LOGICAL CONSISTENCY UNDER PRESSURE: PROBING AND REPAIRING CROSS-QUERY CONTRADICTIONS IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models answer individual logic questions with reasonable accuracy, yet frequently contradict themselves across logically related queries, affirming a conditional while denying its contrapositive, or endorsing a transitive chain while rejecting the implied conclusion. We introduce CONSISTENCYBENCH, a benchmark of 493 logically entailed question sets (1,904 questions) spanning six categories of formal and commonsense reasoning, designed to measure *cross-query logical consistency*. We evaluate eighteen frontier LLMs, including GPT-5.2, GPT-4.1, Claude Opus 4.6, Gemini 2.5 Pro, DeepSeek-R1, o3, and Qwen 2.5 72B, and find that even the strongest model (GPT-4.1) achieves only 46.7% set-level consistency despite 83.0% individual accuracy, revealing consistency gaps of **36–57 percentage points** across all models tested. We propose *Consistency-Guided Decoding* (CGD), a training-free, model-agnostic inference-time method that detects and repairs cross-query contradictions via NLI-based checking. Across 17 models, CGD improves set-level consistency by **+6.6pp** on average (up to **+19.7pp** for GPT-4o), while simultaneously improving individual accuracy by +2.8pp on average, demonstrating that cross-query consistency is a tractable target for inference-time intervention.

1 INTRODUCTION

Logical reasoning is fundamental to reliable AI systems. Real-world reasoning requires not just answering individual questions correctly, but maintaining *logical coherence across related inferences*. A physician who diagnoses hypertension but denies the patient has elevated blood pressure has failed not on knowledge but on consistency. Similarly, an LLM-powered legal assistant that affirms "if the contract is breached, damages are owed" but denies "since damages are not owed, the contract was not breached" makes logically contradictory claims despite both questions arising from the same premise.

Large language models (LLMs) have achieved impressive reasoning performance on isolated benchmarks (Wei et al., 2022; Wang et al., 2023; OpenAI, 2024b), yet their cross-query consistency remains poorly understood. The "reversal curse" (Berglund et al., 2023) demonstrated that models trained on "A is B" fail to generalize to "B is A." We show that this problem extends far beyond pairwise reversal: LLMs *systematically* contradict themselves across diverse logical relationships including contrapositive equivalence, transitivity, syllogistic reasoning, negation, modus tollens, and commonsense entailment (Figure 1).

We make three contributions: **(1)** CONSISTENCYBENCH, a benchmark of 493 question sets (1,904 questions) spanning six logical reasoning categories, with three hierarchical consistency metrics (§3); **(2)** a systematic evaluation of 18 frontier LLMs revealing that all models exhibit consistency gaps of 36–57 percentage points between individual accuracy and set-level consistency, a finding that holds across model families, scales, and training paradigms including reasoning-specialized models (§5); and **(3)** Consistency-Guided Decoding (CGD), a training-free, model-agnostic inference-time method that improves consistency in 16 out of 17 models tested, with a mean improvement of +6.6pp and a maximum of +19.7pp (§4).

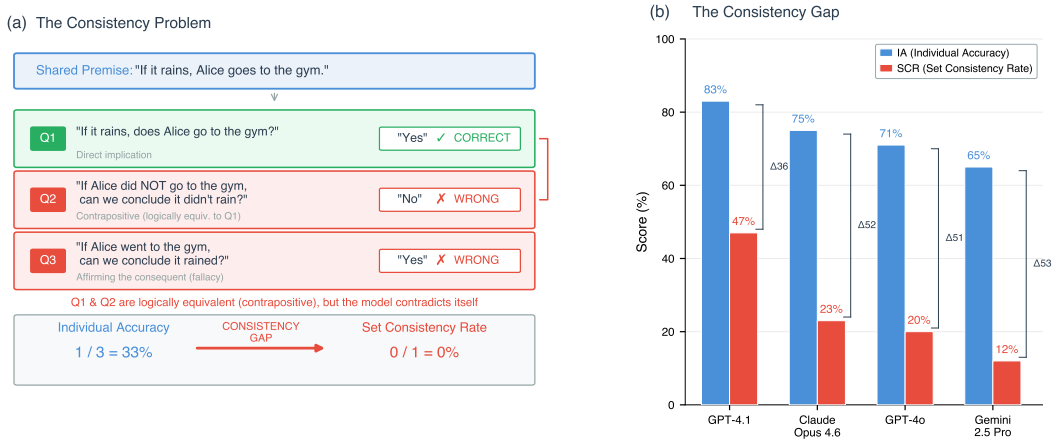

Figure 1: **The cross-query consistency problem. (a)** A model correctly answers Q1 (direct implication) but contradicts itself on Q2 (the logically equivalent contrapositive) and incorrectly affirms Q3 (the invalid "affirming the consequent" fallacy). Despite 33% individual accuracy, the set-level consistency rate is 0%. **(b)** The consistency gap across four representative models: individual accuracy (IA) far exceeds set-level consistency (SCR), with gaps of 36–53 percentage points.

2 Related Work

LLM Logical Reasoning. Recent benchmarks evaluate LLMs on propositional logic (Saparov & He, 2023), first-order logic (Han et al., 2022), and syllogistic reasoning (Dasgupta et al., 2022), but focus on single-query accuracy. These benchmarks measure whether a model can answer a logic question correctly in isolation. We measure a complementary and distinct property: *whether a model maintains consistency across logically related queries derived from the same premises*.

Consistency Failures. Berglund et al. (2023) exposed the reversal curse, showing that training on "A is B" does not yield "B is A." Raj et al. (2023) studied semantic consistency under paraphrasing, measuring whether models give equivalent answers to rephrased questions. Shi et al. (2023) showed sensitivity to irrelevant context. Our work extends the consistency analysis to multi-step logical relationships at scale, testing six distinct categories of logical entailment across 18 models.

Inference-Time Reasoning. Chain-of-thought prompting (Wei et al., 2022) and self-consistency decoding (Wang et al., 2023) improve single-query accuracy but do not enforce cross-query coherence, since each question is processed independently. Maieutic prompting (Jung et al., 2022) detects self-contradictions within a single response but requires expensive abductive/deductive candidate generation. Neuro-symbolic approaches (Pan et al., 2023; Xu et al., 2024) require structured formal representations. In contrast, CGD operates entirely in natural language, is training-free, and works with black-box APIs without access to model internals.

3 ConsistencyBench: A Benchmark for Cross-Query Logical Consistency

ConsistencyBench contains 493 question sets comprising 1,904 individual questions across six categories (Table 1). Each set shares a common premise and contains 3–4 questions that test logically related inferences. Ground-truth answers are derived from classical propositional and first-order logic for the formal categories and from manually curated entailment rules for commonsense reasoning. Questions are generated programmatically from curated pools of 24 person names, 20 factual statements, 10 activity pairs, 8 condition pairs, 10 abstract categories, and 70 commonsense entailment rules, ensuring diversity while maintaining logical soundness.

Metrics. We define three hierarchical metrics with increasing strictness. **Individual Accuracy (IA)** measures the fraction of individual questions answered correctly. **Pairwise Consistency Rate**

Table 1: The six categories of CONSISTENCYBENCH. Each category tests a specific logical pattern and includes a question designed to detect a common fallacy.

Category	Pattern	#Sets	Fallacy Tested
Contrapositive	$P \to Q \equiv \neg Q \to \neg P$	85	Affirm. consequent
Transitivity	$A \to B, B \to C \vdash A \to C$	85	Chain break
Syllogistic	$\forall x : A(x) \to B(x)$	85	Converse error
Negation	P vs. $\neg P$	85	Neg. incoherence
Modus Tollens	$P \to Q, \neg Q \vdash \neg P$	85	Deny antecedent
Commonsense	Everyday entailment	68	Reverse entailment
Total		**493**	

Algorithm 1 Consistency-Guided Decoding (CGD)

Require: Question set $\mathcal{Q} = \{q_1, \ldots, q_n\}$, premise P, LLM $\mathcal{M}$, NLI checker $\mathcal{C}$
history $\leftarrow$ []
for $i = 1$ to n **do**
  $a_i \leftarrow \mathcal{M}(q_i, P)$ {Generate}
  for $(q_j, a_j) \in$ history **do**
    if $\mathcal{C}(a_j, a_i) =$ CONTRADICTION **then**
      $a_i \leftarrow \mathcal{M}(\text{REVISE}(q_i, q_j, a_j, P))$ {Repair}
      break
    end if
  end for
  history.append((q_i, a_i))
end for
return $\{a_1, \ldots, a_n\}$

(PCR) measures the fraction of within-set question pairs where both are answered correctly. **Set-Level Consistency Rate (SCR)** measures the fraction of sets where *all* questions are answered correctly. The *consistency gap*, defined as IA $-$ SCR, quantifies how much single-query accuracy metrics overestimate a model's true logical reasoning capability.

4 CONSISTENCY-GUIDED DECODING

We propose CGD, a training-free inference-time method for improving cross-query logical consistency. For each question q_i in a set, CGD operates in three steps (Algorithm 1):

(1) Generate: The target LLM $\mathcal{M}$ produces an initial answer a_i for question q_i. **(2) Check**: An NLI contradiction checker $\mathcal{C}$ (GPT-4o-mini) tests whether a_i contradicts any prior answer a_j in the answer history. **(3) Repair**: If a contradiction is detected, a revision prompt that includes the conflicting prior answer a_j, the corresponding question q_j, and the shared premise is sent to $\mathcal{M}$ to elicit a corrected answer a_i'.

CGD is model-agnostic (works with any LLM including black-box APIs), training-free (requires no fine-tuning or parameter updates), and efficient (adds at most $O(n)$ NLI checks per question, where n is the set size, typically 3–4). Figure 2 in Appendix A illustrates the full pipeline.

5 EXPERIMENTS

5.1 SETUP

Models. We evaluate 18 frontier LLMs spanning six model families: **OpenAI**: GPT-5.2, GPT-5, GPT-5 Mini, GPT-4.1, GPT-4o (OpenAI, 2024a), o3 (OpenAI, 2025b;a); **Anthropic**: Claude Opus 4.6, Sonnet 4.6, Sonnet 4.5, 3.5 Sonnet (Anthropic, 2025; 2024); **Google**: Gemini 2.5 Pro, 2.5 Flash, 2.0 Flash (Google DeepMind, 2025; 2024); **DeepSeek**: V3.2, R1 (DeepSeek-AI, 2024; 2025); **Meta**: Llama 3.3 70B, 3.1 70B (Meta AI, 2024; Dubey et al., 2024); **Qwen**: 2.5 72B (Qwen

Table 2: Direct prompting vs. CGD on CONSISTENCYBENCH (300 sets, 1,150 questions). Gap = IA − SCR. ΔSCR = absolute SCR change with CGD. Models sorted by vanilla SCR. Best SCR per model in **bold**. Full per-category results in Appendix E.

	Direct			CGD (Ours)			
Model	**IA**	**SCR**	**Gap**	**IA**	**SCR**	**Gap**	**ΔSCR**
GPT-4.1	83.0	46.7	36.4	85.2	**54.0**	31.2	+7.3
Qwen 2.5 72B	75.0	32.7	42.3	73.7	**33.0**	40.7	+0.3
Cl. Opus 4.6	74.8	23.3	51.4	79.1	**32.3**	46.8	+9.0
DeepSeek-R1	70.2	**21.7**	48.5	–	–	–	–
GPT-4o	70.8	20.0	50.8	79.0	**39.7**	39.4	+19.7
Llama 3.1 70B	65.3	**19.0**	46.3	61.6	17.0	44.6	−2.0
Gemini 2.5 Flash	67.6	17.7	49.9	69.2	**22.3**	46.9	+4.7
Cl. 3.5 Sonnet	71.6	17.3	54.2	73.5	**22.0**	51.5	+4.7
Gemini 2.5 Pro	65.4	12.3	53.1	67.7	**21.7**	46.0	+9.3
GPT-5.2	69.3	12.0	57.3	73.2	**23.7**	49.6	+11.7
o3	67.2	12.0	55.2	70.3	**21.0**	49.3	+9.0
Llama 3.3 70B	56.8	10.7	46.1	57.0	**13.3**	43.6	+2.7
Cl. Sonnet 4.5	51.9	10.7	41.2	54.7	**12.0**	42.7	+1.3
Cl. Sonnet 4.6	57.5	10.3	47.1	67.0	**24.3**	42.6	+14.0
DeepSeek V3.2	59.2	10.0	49.2	61.7	**19.3**	42.4	+9.3
GPT-5	62.9	9.3	53.5	63.6	**11.3**	52.2	+2.0
Gemini 2.0 Flash	44.7	7.3	37.4	55.7	**12.3**	43.4	+5.0
GPT-5 Mini	58.4	4.0	54.4	57.4	**9.0**	48.4	+5.0
Mean	65.1	16.5	48.6	67.6	22.8	44.8	+6.6

Team, 2024). This selection spans multiple scales, architectures, and training paradigms, including reasoning-specialized models (o3, DeepSeek-R1).

Evaluation protocol. All 18 models are evaluated under **Direct** (vanilla) prompting with concise yes/no/cannot-determine answers; 17 models under CGD (DeepSeek-R1 excluded due to API constraints); and 2 models under **Chain-of-Thought (CoT)** prompting (GPT-4o, Qwen 2.5 72B). Evaluation uses a stratified sample of 300 sets (50 per category, 1,150 questions total).

5.2 MAIN RESULTS

Table 2 presents the central finding: *every model exhibits a large gap between individual accuracy and set-level consistency*, and CGD improves SCR for 16 of 17 models tested.

5.3 KEY FINDINGS

(1) The consistency gap is universal. All 18 models exhibit IA $>$ SCR, with gaps ranging from 36.4pp (GPT-4.1) to 57.3pp (GPT-5.2). Even the best-performing model, GPT-4.1, achieves only 46.7% SCR despite 83.0% IA. The mean gap across all models is 48.6pp, indicating that standard accuracy metrics dramatically overestimate logical reasoning capability.

(2) Scaling and reasoning tokens do not solve consistency. GPT-5.2, one of the newest models, reaches only 12.0% SCR (gap: 57.3pp). The reasoning-specialized o3 also reaches only 12.0% SCR despite extended inference-time computation. GPT-5 achieves 9.3% SCR, and GPT-5 Mini only 4.0%. These results suggest that scale and reasoning-time compute address different failure modes than cross-query consistency.

(3) CGD works broadly. Across 17 models, CGD improves SCR for 16 (94%), with a mean improvement of +6.6pp and a median of +5.0pp. The largest gains are observed for GPT-4o (+19.7pp), Claude Sonnet 4.6 (+14.0pp), GPT-5.2 (+11.7pp), and Gemini 2.5 Pro (+9.3pp). Only Llama 3.1 70B shows a decrease (-2.0pp), suggesting that weaker models may propagate errors through the revision mechanism.

(4) CGD simultaneously improves accuracy. Beyond consistency, CGD also improves IA by +2.8pp on average. The contradiction-aware revision process does not merely enforce agreement across answers but helps the model arrive at more correct answers, suggesting that the knowledge needed for consistency is often latent but not reliably activated during independent generation.

Chain-of-Thought comparison. On GPT-4o, CoT improves IA from 70.8% to 75.0% and SCR from 20.0% to 28.3%, whereas CGD achieves 79.0% IA and 39.7% SCR. On Qwen 2.5 72B, CoT improves IA from 75.0% to 77.1% but *slightly decreases* SCR from 32.7% to 32.3%. This confirms that independent reasoning paths, even when they individually improve accuracy, can still be mutually inconsistent. The full strategy comparison is in Appendix B.

6 Discussion and Limitations

Our results reveal that cross-query logical consistency is a fundamental weakness of current LLMs that is not addressed by scaling, reasoning tokens, or chain-of-thought prompting. The consistency gap persists across all six model families tested, from the latest GPT-5 series to open-weight models like Llama and Qwen, and from general-purpose models to reasoning-specialized ones like o3 and DeepSeek-R1.

CGD demonstrates that lightweight inference-time intervention can substantially narrow this gap. The approach is practical for deployment: it requires no fine-tuning, works with black-box APIs, and the NLI checker (GPT-4o-mini) adds minimal computational cost. The fact that models can often correct themselves when shown a contradiction suggests that consistency knowledge is encoded in the model's parameters but is not reliably activated during independent generation. This "latent consistency" hypothesis points toward promising directions for both training-time and inference-time solutions.

Limitations. ConsistencyBench covers propositional and first-order logic patterns but does not include temporal, probabilistic, modal, or domain-specific reasoning. CGD depends on the NLI checker's quality; false positives could degrade performance, as possibly evidenced by the negative result on Llama 3.1 70B (-2.0pp). The method adds inference latency proportional to the set size. CoT was evaluated on only 2 of 18 models due to cost constraints. All models are accessed via commercial APIs, and results may vary across API versions or updates.

7 Conclusion

We presented ConsistencyBench, a benchmark for measuring cross-query logical consistency across six reasoning categories, and CGD, a training-free inference-time method for improving it. Across 18 frontier LLMs, we found a universal consistency gap of 36–57 percentage points between individual accuracy and set-level consistency. CGD narrows this gap for 16 of 17 models tested (mean +6.6pp, max +19.7pp) through NLI-based contradiction detection and revision, while simultaneously improving individual accuracy. Our findings demonstrate that cross-query logical consistency is both a significant blind spot in current LLM evaluation and a tractable target for inference-time intervention.

References

Anthropic. The Claude model family. *Anthropic Technical Report*, 2024.

Anthropic. The Claude 4 model family. *Anthropic Technical Report*, 2025.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". *arXiv preprint arXiv:2309.12288*, 2023.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Google DeepMind. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2024.

Google DeepMind. Gemini 2.5: Advancing multimodal reasoning. *Google DeepMind Technical Report*, 2025.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. FOLIO: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*, 2022.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

Meta AI. Llama 3.3: Efficient multilingual large language models. *Meta AI Blog*, 2024.

OpenAI. GPT-4o system card. *OpenAI Technical Report*, 2024a. URL .

OpenAI. Learning to reason with LLMs. *OpenAI Blog*, 2024b. URL .

OpenAI. GPT-5 technical report. *OpenAI Technical Report*, 2025a. URL .

OpenAI. Deliberative alignment in o3. *OpenAI Technical Report*, 2025b. URL .

Liangming Pan, Michael Alber, Lajanugen Wan, Wenhu Chen, and Nanyun Peng. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Harsh Raj, Vipul Gupta, Domenic Rosati, and Subhabrata Majumdar. Semantic consistency for assuring reliability of large language models. *arXiv preprint arXiv:2308.09138*, 2023.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *International Conference on Learning Representations*, 2023.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Hanmeng Xu, Zhaoyang Liu, et al. SymbCoT: Symbolic chain-of-thought meets first-order logic. *arXiv preprint arXiv:2405.18357*, 2024.

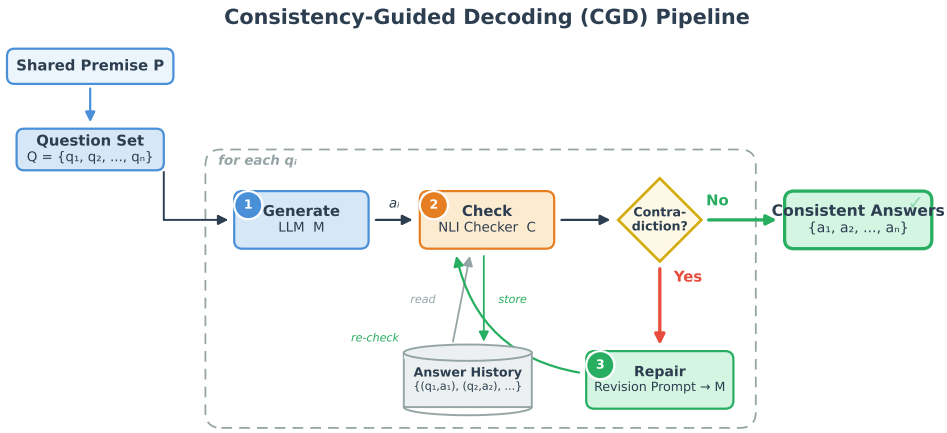

Figure 2: **Consistency-Guided Decoding (CGD) pipeline.** For each question q_i in a set, the LLM generates an initial answer (Step 1), which is checked against all prior answers in the answer history via an NLI contradiction checker (Step 2). If a contradiction is found, a revision prompt containing the conflicting prior answer and the shared premise is sent back to the LLM (Step 3). The revised answer is then re-checked. Consistent answers are stored in the history for future checks.

Table 3: Detailed strategy comparison on representative models. CoT = Chain-of-Thought. Bold indicates best result per model per metric.

Model	Strategy	IA (%)	PCR (%)	SCR (%)
GPT-4o	Direct	70.8	57.0	20.0
	CoT	75.0	63.5	28.3
	CGD (Ours)	**79.0**	**69.5**	**39.7**
Qwen 2.5 72B	Direct	75.0	65.0	32.7
	CoT	**77.1**	65.3	32.3
	CGD (Ours)	73.7	**67.8**	**33.0**
GPT-4.1	Direct	83.0	74.0	46.7
	CGD (Ours)	**85.2**	**77.6**	**54.0**
GPT-5.2	Direct	69.3	52.3	12.0
	CGD (Ours)	**73.2**	**58.4**	**23.7**
o3	Direct	67.2	51.8	12.0
	CGD (Ours)	**70.3**	**56.7**	**21.0**
Cl. Sonnet 4.6	Direct	57.5	47.4	10.3
	CGD (Ours)	**67.0**	**57.7**	**24.3**

A CGD Pipeline Diagram

B Full Strategy Comparison

Table 3 provides a detailed comparison across all evaluated strategies (Direct, CoT, CGD) for models where multiple strategies were tested. The results demonstrate that CGD outperforms both Direct and CoT prompting on set-level consistency across all models tested with multiple strategies.

On GPT-4o, which is the only model evaluated under all three strategies, CGD achieves 39.7% SCR compared to 28.3% for CoT and 20.0% for Direct prompting, representing a 99% relative improvement over Direct and a 40% improvement over CoT. Importantly, CGD also achieves the

highest IA (79.0%) and PCR (69.5%) on this model, demonstrating that consistency-aware revision benefits all metrics simultaneously.

For Qwen 2.5 72B, the results reveal an instructive pattern: CoT improves IA from 75.0% to 77.1% but slightly *decreases* SCR from 32.7% to 32.3%, while CGD improves SCR to 33.0%. This illustrates that independent reasoning improvements (CoT) do not automatically translate to cross-query consistency.

C Per-Category SCR Heatmap

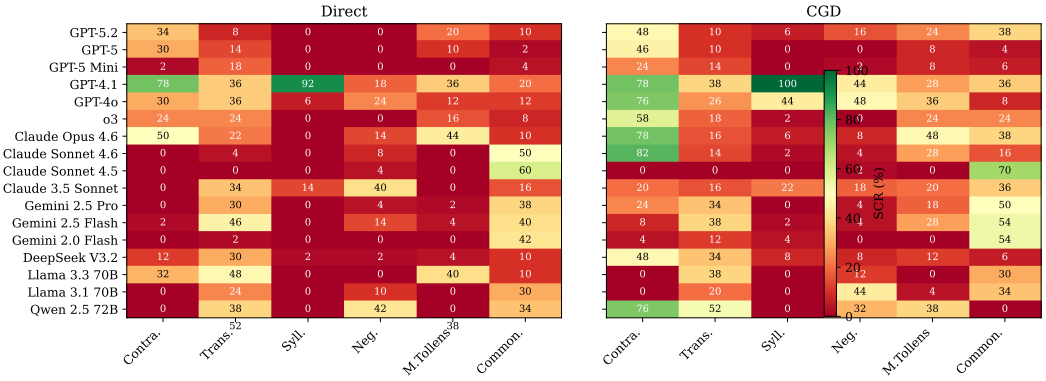

Figure 3: Set-level consistency rate (SCR%) by category for Direct prompting (left) and CGD (right) across all evaluated models. CGD provides improvements across most model-category combinations, with the largest gains on contrapositive and modus tollens categories that specifically test logical equivalence.

The heatmap reveals several patterns. Syllogistic reasoning is generally the hardest category, with many models achieving 0% SCR under Direct prompting. Transitivity and contrapositive categories show moderate scores for the strongest models. Commonsense entailment presents a distinct pattern: some models that perform poorly on formal logic categories achieve relatively higher scores here (e.g., Gemini 2.0 Flash: 42% commonsense vs. 0% on four formal categories), likely because commonsense reasoning patterns are better represented in pre-training data.

D Consistency Gap Visualization

The consistency gap visualization reinforces two key findings. First, the gap is *larger* in absolute terms for models with higher IA, because their higher individual accuracy creates more room for SCR to lag behind. For instance, GPT-5.2 has both the highest gap (57.3pp) and a relatively high IA (69.3%). Second, CGD reduces the gap primarily by improving SCR more than it changes IA, confirming that the method specifically targets the consistency failure mode rather than broadly improving reasoning.

E Complete Per-Category Breakdown

Table 4 provides the complete per-category SCR breakdown for all models evaluated under Direct prompting. Each cell shows the percentage of question sets within that category where the model answered all questions correctly.

Several patterns emerge from this breakdown. GPT-4.1 achieves an exceptional 92% on syllogistic reasoning, far above all other models, yet scores only 18% on negation and 20% on commonsense entailment. Qwen 2.5 72B shows an extreme pattern: 78% on contrapositive and 52% on transitivity, but exactly 0% on syllogistic and commonsense. This highly uneven performance across categories suggests that models acquire logical patterns idiosyncratically rather than developing general logical competence. Gemini 2.0 Flash presents a particularly striking case: it achieves 0% SCR on four of

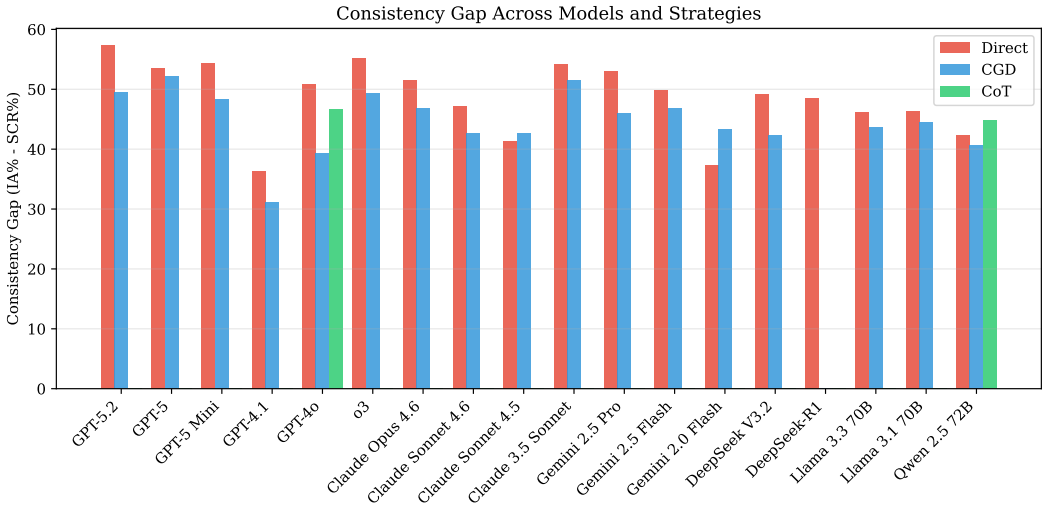

Figure 4: Consistency gap (IA% − SCR%) across all models and strategies. Smaller bars indicate better cross-query consistency. CGD consistently narrows the gap compared to Direct prompting.

Table 4: Per-category set-level consistency rate (SCR%) under Direct prompting for select models. Each category has 50 question sets. Values represent the percentage of sets where all questions were answered correctly.

Model	**Contra.**	**Trans.**	**Syll.**	**Neg.**	**M.Toll.**	**Comm.**
GPT-4.1	78.0	36.0	92.0	18.0	36.0	20.0
Qwen 2.5 72B	78.0	52.0	0.0	28.0	38.0	0.0
Cl. Opus 4.6	50.0	22.0	0.0	14.0	44.0	10.0
GPT-4o	30.0	36.0	6.0	24.0	12.0	12.0
GPT-5.2	34.0	8.0	0.0	0.0	20.0	10.0
o3	24.0	24.0	0.0	0.0	16.0	8.0
Gemini 2.0 Flash	0.0	2.0	0.0	0.0	0.0	42.0

six formal categories but 42% on commonsense entailment, suggesting that its performance is driven by surface pattern matching on commonsense patterns rather than logical reasoning.

F Qualitative Error Analysis

Table 5 shows representative examples of cross-query contradictions from the Direct prompting evaluation.

Common error patterns. We identify three dominant error patterns:

1. **Contrapositive blindness** (most common): Models affirm $P \to Q$ but deny $\neg Q \to \neg P$, treating the contrapositive as a separate and uncertain claim rather than a logical equivalence.
2. **Fallacy acceptance**: Models frequently affirm the consequent ("Q is true, therefore P is true") and deny the antecedent ("not P, therefore not Q"), treating these invalid inferences as valid.
3. **Over-cautious hedging**: On some questions, models output "Cannot be determined" when the answer is logically certain, particularly for contrapositive and transitive inferences.

CGD repair examples. When CGD detects a contradiction (e.g., answering "No" to the contrapositive after answering "Yes" to modus ponens), the revision prompt surfaces the inconsistency. In

Table 5: Representative cross-query contradictions. The model answers Q1 correctly but contradicts itself on Q2, despite Q2 being logically entailed by the same premises.

Category	Question (abbreviated)	Exp.	Pred.
Common.	Q1: Given that "The election results were announced", is it reasonable to conclude t...	Yes	Yes
	Q2: If we know that it is NOT the case that "an election took place", is it possible...	No	N/A
Neg.	Q1: Is the following statement true? "Gravity pulls objects toward the center of the...	Yes	Yes
	Q2: Is the negation of the following statement true? "It is NOT the case that gravit...	No	Yes
M.Tollens	Q1: Given: "If the road is clear, then Victor goes for a walk." We observe that Vict...	Yes	Yes
	Q2: Given: "If the road is clear, then Victor goes for a walk." We observe that Vict...	No	No

the majority of cases, the model corrects itself when shown the contradiction, suggesting that the knowledge of logical equivalence is latent but not reliably activated during independent generation.

G Full ConsistencyBench Statistics

ConsistencyBench contains 493 question sets with 1,904 individual questions. The benchmark is designed for measuring cross-query logical consistency and tests six categories of reasoning. Table 6 summarizes the benchmark statistics.

Table 6: Full ConsistencyBench statistics. The evaluation sample uses a stratified subset of 300 sets (50 per category).

Statistic	Value
Total question sets	493
Total individual questions	1,904
Questions per set	3–4
Number of categories	6
Sets per formal category	85
Sets in commonsense category	68
Evaluation sample: sets	300
Evaluation sample: questions	1,150
Sets per category (eval)	50
Entity pools: person names	24
Entity pools: factual statements	20
Entity pools: activity pairs	10
Entity pools: condition pairs	8
Entity pools: abstract categories	10
Entity pools: commonsense rules	70
Total models evaluated	18
Total evaluation runs	37
Models with Direct results	18
Models with CGD results	17
Models with CoT results	2

Questions are generated programmatically from curated entity pools. For formal categories, the premise structure follows classical logic (e.g., "If P then Q") and questions test the direct inference, contrapositive, converse, and inverse. For commonsense entailment, premises describe everyday sit-

uations (e.g., "The election results were announced") and questions test whether the model correctly identifies entailed conclusions and rejects non-entailed ones. All ground-truth labels are verified by the authors.

H PROMPTING TEMPLATES

We use three prompting strategies across our evaluation. The exact system prompts and formatting templates are reproduced below.

Direct (vanilla) prompting.

> *System*: You are a precise logical reasoning assistant. Answer questions based strictly on the given premises and logical rules. Be concise and clear. Always start your final answer with one of: "Yes", "No", or "Cannot be determined".
>
> *User*: Premise: [premise text]
>
> Question: [question text]

Chain-of-Thought (CoT) prompting.

> *System*: You are a precise logical reasoning assistant. Think step by step through the logical reasoning before giving your final answer. State your final answer starting with "FINAL ANSWER:" followed by "Yes", "No", or "Cannot be determined".
>
> *User*: Premise: [premise text]
>
> Question: [question text]
>
> Think through this step by step:

CGD NLI contradiction check.

> *System*: You are a logical consistency checker. Given two statements derived from the same premise, determine if they contradict each other. Answer "CONTRADICTION" or "NO CONTRADICTION" only.
>
> *User*: Premise: [shared premise]
>
> Statement 1 (prior answer): [prior question] $\rightarrow$ [prior answer]
>
> Statement 2 (current answer): [current question] $\rightarrow$ [current answer]
>
> Do these two answers contradict each other given the shared premise?

CGD revision prompt.

> *User*: You previously answered a related question as follows:
>
> Question: [prior question]
>
> Your answer: [prior answer]
>
> However, your current answer to the following question may be logically inconsistent with your previous answer:
>
> Current question: [current question]
>
> The logical relationship is based on this shared premise: [shared premise]
>
> Please reconsider your answer to the current question, ensuring logical consistency with your previous answer. Start your final answer with "Yes", "No", or "Cannot be determined".

I PER-MODEL FAMILY ANALYSIS

Table 7 summarizes the results by model family, showing that the consistency gap is a cross-family phenomenon.

Table 7: Results aggregated by model family. Values show the range (min–max) across models within each family for Direct prompting.

Family	#Models	IA Range	SCR Range	Gap Range
OpenAI	6	58.4–83.0	4.0–46.7	36.4–57.3
Anthropic	4	51.9–74.8	10.3–23.3	41.2–54.2
Google	3	44.7–67.6	7.3–17.7	37.4–53.1
DeepSeek	2	59.2–70.2	10.0–21.7	48.5–49.2
Meta (Llama)	2	56.8–65.3	10.7–19.0	46.1–46.3
Qwen	1	75.0	32.7	42.3
All	**18**	44.7–83.0	4.0–46.7	36.4–57.3

OpenAI family. The OpenAI models show the widest performance range. GPT-4.1 is the clear best performer overall (SCR 46.7%), while GPT-5 Mini is the weakest in the entire evaluation (SCR 4.0%). Notably, the newer GPT-5 series models (GPT-5.2: 12.0%, GPT-5: 9.3%, GPT-5 Mini: 4.0%) perform substantially *worse* on consistency than the older GPT-4.1 (46.7%), despite likely having higher general capability. The reasoning-specialized o3 (12.0%) does not outperform GPT-4.1 on consistency. Under CGD, GPT-4o shows the largest improvement of any model (+19.7pp), reaching 39.7% SCR.

Anthropic family. Claude Opus 4.6 leads the Anthropic family with 23.3% SCR under Direct prompting and benefits substantially from CGD (+9.0pp to 32.3%). Claude Sonnet 4.6 shows an interesting pattern: relatively low vanilla SCR (10.3%) but one of the largest CGD improvements (+14.0pp to 24.3%), suggesting particularly strong latent consistency knowledge that is activated by the revision mechanism.

Google family. The Gemini models show moderate IA but low SCR. Gemini 2.0 Flash has the lowest IA in the evaluation (44.7%) but benefits from CGD with the largest IA improvement of any model (+11.0pp), suggesting that the contradiction-aware revision helps this model avoid systematic errors.

DeepSeek family. DeepSeek-R1, a reasoning-specialized model, achieves a respectable 21.7% SCR, outperforming most models despite being designed primarily for extended reasoning chains rather than cross-query consistency. DeepSeek V3.2 benefits substantially from CGD (+9.3pp).

Meta and Qwen. The Llama models show moderate performance. Llama 3.1 70B is the only model where CGD hurts (−2.0pp), possibly because the model's limited reasoning capacity leads to error propagation during revision. Qwen 2.5 72B achieves the second-best vanilla SCR (32.7%) and shows minimal but positive improvement with CGD (+0.3pp), suggesting that this model's consistency is already relatively well-calibrated.

J Summary of CGD Improvements

Table 8: Summary statistics for CGD improvements across 17 models.

Metric	ΔSCR	ΔIA
Mean	+6.6pp	+2.8pp
Median	+5.0pp	+2.3pp
Max	+19.7pp (GPT-4o)	+11.0pp (Gem. 2.0 Flash)
Min	−2.0pp (Llama 3.1 70B)	−3.7pp (Llama 3.1 70B)
Models improved (SCR)	16 / 17 (94%)	13 / 17 (76%)
Models improved (IA)	13 / 17	–

The improvement distribution is right-skewed: four models gain more than +9pp SCR (GPT-4o, Claude Sonnet 4.6, GPT-5.2, and tied at +9.0pp: Claude Opus 4.6 and o3), while five models gain less than +3pp. This suggests that CGD is most effective for models that have strong individual reasoning capability but fail to apply it consistently across queries, a condition met by most frontier models.

K Reproducibility Details

All models were accessed via their respective commercial APIs between January and February 2026. We used temperature 0 (or the minimum available temperature) for all evaluations to ensure deterministic outputs. The NLI contradiction checker in CGD uses GPT-4o-mini with temperature 0. Evaluation scripts are implemented in Python and process each question set sequentially. Answer extraction uses exact-match parsing of the first word ("Yes", "No", or "Cannot be determined") from the model's response.

The evaluation cost was approximately $850 across all 37 evaluation runs (18 models $\times$ Direct + 17 $\times$ CGD + 2 $\times$ CoT). CGD runs cost approximately 1.5$\times$ the corresponding Direct run due to the additional NLI checks and occasional revision prompts.