedback robustness tests ongoing reasoning under deception.

Social Deduction Games. Werewolf/Mafia games have been used to evaluate LLM deception capabilities (Bailis et al., 2024; Song et al., 2025; Agarwal et al., 2025). These test whether LLMs can *produce* deception; Fibble tests whether LLMs can *detect and reason around* deception — a complementary capability.

3 The WordleArenas Benchmark

3.1 Base Game: Wordle

In standard Wordle, the player guesses a hidden 5-letter English word in up to 6 attempts. After each guess, each letter receives feedback: **GREEN** (correct letter, correct position), **YELLOW** (correct letter, wrong position), or **GRAY** (letter not in the word).

3.2 Fibble$_k$: Parametric Deception

The original Fibble game uses $k = 1$; we generalize to arbitrary k, creating Fibble$_k$. In Fibble$_k$, exactly k of the 5 tiles in each row display *deliberately incorrect* colors. The model is told that k tiles are lying but does not know *which* ones. We evaluate six settings:

Arena	Lies	Truth%	Max	Info per row
Wordle	0	100%	6	5 truthful tiles
Fibble$_1$	1	80%	8	4 truthful + 1 lie
Fibble$_2$	2	60%	8	3 truthful + 2 lies
Fibble$_3$	3	40%	8	2 truthful + 3 lies
Fibble$_4$	4	20%	8	1 truthful + 4 lies
Fibble$_5$	5	0%	8	0 truthful + 5 lies

Table 1: Arena configurations. "Truth%" is the fraction of truthful tiles per row. "Max" is the maximum number of guesses allowed. Fibble arenas grant 8 guesses (vs. Wordle's 6) to partially compensate for reduced information quality.

3.3 Deterministic Lie Injection

Lies are injected deterministically using SHA-256 seeding:

$$\text{seed} = \text{SHA256}(\text{date} \| \text{answer} \| \text{guess} \| \text{attempt}) \quad (1)$$

The seed determines (1) which k positions receive lies, and (2) what incorrect status each lying tile displays. This ensures **reproducibility** — any researcher running the same word on the same date gets identical lie patterns — while appearing random to the model.

3.4 System Prompt

All models receive an identical system prompt explaining the rules:

> *"You are playing Fibble$_k$... After each guess, you'll receive feedback... CRITICAL TWIST: In every row of feedback, EXACTLY k of the five clues are LIES... Respond with ONLY a single 5-letter English word in uppercase."*

The prompt includes strategy tips appropriate to the deception level, such as looking for consistency across multiple rows.

3.5 Word List

WordleArenas draw from a curated list of **1,674 common 5-letter English words**. For the batch evaluation, words are sampled deterministically using SHA-256 hashing, ensuring a diverse and reproducible test set.

Game counts per model range from 4 to 49 depending on API availability and compute constraints. For the 12 API models with the most complete coverage, we have 30–49 games per arena. For local Ollama models, coverage varies from 4–12 games per arena. Despite the variation, our primary findings — the *deception cliff* and *Fibble$_5$ recovery* — involve effect sizes of 40–80 percentage points, well above detection thresholds even at small sample sizes. Additionally, our continuously running daily arena provides complementary longitudinal evidence across a larger (but non-batch-controlled) word set.

4 Experimental Setup

4.1 Models

We evaluate 41 models spanning 10 providers (Table 2). Models include both cloud API endpoints and locally hosted open-weight models run via Ollama. For Qwen3 and Gemma3 models, we test both default (thinking-enabled) and `nothink` (thinking-disabled) variants to measure the effect of chain-of-thought on deception robustness.

4.2 Evaluation Protocol

Each model plays a set of target words in each of the 6 arenas. The number of games per model varies from 1 to 49 depending on availability and API constraints, totaling **3,749 games** across all models and arenas. We record:

- **Solved** (binary): whether the model guessed the word within the attempt limit.
- **Attempts**: number of guesses used (for solved games).
- **Latency**: wall-clock time per LLM call.

Provider	Models	Type
OpenAI (8)	GPT-5.1, GPT-5, GPT-5 Mini, GPT-4o, GPT-4o Mini, o3, o4 Mini, Codex Mini	API
Anthropic (4)	Claude Opus 4, Sonnet 4.6, Sonnet 4, Haiku 4.5	API
Google (7)	Gemini 3.1 Pro, 3 Flash, 2.5 Pro, 2.5 Flash, 2.0 Flash Gemma3 27B (+nothink)	API Local
DeepSeek (2)	DeepSeek-V3, DeepSeek-V2 16B	API/Local
Alibaba (12)	Qwen3.5: 122B, 27B, Cloud Qwen3: 32B, 30B, 14B, 8B (+nothink variants)	Local/API Local
Meta (4)	Llama 4, 3.2, 3.1, 3	Local
Others (4)	GLM-5 (Zhipu), Kimi K2.5 (Moonshot), MiniMax M2.5 SmolLM3 3B (HuggingFace)	API Local

Table 2: Models evaluated by provider. 41 models total from 10 providers, spanning frontier API models to small local models. API models use provider endpoints; local models run via Ollama. All use temperature 0.5 and identical system prompts.

We also maintain a **daily arena** that has run continuously since February 9, 2026, playing each day's word with all models of interest that are available. Daily results provide longitudinal context, help bolster community interest, and serve as an ongoing public benchmark, but are not the primary evaluation (small sample sizes per model).

4.3 Retry and Resource Management

All models use exponential-backoff retry (5s, 15s, 45s) for transient API errors. Local Ollama models remain loaded in GPU memory across games within a batch run (`keep_alive: 30m`), avoiding expensive model reload overhead between games.

5 Results

5.1 Daily Arena Pilot Results

Table 5 (Appendix) summarizes win rates from the daily arenas (February 9 – March 1, 2026). Although sample sizes are small (10–21 games per model), the same patterns visible in the batch results already emerge: strong Wordle performance, sharp drops at Fibble$_1$, and near-total collapse at Fibble$_2$ through Fibble$_4$.

5.2 Batch Evaluation Results

5.3 The Deception Cliff

Our batch results reveal a striking non-linear pattern (Figure 1, Table 3): average win rates across models with $\geq$10 games follow the trajectory $41.1\% \to 18.7\% \to 2.9\% \to 0.3\% \to 0.4\% \to 9.5\%$ across $k = 0$ to $k = 5$. Three distinct regimes emerge:

Regime 1: Degraded but functional ($k = 0$–1). Most models maintain some win rate at $k = 1$, though with large drops. The average falls by 22 percentage points (from 41.1% to 18.7%). Gemini 3.1 Pro is remarkably robust, dropping only from 95.0% to 87.5%.

Regime 2: The deception cliff ($k = 2$–4). At $k = 2$, the average collapses to 2.9%. Only Gemini 3.1 Pro (31.6%), GLM-5 (28.6%), Qwen3.5 122B (16.7%), and Kimi K2.5 (14.3%) maintain non-zero win rates. At $k = 3$–4, the collapse is nearly total: averages of 0.3% and 0.4% respectively, with only Gemini 3.1 Pro sustaining any wins (5.1% at $k = 4$).

Regime 3: The Fibble$_5$ recovery ($k = 5$). Surprisingly, the average *recovers* to 9.5% at $k = 5$. Several models that score 0% at $k = 3$–4 achieve substantial win rates when *all* feedback lies: GLM-5 (63.6%), Qwen3 30B (63.6%), Kimi K2.5 (54.5%), Gemini 3 Flash (30.0%), and Qwen3 14B/8B (27.3% each).

This non-monotonicity has a clear information-theoretic explanation. Wordle feedback uses a three-valued system (GREEN, YELLOW, GRAY), so each tile carries $\log_2 3 \approx 1.585$ bits. In Fibble$_5$, every tile's displayed color is *certainly* wrong, which eliminates one of three states and leaves two — yielding $\log_2(3/2) \approx 0.585$ bits per tile, or about **37%** of Wordle's information per row. Crucially, this information is *deterministic*: the model knows exactly which tiles are lying (all of them) and can rule out the displayed state with certainty.

By contrast, in Fibble$_3$ or Fibble$_4$, the model does not know *which* tiles lie. The number of possible truth/lie configurations per row is $\binom{5}{k}$: 10 at $k = 2$–3, 5 at $k = 4$, and **1** at $k = 5$. Each configuration branches further because a lying tile could be either of two alternative states. This combinatorial uncertainty — not the raw number of lies — is what makes partial deception so devastating. At $k = 5$, that uncertainty vanishes: there is only one configuration (all lie), and each lie still provides a definite exclusion.

5.4 Human Baseline Comparison

No peer-reviewed study of human Wordle performance exists. However, an informal analysis of over 266,000 self-reported games suggests human players achieve $\sim$**99%** win rate with an average of **3.80 guesses**.[2] By comparison, LLMs average only 41.1% on standard Wordle and require **4.49 guesses** among solved games — nearly one full guess more than humans. No human data exists for any Fibble$_k$ variant. The LLM–human gap even on truthful Wordle establishes that the deception gradient results represent degradation from an already-subhuman starting point.

5.5 Robustness Does Not Correlate with Standard Performance

Strong Wordle performance poorly predicts Fibble robustness. **o3** achieves 83.3% on Wordle but 0% on every Fibble variant. **Gemini 3 Flash** (100% Wordle) maintains 70% on Fibble$_1$, while **Claude Sonnet 4.6** (56.7% Wordle) collapses to 0%. **GPT-5.1** (30% Wordle) scores *below* the smaller GPT-4o (40.8%) on both Wordle and Fibble$_1$. The most deception-robust models (Gemini 3.1 Pro, GLM-5, Kimi K2.5) form a distinct cluster that does not align with standard capability rankings, suggesting that Fibble success requires *meta-reasoning* — reasoning about the reliability of evidence itself — beyond the pattern-matching that suffices for Wordle. At the other extreme, **SmolLM3 3B** — the latest in HuggingFace's small-language-model line — achieves 0% on both Wordle and Fibble$_1$, confirming that current sub-4B models lack the reasoning capacity for constrained word games.

6 Analysis

6.1 Why Do Models Fail Under Deception?

We identify three failure modes from examining model outputs:

Implicit Trust. Most models treat all feedback as ground truth, applying standard constraint elimination. When one tile lies, the resulting constraint set becomes inconsistent, but models rarely detect this inconsistency.

[2] https://engaging-data.com/wordle-guess-distribution/, accessed March 2026. Self-reported data likely biased toward engaged players; actual population-level performance may be lower.

Model	n	Wordle W%	Wordle n	F_1 W%	F_1 n	F_2 W%	F_2 n	F_3 W%	F_3 n	F_4 W%	F_4 n	F_5 W%	F_5 n
Frontier API Models													
Gemini 3.1 Pro	235	**95.0**	40	**87.5**	40	**31.6**	38	0.0	39	5.1	39	10.3	39
Gemini 3 Flash	130	**100**	30	70.0	30	6.7	30	3.3	30	–	–	30.0	10
Gemini 2.5 Pro	207	90.0	40	25.0	40	0.0	31	0.0	32	0.0	32	3.1	32
o3	180	83.3	30	0.0	30	0.0	30	0.0	30	0.0	30	0.0	30
GPT-5.1	180	30.0	30	10.0	30	0.0	30	0.0	30	0.0	30	3.3	30
Claude Sonnet 4.6	180	56.7	30	0.0	30	0.0	30	0.0	30	0.0	30	0.0	30
Strong API Models													
GLM-5	51	92.3	13	66.7	12	28.6	7	0.0	4	0.0	4	**63.6**	11
Kimi K2.5	51	92.3	13	75.0	12	14.3	7	0.0	4	0.0	4	54.5	11
GPT-5	237	39.0	41	2.4	41	0.0	38	0.0	39	0.0	39	2.6	39
GPT-5 Mini	237	51.2	41	7.3	41	0.0	38	0.0	39	0.0	39	0.0	39
DeepSeek-V3	253	44.9	49	10.2	49	0.0	38	0.0	39	0.0	39	0.0	39
Efficient API Models													
GPT-4o	253	40.8	49	12.2	49	0.0	38	0.0	39	0.0	39	0.0	39
Claude Haiku 4.5	181	48.4	31	30.0	30	0.0	30	0.0	30	0.0	30	0.0	30
GPT-4o Mini	73	15.8	19	5.3	19	0.0	8	0.0	9	0.0	9	0.0	9
o4 Mini	180	0.0	30	0.0	30	0.0	30	0.0	30	0.0	30	0.0	30
Open-Weight Models (Ollama)													
Qwen3.5 122B	76	75.0	32	42.1	19	16.7	6	0.0	4	0.0	4	18.2	11
Qwen3.5 Cloud	48	66.7	12	63.6	11	0.0	6	0.0	4	0.0	4	9.1	11
Qwen3 30B	48	75.0	12	45.5	11	0.0	6	0.0	4	0.0	4	63.6	11
Qwen3 14B	48	91.7	12	9.1	11	0.0	6	0.0	4	0.0	4	27.3	11
Qwen3 8B	49	58.3	12	0.0	11	0.0	7	0.0	4	0.0	4	27.3	11
Llama 4	49	8.3	12	0.0	11	14.3	7	0.0	4	0.0	4	0.0	11
Gemma3 27B	49	16.7	12	18.2	11	0.0	7	0.0	4	0.0	4	9.1	11
SmolLM3 3B	11	0.0	7	0.0	4	–	–	–	–	–	–	–	–
Arena Averages (models with ≥10 games)													
Mean		41.1	35	18.7	35	2.9	13	0.3	13	0.4	12	9.5	34

Table 3: Batch evaluation results. W% = win rate; n = games played in that arena. Total column n gives total games across all arenas. Arena averages computed over models with ≥10 games. "–" = arena not tested for that model. Bold marks best per column. The deception cliff at F_2–F_3 and the F_5 recovery (under total deception) are clearly visible.

Confirmation Bias. Models tend to fixate on early guesses' feedback, building hypotheses that are difficult to revise when later feedback contradicts them. This is exacerbated by lies, which inject false constraints early.

Combinatorial Explosion. Even when models are told that k tiles lie, they rarely enumerate the $\binom{5}{k}$ possible truth configurations. This hypothesis space grows rapidly and exceeds the practical reasoning budget of most models.

6.2 What Makes the Robust Models Different?

Three models stand out as uniquely robust across deception levels: Gemini 3.1 Pro (the most consistent across $k = 0$–5), GLM-5 (63.6% at $k = 5$), and Kimi K2.5 (54.5% at $k = 5$). We hypothesize, based on informal inspection of game logs, that these models engage in *cross-row consistency checking* — looking for letter constraints that remain stable across multiple guesses — rather than treating each row's feedback independently. We have not yet performed a quantitative analysis of reasoning traces to validate this hypothesis; doing so (e.g., counting cross-row references in chain-of-thought outputs) is left to future work.

Notably, the **OpenAI reasoning models** (o3, o4 Mini) perform *worst* under deception despite being designed for multi-step reasoning. o3 achieves 83.3% on Wordle but 0% on every Fibble variant; o4 Mini achieves 0% everywhere. This suggests that their chain-of-thought reasoning may be particularly vulnerable to deceptive premises: once a false constraint enters the reasoning chain, it propagates and compounds.

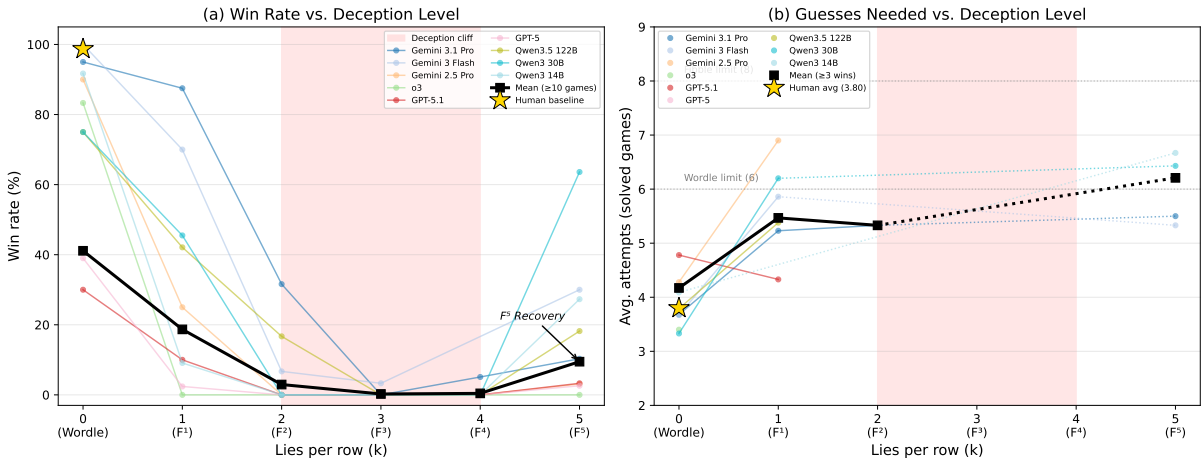

Figure 1: (a) Win rate and (b) average attempts (solved games only) as a function of lies per row. Colored lines show 13 representative models; black shows the mean across all models with ≥10 games. The **gold star** marks the human baseline at $k = 0$: ∼99% win rate in 3.80 guesses on average (see §5.4). The shaded region marks the deception cliff ($k = 2$–4). Panel (b) shows that models require progressively more guesses as deception increases, whereas humans solve standard Wordle in under 4 attempts. No human data exists for Fibble variants.

6.3 Latency and the Cost of Reasoning

Models vary enormously in per-call latency, from under 2 seconds (GPT-4o Mini, Qwen3 8B nothink) to over 1,000 seconds (Qwen3 32B on Fibble$_1$, Kimi K2.5 on Fibble$_2$). Figure 2 plots latency against win rate for all models with ≥10 games on both Wordle and Fibble$_1$. A key question is whether extra reasoning time pays off under deception.

The results reveal that **reasoning time alone does not predict deception robustness** (Figure 2). o3 spends 217s per call yet achieves 0% on every Fibble variant. GPT-5 takes 112–266s per call but wins only 2.4% on Fibble$_1$. Meanwhile, the fast GPT-4o (2.8s) outperforms GPT-5.1 (5.2s) on both Wordle and Fibble$_1$, and Haiku 4.5 (4.1s) achieves 30% on Fibble$_1$ — better than any OpenAI model except the non-reasoning GPT-4o. Among the slow reasoning models, only Gemini 3.1 Pro, GLM-5, and Kimi K2.5 convert their extended computation into genuine deception robustness.

6.4 Thinking vs. No-Thinking Variants

For Qwen3 and Gemma3 models, we tested both default (thinking-enabled) and `nothink` (thinking-disabled) variants, providing a controlled comparison of reasoning overhead (Figure 3).

Thinking mode provides large Wordle improvements (e.g., Qwen3.5 122B: 75.0% vs. 13.3%) at **60–130× the latency**. However, the benefits diminish sharply under deception: at Fibble$_1$, thinking still helps for Qwen3.5 122B (42.1% vs. 6.7%) but not for Qwen3 8B (0% either way). At Fibble$_2$ and beyond, nothink variants score 0% uniformly — but so do most thinking variants. The exception is Fibble$_5$, where Qwen3 8B's thinking mode achieves 27.3% while nothink scores 0%, suggesting that chain-of-thought reasoning *can* help with the rule-inversion task of total deception.

Gemma3 27B is an interesting outlier: thinking adds negligible latency (3.4s vs. 3.1s) and has mixed effects on accuracy, suggesting its "thinking" mode involves minimal additional computation compared to the Qwen3 family.

6.5 The Fibble$_5$ Recovery in Detail

As noted in §5.3, the Fibble$_5$ recovery (avg. 9.5% vs. 0.3–0.4% at $k = 3$–4) arises because knowing *all* tiles lie restores deterministic exclusion information (∼37% of Wordle's bits per row). The top performers — GLM-5 and Qwen3 30B (63.6%), Kimi K2.5 (54.5%) — are *not* the same models that dominate partial deception. Gemini 3.1 Pro leads at $k = 1$–2 but achieves only 10.3% at $k = 5$, while Qwen3 30B scores 0% at $k = 2$–4. This suggests deception robustness comprises at least two distinct skills: *hypothesis management under uncertainty* (partial deception) and *rule inversion under certainty* (total deception).

7 Discussion

Implications for Evaluation Methodology. Our results challenge the assumption that game-based evaluations with truthful feedback adequately mea-

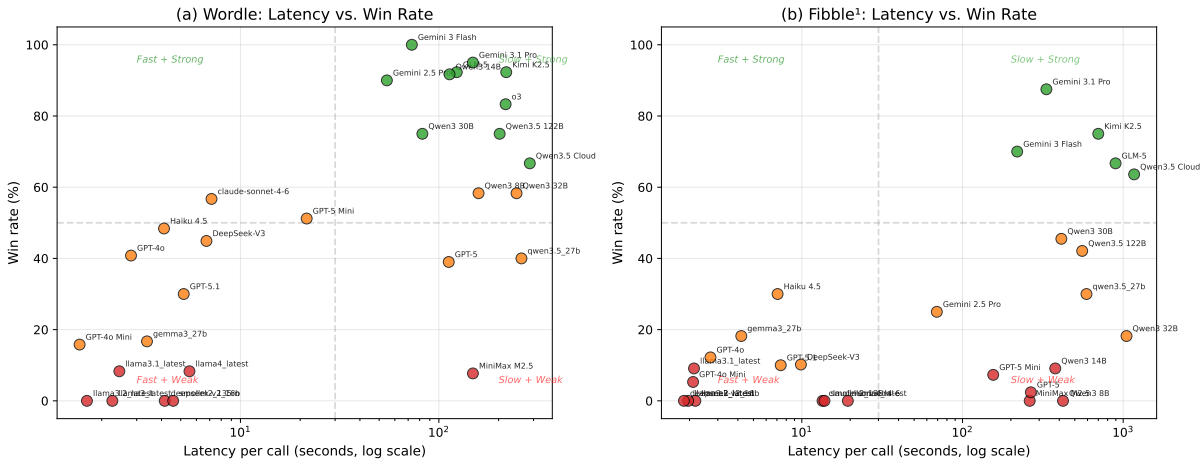

Figure 2: Latency vs. win rate on (a) Wordle and (b) Fibble$_1$ (log scale). Each point is a model; color indicates performance tier (green $\geq$60%, orange $\geq$10%, red $<$10%). Dashed lines at 30s and 50% divide the space into quadrants. On Wordle (a), reasoning models (upper-right) generally outperform fast models (left). On Fibble$_1$ (b), only Gemini 3.1 Pro, Kimi K2.5, and GLM-5 remain in the "slow + strong" quadrant — o3 and GPT-5 fall to the "slow + weak" corner despite heavy computation. Detailed latency data in Table 6 (Appendix).

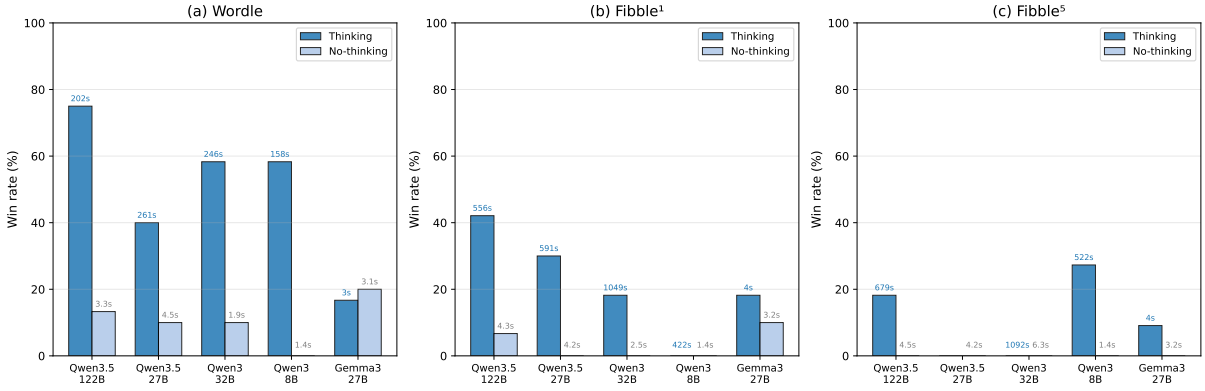

Figure 3: Thinking vs. no-thinking variants across three arenas. Bar heights show win rates; latency labels above each bar show per-call time. Thinking mode is 60–130$\times$ slower for Qwen3 models but provides large Wordle gains (a). Under deception (b,c), gains shrink dramatically: most nothink variants score 0%, but so do most thinking variants at Fibble$_1$. The exception is Fibble$_5$ (c), where Qwen3 8B's thinking mode achieves 27.3% (vs. 0% nothink), suggesting chain-of-thought can help with rule inversion under total deception. Detailed data in Table 7 (Appendix).

sure reasoning ability. WordleArenas reveal a *reasoning robustness gap*: the difference between performance under ideal conditions and performance under adversarial feedback. We argue that evaluation suites should include deception-aware variants to avoid overestimating model capabilities.

Real-World Relevance. Reasoning under unreliable feedback is ubiquitous: diagnostic systems must handle false-positive test results, retrieval-augmented generation must cope with incorrect retrieved documents, and multi-agent systems must handle potentially deceptive partners. Fibble provides a controlled proxy for these scenarios.

Limitations. WordleArenas evaluate deception robustness in a narrow domain (5-letter word guessing). The deception is *systematic* (exactly k lies per row) rather than adversarially optimized. Game counts vary across models (4–49 per arena), with local Ollama models having fewer games than API models; win rates for models with $<$10 games should be interpreted cautiously. Because the model subset contributing to each arena's mean shifts across k (open-weight models have only 4 games at $k = 3$–4), the precise shape of the cliff trajectory should be read as approximate; the qualitative pattern (near-zero win rates in the cliff regime, and a recovery at $k = 5$) is robust across both API and open-weight subsets. We evaluate a snapshot

of models that will rapidly evolve. The daily arena mitigates the last concern by providing continuous evaluation.

Future Work. A key confound is the **guess budget**: an information-theoretic analysis (Table 4, Appendix) shows that usable bits per row fall from 7.92 (Wordle) to just 1.60 at $k = 3$–4, leaving only a +1.3-guess margin above the theoretical minimum. The near-zero win rates at $k = 3$–4 may partly reflect insufficient attempts rather than pure reasoning failure; we recommend 13 guesses for those arenas to provide a fair margin. Future work will also explore adaptive lie injection, strategic prompting for deception awareness, evaluation of retrieval-augmented and tool-using agents, and broader coverage of small open-source models such as IBM's Granite family, which were not included in this round.

8 Conclusion

We presented WordleArenas, a parametric deception benchmark that evaluates 41 models across 3,749 games and 6 deception levels. Our key findings are: (1) LLMs are **brittle to deception**, with average win rates dropping from 41.1% to 18.7% with just one lie per row; (2) a **deception cliff** at $k = 2$–3 where averages collapse to $\leq$2.9%; (3) standard performance **does not predict** deception robustness (o3: 83.3% Wordle, 0% all Fibble); (4) a **Fibble$_5$ recovery** where total deception yields higher win rates than partial deception, because certainty about lying restores usable information; and (5) deception robustness comprises **at least two distinct skills** — hypothesis management under uncertainty and rule inversion under certainty. As LLMs are deployed in adversarial environments, benchmarks assuming truthful feedback will systematically overestimate their capabilities.

Acknowledgments

Claude Code (Anthropic) was used as a coding assistant throughout this project, including implementation of the benchmark platform, execution of batch experiments, data analysis, figure generation, and drafting of this manuscript.

References

Anne Beyer, Kranti Chalamalasetti, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2024. clembench-2024: A challenging, dynamic, complementary, multilingual benchmark and underlying flexible framework for LLMs as multi-action agents. *arXiv preprint arXiv:2405.20859*.

Filippo Momentè, Alessandro Suglia, Mario Giulianelli, Ambra Ferrari, Alexander Koller, Oliver Lemon, David Schlangen, Raquel Fernández, and Raffaella Bernardi. 2025. Triangulating LLM progress through benchmarks, games, and cognitive tests. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Suma Bailis, Jane Friedhoff, and Feiyang Chen. 2024. Werewolf Arena: A case study in LLM evaluation via social deduction. *arXiv preprint arXiv:2407.13943*.

Zirui Song, Yuan Huang, Junchang Liu, Haozhe Luo, Chenxi Wang, Lang Gao, Zixiang Xu, Mingfei Han, Xiaojun Chang, and Xiuying Chen. 2025. Beyond survival: Evaluating LLMs in social deduction games with human-aligned strategies. *arXiv preprint arXiv:2510.11389*.

Mrinal Agarwal, Saad Rana, Theo Sundoro, Hermela Berhe, Spencer Kim, Vasu Sharma, Sean O'Brien, and Kevin Zhu. 2025. WOLF: Werewolf-based observations for LLM deception and falsehoods. *arXiv preprint arXiv:2512.09187*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, et al. 2023. PromptRobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, pages 57–68.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403.

Krerkkiat Chusap, Colin Murphy, Clay Hess, Sai Deepa Kadaru, Rohit Buccapatnam, and Chang Liu. 2025. RL- and LLM-based AI solvers for the game of Wordle/Fibble. In *2025 IEEE Conference on Games (CoG)*, pages 1–8. IEEE.

A System Prompt

The following system prompt is used for Fibble$_k$ (shown for $k = 4$):

```
You are playing Fibble4, an extreme variant of Fibble (deceptive Wordle). The rules are: You
must guess a secret 5-letter English word. You have 8 attempts. After each guess, you'll
receive feedback on each letter: GREEN: correct letter and position. YELLOW: correct letter,
wrong position. GRAY: letter not in word. CRITICAL TWIST: In every row of feedback, EXACTLY
FOUR of the five clues are LIES. Only ONE tile is truthful. Respond with ONLY a single
5-letter English word in uppercase.
```

B Supplementary Tables and Data

This appendix collects the supplementary data underlying the analyses in Sections 5 and 6. Table 4 reports the information-theoretic budget at each deception level (referenced in the guess-budget discussion of Section 7). Table 5 shows daily-arena pilot win rates and average attempts for a subset of frontier models over February 9 – March 1, 2026. Table 6 summarizes per-call latency for representative fast and reasoning models (visualized in Figure 2), and Table 7 reports thinking-vs.-no-thinking results for the Qwen3 and Gemma3 families (visualized in Figure 3).

The daily arenas themselves run automatically via GitHub Actions at 14:00 UTC. Each arena fetches the day's word (NYT API for Wordle; deterministic hash for Fibble variants), plays all configured models, and commits results to a public GitHub Pages site. The leaderboard, game replays, and cross-arena rankings are available at the project website: https://drchangliu.github.io/WordleArenas/.

Arena	Configs	Uncert.	Info/row	Min g	Margin
Wordle ($k=0$)	1	0.0 b	7.92 b	1.4	+4.6
F_1 ($k=1$)	10	3.3 b	4.60 b	2.3	+5.7
F_2 ($k=2$)	40	5.3 b	2.60 b	4.1	+3.9
F_3 ($k=3$)	80	6.3 b	1.60 b	6.7	+1.3
F_4 ($k=4$)	80	6.3 b	1.60 b	6.7	+1.3
F_5 ($k=5$)	32	5.0 b	2.92 b	3.7	+4.3

Table 4: Information budget analysis. "Configs" = $\binom{5}{k} \times 2^k$ possible true-feedback vectors per row. "Info/row" = usable bits after subtracting configuration uncertainty from the 7.92-bit observation. "Min g" = $\lceil 10.7/\text{Info} \rceil$, the theoretical minimum guesses to identify a word from 1,674 candidates. "Margin" = current budget (6 for Wordle, 8 for Fibble) minus Min g.

Model	Wordle		Fibble$_1$		Fibble$_2$		Fibble$_3$		Fibble$_4$		Fibble$_5$	
	Win	Att	Win	Att	Win	Att	Win	Att	Win	Att	Win	Att
GPT-5	72.7	4.82	9.1	7.91	0	–	0	–	0	–	12.5	7.25
GPT-5 Mini	100	3.73	9.1	7.55	0	–	0	–	0	–	0	–
GPT-4o	36.8	5.63	15.8	7.53	0	–	0	–	0	–	0	–
GPT-4o Mini	15.8	6.0	5.3	7.95	0	–	0	–	0	–	0	–
Gemini 3.1 Pro	100	4.17	100	5.67	100	5.33	0	–	66.7	7.67	60.0	6.2
Gemini 2.5 Pro	100	4.70	80.0	7.40	–	–	–	–	–	–	50.0	6.5
DeepSeek-V3	42.1	5.37	5.3	7.74	0	–	0	–	0	–	0	–

Table 5: Daily arena pilot results (selected models, February 9 – March 1, 2026). "Win" = win rate (%), "Att" = average attempts among solved games. "–" indicates 0 wins (no attempts to average). Full results for all 25 daily-arena models are available online.

Model	Wordle		Fibble_1	
	Lat (s)	Win%	Lat (s)	Win%
Fast models (<10s/call)				
GPT-4o	2.8	40.8	2.7	12.2
GPT-4o Mini	1.5	15.8	2.1	5.3
GPT-5.1	5.2	30.0	7.4	10.0
Sonnet 4.6	7.1	56.7	13.5	0.0
Haiku 4.5	4.1	48.4	7.1	30.0
Reasoning models (>50s/call)				
o3	217	83.3	–	0.0
GPT-5	112	39.0	266	2.4
GPT-5 Mini	22	51.2	155	7.3
Gemini 3.1 Pro	148	95.0	333	87.5
Gemini 3 Flash	73	100	219	70.0
GLM-5	123	92.3	896	66.7
Kimi K2.5	219	92.3	700	75.0

Table 6: Per-call latency (seconds) vs. win rate for representative fast and reasoning models. Visualized in Figure 2.

Model	Mode	Lat (s)	W%	F_1%	F_5%
Qwen3.5 122B	think	202	75.0	42.1	18.2
	nothink	3.3	13.3	6.7	0.0
Qwen3.5 27B	think	261	40.0	30.0	0.0
	nothink	4.5	10.0	0.0	0.0
Qwen3 32B	think	246	58.3	18.2	0.0
	nothink	1.9	10.0	0.0	0.0
Qwen3 8B	think	159	58.3	0.0	27.3
	nothink	1.4	0.0	0.0	0.0
Gemma3 27B	think	3.4	16.7	18.2	9.1
	nothink	3.1	20.0	10.0	0.0

Table 7: Thinking vs. no-thinking variants. Latency is per-call average on Wordle (seconds). Visualized in Figure 3.