AI Text Detector Report
==========
Overall: AI-Generated (99.95%)

Stats:
  Avg Sentence Length: 22.10
  Burstiness: 10.74
  Function Word Ratio: 0.31
  Lexical Diversity: 0.37
  Readability: 12.20
  Noun Ratio: 0.28
  Verb Ratio: 0.14
  Adj Ratio: 0.13

AI-Generated (Confidence: 99.95%)
Sentence Breakdown: High:126 Medium:9 Low:0
Avg Sentiment: 0.08, Avg Perplexity: 253.82

Sentence Analysis:
1. CS 7650 Midterm Report
Scott Matkosky
Spring 2025
Read the following paper:
Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H.
Chi, Quoc V .
  → AI-Generated (96.96%), Perplexity: 309.85, Sentiment: 0.00
2. Le, and Denny Zhou (2023).
  → Human-Written (77.88%), Perplexity: 150.54, Sentiment: 0.00
3. Chain-of-Thought Prompting Elicits Reasoning
in Large Language Models .
  → AI-Generated (99.68%), Perplexity: 464.19, Sentiment: 0.00
4. Proceedings of the 36th Conference on Neural Information
Processing Systems.
  → Human-Written (78.68%), Perplexity: 119.37, Sentiment: 0.00
5. This document provides the required prompts.
  → AI-Generated (99.96%), Perplexity: 290.88, Sentiment: 0.00
6. Answer each in the space provided.
  → AI-Generated (71.06%), Perplexity: 115.63, Sentiment: 0.00
7. Save the
final version as a PDF and submit to Canvas.
  → AI-Generated (99.84%), Perplexity: 190.48, Sentiment: 0.49
8. 1.
  → AI-Generated (70.98%), Perplexity: 6.61, Sentiment: 0.00
9. Research Challenge
What is the challenge that the authors are trying to address?
  → AI-Generated (98.90%), Perplexity: 29.75, Sentiment: 0.15
10. Large language models have shown notable performance in a variety of language tasks.
  → AI-Generated (96.08%), Perplexity: 86.58, Sentiment: 0.00
11. How-
ever, they tend to falter on problems that require a sequence of reasoning steps rather than
a simple next-word prediction.
  → AI-Generated (99.84%), Perplexity: 158.65, Sentiment: -0.40
12. The challenge, as outlined in the paper, is that these models
often produce a final answer without outlining the intermediate logical steps that lead to that
answer.
  → AI-Generated (99.98%), Perplexity: 110.63, Sentiment: 0.08
13. This absence of a chain of thought makes it difficult to verify the reasoning process and
trust the final output.
  → AI-Generated (99.91%), Perplexity: 133.87, Sentiment: 0.20
14. The authors propose a method called chain-of-thought prompting, where

models are given examples that include intermediate reasoning steps.
→ AI-Generated (98.95%), Perplexity: 261.20, Sentiment: 0.00
15. This approach prompts
the model to decompose complex problems into logical phases, which not only aids in arriving
at the correct answer but also provides transparency.
→ AI-Generated (99.98%), Perplexity: 186.33, Sentiment: -0.21
16. Reasoning is central to human problem-
solving, so for models to be genuinely valuable in serious applications, they should exhibit a
similar breakdown of thought.
→ AI-Generated (99.97%), Perplexity: 157.04, Sentiment: 0.36
17. Without it, models might land on an output by chance or pattern
matching.
→ AI-Generated (99.46%), Perplexity: 483.47, Sentiment: 0.25
18. Chain-of-thought prompting addresses this gap, making reasoning more explicit and
trackable.
→ AI-Generated (99.99%), Perplexity: 468.84, Sentiment: 0.00
19. If fully solved, these models could tackle multi-step mathematical proofs, elaborate
planning, and nuanced decisions across many fields, without requiring extensive fine-tuning.
→ AI-Generated (99.98%), Perplexity: 141.50, Sentiment: 0.34
20. Example of the Challenge:
Suppose a model is asked, “If each table seats four and you have 18 guests, how many tables do
you need?” The model might say “4” but fail to explain whether it considered leftover guests
or partial tables.
→ AI-Generated (98.00%), Perplexity: 67.98, Sentiment: -0.68
21. With chain-of-thought prompting, you’d see it multiply 4 by the number of
1 tables, compare to the total guests, and conclude the correct count—thereby revealing each
step.
→ AI-Generated (99.99%), Perplexity: 200.91, Sentiment: 0.08
22. Why is this challenge important?
→ AI-Generated (99.85%), Perplexity: 65.66, Sentiment: 0.29
23. The difference between how humans think step-by-step and how models often jump to con-
clusions can create serious issues in practice.
→ AI-Generated (99.95%), Perplexity: 225.86, Sentiment: 0.20
24. For instance, if a model proposes a household
budget but overlooks tax deductions or fees, it might produce a neat total that’s far off reality.
→ AI-Generated (99.99%), Perplexity: 228.39, Sentiment: 0.61
25. In healthcare, a model could appear correct while skipping a crucial symptom or dosage note,
leading to a bad prescription.
→ AI-Generated (99.99%), Perplexity: 565.66, Sentiment: -0.54
26. These examples show how a decisive-sounding answer can hide
major flaws if intermediate steps stay hidden.
→ AI-Generated (99.98%), Perplexity: 593.56, Sentiment: -0.18
27. Asking the model to disclose each reasoning
phase helps users detect mistakes and trust final answers more.
→ AI-Generated (99.91%), Perplexity: 962.33, Sentiment: 0.53
28. It also generalizes more nat-
urally—models guided by chain-of-thought prompting can handle tasks demanding multi-step
logic, without heavy or domain-specific re-training.
→ AI-Generated (99.88%), Perplexity: 514.87, Sentiment: -0.23
29. What would one be able to do if the challenge were solved completely?
→ AI-Generated (85.44%), Perplexity: 44.49, Sentiment: 0.34
30. If this challenge were fully solved, models would systematically present coherent reasoning
steps behind their answers.
→ AI-Generated (99.96%), Perplexity: 984.83, Sentiment: 0.40
31. They could serve as transparent teaching tools in education, clari-
fying detailed lines of thought, or act as dependable assistants in domains like finance or law,
where verifying each assumption is vital.
→ AI-Generated (99.83%), Perplexity: 319.44, Sentiment: 0.57
32. They’d also adapt easily to new challenges that in-
volve deeper analysis, like multi-stage scientific research or planning complex projects.

→ AI-Generated (99.96%), Perplexity: 189.36, Sentiment: 0.64
33. And
because the reasoning would be laid out, anyone could audit or refine the model's approach,
shifting the paradigm from "black box" to "collaborative partner.
→ AI-Generated (99.93%), Perplexity: 61.25, Sentiment: 0.00
34. "
2.
→ Human-Written (61.30%), Perplexity: 31.28, Sentiment: 0.00
35. Contributions and Key Idea
What do the authors claim to be the central novel technical, theoretical, or conceptual
contribution of the paper?
→ Human-Written (79.84%), Perplexity: 110.57, Sentiment: 0.32
36. The central contribution is a method called chain-of-thought prompting, which encourages
large language models to write out a sequence of reasoning steps before giving a final answer.
→ AI-Generated (99.71%), Perplexity: 160.68, Sentiment: 0.65
37. These intermediate steps don't rely on labor-intensive annotations; they emerge when the model
is shown a short set of sample problems that already illustrate step-by-step thinking.
→ AI-Generated (96.75%), Perplexity: 120.57, Sentiment: -0.40
38. This
approach harnesses a model's ability to learn from context: by offering examples where each
phase of problem-solving is spelled out, the model starts to mimic that style in new tasks.
→ AI-Generated (99.99%), Perplexity: 111.23, Sentiment: 0.32
39. In
essence, the authors argue that multi-step reasoning becomes more visible once prompts are
2 crafted to emphasize logical sequences.
→ AI-Generated (99.98%), Perplexity: 433.21, Sentiment: -0.34
40. Unlike rigid, domain-specific methods (i.e., systems
that use carefully tailored rules or specialized data for a single niche and don't generalize
easily), chain-of-thought prompting uses plain language for each intermediate stage, which
makes it more adaptable.
→ AI-Generated (64.42%), Perplexity: 129.97, Sentiment: 0.00
41. The authors also demonstrate that adjusting the prompt format can
unlock capabilities hidden in large models—hence the repeated emphasis on "em dashes" to
show how minimal the extra effort can be.
→ AI-Generated (99.94%), Perplexity: 205.46, Sentiment: 0.00
42. Example around main contribution:
Imagine a puzzle about assigning workloads among five employees, each with different dead-
lines and skill sets.
→ AI-Generated (99.99%), Perplexity: 305.97, Sentiment: -0.65
43. If you just ask for the final allocation, the model might provide an answer
but leave out its rationale.
→ AI-Generated (99.93%), Perplexity: 166.22, Sentiment: -0.08
44. With chain-of-thought prompting, the model sees examples that de-
tail each deciding factor—like who has design expertise or who's already at capacity—then
follows a similar pattern when it tackles the new puzzle.
→ AI-Generated (99.98%), Perplexity: 270.42, Sentiment: 0.00
45. It describes the constraints and how it
integrates them, letting you check each step for consistency.
→ AI-Generated (99.99%), Perplexity: 205.89, Sentiment: 0.00
46. What is the key idea that allowed progress on the research challenge?
→ AI-Generated (99.92%), Perplexity: 88.45, Sentiment: 0.48
47. The essential insight is that very large language models have a latent capacity for multi-step
reasoning, but standard prompts don't push them to reveal it.
→ AI-Generated (99.99%), Perplexity: 136.49, Sentiment: 0.00
48. In most "standard prompts, "
which are short requests or single questions, the model jumps straight to a final answer without
showing how it arrived there.
→ AI-Generated (99.74%), Perplexity: 260.86, Sentiment: 0.23
49. Once the authors introduced chain-of-thought prompts—examples
that explicitly lay out each step of the reasoning—the model began to imitate that structured

approach, exposing a detailed thought process it would otherwise skip.
→ AI-Generated (99.99%), Perplexity: 180.33, Sentiment: -0.27
50. By “latent capacity, ”
the authors mean these large models, often exceeding 100 billion parameters, already represent reasoning internally but usually don’t surface it.
→ AI-Generated (98.68%), Perplexity: 429.63, Sentiment: 0.00
51. Notably, smaller models (well below
100B) don’t exhibit the same effect, suggesting there’s a threshold of scale for chain-of-thought prompting to succeed.
→ AI-Generated (99.98%), Perplexity: 89.87, Sentiment: 0.49
52. The core realization, then, is that big models can handle multi-step
logic, yet they need prompts constructed to highlight every incremental step.
→ AI-Generated (99.97%), Perplexity: 229.32, Sentiment: 0.34
53. Once given such
examples, the model naturally takes on a similar format for new questions, enabling a level of transparency and logical structure absent from typical short-answer prompts.
→ AI-Generated (99.98%), Perplexity: 193.71, Sentiment: 0.00
54. 3.
→ AI-Generated (73.42%), Perplexity: 6.72, Sentiment: 0.00
55. Prior Work and Alternatives
What was the state of the art before the paper?
→ Human-Written (99.93%), Perplexity: 76.00, Sentiment: 0.00
56. Before this paper, people often relied on “few-shot prompts, ” where only a handful of examples
3 (sometimes just one or two) demonstrate how to do a task.
→ AI-Generated (99.85%), Perplexity: 96.31, Sentiment: 0.00
57. While that works for simpler queries,
it typically falls short when multiple reasoning steps are needed.
→ AI-Generated (99.96%), Perplexity: 262.77, Sentiment: 0.00
58. Researchers tackled the gap
by either fine-tuning on large sets of explanations or examples (which demands a lot of human labor) or by building specialized symbolic systems that break complex problems into smaller chunks—each chunk handled by a separate rule or sub-module.
→ AI-Generated (99.94%), Perplexity: 160.81, Sentiment: -0.40
59. These methods could succeed
in constrained scenarios but required significant engineering effort or domain-specific data.
→ AI-Generated (99.89%), Perplexity: 319.78, Sentiment: 0.48
60. Chain-of-thought prompting, by contrast, seeks a more universal solution: the model observes a small set of step-by-step examples within the prompt, then applies that structured approach to new, complex problems without needing big architectural changes or massive new datasets.
→ AI-Generated (99.98%), Perplexity: 129.71, Sentiment: -0.03
61. What-If Example:
Suppose you want the AI to handle a multi-step puzzle about scheduling recurring events for several team members—each with unique time windows and constraints.
→ AI-Generated (99.96%), Perplexity: 97.83, Sentiment: 0.08
62. One older method
might split the puzzle: (1) list each member’s available slots, (2) compare them in pairs, (3) merge overlaps, then (4) finalize the schedule.
→ AI-Generated (99.44%), Perplexity: 63.03, Sentiment: 0.00
63. Another approach might involve fine-tuning the
model on a large dataset of human-made scheduling rationales.
→ AI-Generated (99.95%), Perplexity: 146.76, Sentiment: 0.00
64. Both are viable but demand
either specialized code or intensive data.
→ AI-Generated (99.93%), Perplexity: 1025.67, Sentiment: -0.19
65. In contrast, chain-of-thought prompting would give
the model a couple of examples that outline, in text, how to consider each constraint sequentially.
→ Human-Written (59.93%), Perplexity: 314.88, Sentiment: 0.00
66. The model then follows this style when tackling fresh scheduling problems, sidestepping heavy domain engineering and data collection.

→ AI-Generated (99.98%), Perplexity: 833.72, Sentiment: -0.10
67. Example of Prior Approaches:
Some teams developed “scratchpad” methods that let the model maintain a hidden text record of partial steps—essentially a running commentary the model uses internally before arriving at a final answer.
→ AI-Generated (98.29%), Perplexity: 236.27, Sentiment: 0.00
68. This can make the reasoning more visible or structured, but it often requires special data or configurations to train the model to use that hidden space effectively.
→ AI-Generated (99.96%), Perplexity: 191.92, Sentiment: 0.81
69. Other researchers used symbolic logic pipelines, where predefined rules or algorithms handle arithmetic or puzzle-solving tasks by systematically breaking them down.
→ AI-Generated (99.87%), Perplexity: 338.24, Sentiment: 0.00
70. Although both tactics have produced acceptable outcomes in specific settings, they demand dedicated data or coding for each domain.
→ AI-Generated (99.93%), Perplexity: 644.15, Sentiment: 0.59
71. A scratchpad set up for math word problems might not automatically extend to legal analysis or financial planning, and a symbolic solver for combinatorial puzzles might need an entirely different rule set for scheduling or medical questions.
→ Human-Written (53.41%), Perplexity: 332.16, Sentiment: -0.41
72. That lack of adaptability keeps these methods from seamlessly transferring to broad new tasks.
→ AI-Generated (99.98%), Perplexity: 335.52, Sentiment: -0.32
73. 4 What are the alternative techniques that one might have used prior to this paper, and why might they not have been sufficient?
→ Human-Written (86.44%), Perplexity: 75.77, Sentiment: 0.00
74. Fine-tuning for each domain is expensive, needing large labeled sets for tasks like math or legal analysis.
→ AI-Generated (99.98%), Perplexity: 618.89, Sentiment: 0.46
75. Symbolic frameworks, while accurate for well-defined problems, don’t adapt well to open-ended text or situations with fuzzy details.
→ AI-Generated (99.97%), Perplexity: 163.88, Sentiment: -0.15
76. They also lose natural language’s flexibility.
→ AI-Generated (99.16%), Perplexity: 186.13, Sentiment: 0.30
77. Chain-of-thought prompting solves many of these issues by embedding stepwise logic in normal text, avoiding huge annotation costs or specialized rule-based coding.
→ AI-Generated (99.96%), Perplexity: 383.26, Sentiment: 0.25
78. 4.
→ AI-Generated (72.89%), Perplexity: 7.01, Sentiment: 0.00
79. Evaluation and Evidence
4.
→ Human-Written (55.64%), Perplexity: 146.95, Sentiment: 0.00
80. What evidence is given that the work in the paper has improved the state of the art on the challenge they identified?
→ Human-Written (99.73%), Perplexity: 113.63, Sentiment: 0.53
81. The authors tested chain-of-thought prompts on several math and commonsense benchmarks.
→ AI-Generated (98.13%), Perplexity: 323.33, Sentiment: 0.00
82. For arithmetic, they used GSM8K (grade-school math word problems), SVAMP (varied math tasks with tricky structures), and MAWPS (a collection of diverse arithmetic challenges).
→ AI-Generated (99.07%), Perplexity: 384.51, Sentiment: -0.15
83. In each case, chain-of-thought prompts outperformed standard prompts—often by large margins, sometimes even doubling accuracy on intricate multi-step problems.
→ AI-Generated (99.99%), Perplexity: 294.60, Sentiment: -0.27
84. The paper includes both numerical results and concrete examples showing how chain-of-thought prompting boosts correctness by coaxing the model to articulate intermediate steps.
→ AI-Generated (98.64%), Perplexity: 477.79, Sentiment: 0.32
85. Moreover, the authors compare

these outcomes to older GPT-3 versions that were fine-tuned on reasoning tasks, noting that chain-of-thought prompting can match or exceed those baselines with much less configuration effort.
→ AI-Generated (99.98%), Perplexity: 175.38, Sentiment: 0.00
86. Beyond arithmetic, the paper highlights improvements on commonsense tasks, such as CSQA (a broad test of everyday knowledge questions), StrategyQA (which demands multi-hop strategic reasoning), and specialized sets like Date Understanding and Sports Understanding.
→ AI-Generated (99.96%), Perplexity: 376.59, Sentiment: 0.59
87. In those scenarios, chain-of-thought prompts also lift performance by revealing how the model pieces together facts or logical connections.
→ AI-Generated (99.99%), Perplexity: 627.54, Sentiment: 0.00
88. Example of the Evidence:
In one GSM8K problem about distributing pie slices among multiple guests, chain-of-thought prompting led the model to spell out each fraction or leftover step.
→ AI-Generated (74.75%), Perplexity: 354.16, Sentiment: -0.40
89. Observers could see, at a
glance, whether it handled partial slices or miscounted.
→ AI-Generated (99.91%), Perplexity: 221.20, Sentiment: 0.00
90. Under a plain prompt, the model just
gave a final fraction, leaving no trace of its logic or any mistakes.
→ AI-Generated (99.98%), Perplexity: 177.20, Sentiment: -0.57
91. Similarly, on a StrategyQA
item, the model listed each relevant fact or sub-question before concluding the correct answer—a stark contrast to a simple one-line output.
→ AI-Generated (99.98%), Perplexity: 353.13, Sentiment: 0.00
92. Such transparency not only raises accuracy
5 but also reveals exactly where a slip might occur if the model gets it wrong.
→ AI-Generated (99.95%), Perplexity: 210.06, Sentiment: -0.63
93. Is the evidence convincing and conclusive?
→ AI-Generated (99.54%), Perplexity: 210.01, Sentiment: 0.40
94. Why or why not?
→ AI-Generated (66.64%), Perplexity: 30.99, Sentiment: 0.00
95. The evidence seems strong for tasks tested so far, with clear gains on arithmetic and commonsense benchmarks.
→ AI-Generated (99.98%), Perplexity: 524.22, Sentiment: 0.82
96. Nevertheless, most of these improvements appear mainly when using very
large models—those with hundreds of billions of parameters.
→ AI-Generated (97.10%), Perplexity: 193.18, Sentiment: 0.37
97. Smaller models see fewer benefits, hinting there's a scale threshold for chain-of-thought to show real value.
→ AI-Generated (99.98%), Perplexity: 272.93, Sentiment: 0.34
98. Furthermore,
while arithmetic and commonsense tasks respond well, it remains unclear whether specialized or ambiguous domains—like complex legal analysis—would see the same benefits.
→ AI-Generated (99.98%), Perplexity: 204.56, Sentiment: 0.49
99. Thus, the
findings are compelling within the scope presented, but more wide-ranging studies might be required to confirm generality across all task types.
→ AI-Generated (99.98%), Perplexity: 131.15, Sentiment: 0.12
100. Give at least one reason in support and one reason against.
→ Human-Written (99.60%), Perplexity: 89.41, Sentiment: 0.40
101. Reason in support: One concrete example is how chain-of-thought prompting doubled accuracy on GSM8K for a large model.
→ AI-Generated (97.91%), Perplexity: 509.16, Sentiment: 0.40
102. That's a dramatic jump, suggesting it really does tease out multi-
step logic that normally stays hidden.
→ AI-Generated (99.99%), Perplexity: 286.50, Sentiment: -0.38
103. It also provides readable intermediate steps, which makes
the final answer more trustworthy and easier to debug.
→ AI-Generated (99.74%), Perplexity: 162.31, Sentiment: 0.79

104. Reason against: The method’s reliance on large-scale models means smaller models—those under, say, 100B parameters—gain almost nothing.
→ AI-Generated (99.97%), Perplexity: 220.41, Sentiment: 0.00
105. This restricts who can benefit, since not
every research group or production environment has the resources to host a massive model.
→ AI-Generated (94.69%), Perplexity: 394.70, Sentiment: 0.18
106. Additionally, some domains might require specialized knowledge beyond what chain-of-thought can provide, limiting its overall coverage.
→ AI-Generated (99.99%), Perplexity: 217.06, Sentiment: 0.00
107. 5.
→ AI-Generated (77.75%), Perplexity: 7.09, Sentiment: 0.00
108. Missing Information
5.
→ AI-Generated (78.56%), Perplexity: 184.32, Sentiment: -0.30
109. What is a piece of information that you wish was given but is not in the paper’s evaluation?
→ Human-Written (99.54%), Perplexity: 73.78, Sentiment: 0.21
110. Although the authors demonstrate clear benefits from chain-of-thought prompting, a deeper comparison against “symbolic solvers” or “hybrid neural-symbolic methods” would help contextualize these gains.
→ AI-Generated (98.67%), Perplexity: 95.70, Sentiment: 0.85
111. Symbolic solvers are systems that rely on hand-crafted rules or formal
logic representations, often excelling at tasks like algebra or constraint satisfaction by following precise inference procedures.
→ AI-Generated (99.99%), Perplexity: 382.53, Sentiment: 0.81
112. Hybrid approaches blend these symbolic routines with
neural models, using the former to handle logical structure and the latter for open-ended language tasks.
→ AI-Generated (99.98%), Perplexity: 383.41, Sentiment: 0.00
113. Directly pitting chain-of-thought prompting against these methods—on the same
6 dataset—would show whether a plain-language approach can rival or exceed rule-based logic in complex scenarios.
→ AI-Generated (61.34%), Perplexity: 335.63, Sentiment: 0.00
114. Furthermore, some detail on the runtime or computational overhead of
generating a “stepwise solution” would clarify when chain-of-thought is feasible.
→ AI-Generated (97.99%), Perplexity: 118.36, Sentiment: 0.00
115. A “stepwise
solution” refers to the model enumerating each stage of its reasoning, rather than jumping to an end result.
→ AI-Generated (83.67%), Perplexity: 125.93, Sentiment: 0.00
116. If a problem has many steps (for instance, multi-stage planning or an extended
math proof), the model will generate more tokens, raising inference time and costs.
→ AI-Generated (79.42%), Perplexity: 234.80, Sentiment: -0.40
117. Under-
standing how well chain-of-thought prompting scales with longer reasoning chains—both in terms of accuracy and resource consumption—would help practitioners decide when and how to employ it.
→ AI-Generated (99.54%), Perplexity: 268.21, Sentiment: 0.59
118. Example of Missing Info:
In addition to basic arithmetic problems, one might want to test chain-of-thought prompts on a more elaborate math set—say, an advanced geometry or calculus dataset—and compare performance against a specialized rule-based solver designed for those domains.
→ AI-Generated (99.13%), Perplexity: 103.68, Sentiment: -0.38
119. If the chain-
of-thought approach can rival or outdo a system that uses meticulously coded theorems or symbolic operations, it would confirm that plain-language reasoning can scale to tough, structured tasks without requiring hand-engineered logic.
→ AI-Generated (99.91%), Perplexity: 139.12, Sentiment: -0.13
120. On the other hand, if the symbolic solver
dominates, it would highlight where chain-of-thought prompting still falls short, helping define

the boundaries of its applicability.
→ AI-Generated (99.99%), Perplexity: 165.41, Sentiment: 0.79
121. 7 6.
→ AI-Generated (72.99%), Perplexity: 38.85, Sentiment: 0.00
122. Limitations
6.
→ AI-Generated (79.67%), Perplexity: 296.63, Sentiment: 0.00
123. What are the potential limitations of the approach?
→ AI-Generated (99.87%), Perplexity: 28.97, Sentiment: 0.00
124. That is, where might the proposed
technique not work, or what problems might the technique not apply to?
→ Human-Written (80.64%), Perplexity: 155.53, Sentiment: -0.40
125. Chain-of-thought relies heavily on scaling—massive models are required to see big improvements.
→ AI-Generated (99.83%), Perplexity: 551.03, Sentiment: 0.44
126. Smaller models gain little.
→ AI-Generated (99.43%), Perplexity: 539.22, Sentiment: 0.53
127. Moreover, stepwise logic doesn’t guarantee correctness; a
model might lay out a thorough argument and still reach the wrong conclusion if its steps are flawed.
→ AI-Generated (99.98%), Perplexity: 141.86, Sentiment: -0.54
128. Generating full reasoning also increases token usage, slowing responses and raising costs.
→ AI-Generated (99.85%), Perplexity: 1185.81, Sentiment: 0.00
129. In resource-limited situations or real-time settings, that can be prohibitive.
→ Human-Written (64.77%), Perplexity: 105.50, Sentiment: 0.00
130. Example of a Limitation:
Suppose an LLM is providing financial advice about retirement plans.
→ AI-Generated (98.28%), Perplexity: 146.17, Sentiment: -0.30
131. It might lay out a detailed
chain explaining how to invest in different accounts, referencing tax brackets or contribution limits.
→ AI-Generated (99.99%), Perplexity: 589.15, Sentiment: 0.00
132. At first glance, the argument seems thorough.
→ AI-Generated (99.98%), Perplexity: 102.94, Sentiment: -0.36
133. However, if the model neglects a recent
change in tax law or confuses state-specific rules, the final recommendation could be misguided, even though the chain-of-thought looks convincing.
→ AI-Generated (99.99%), Perplexity: 122.95, Sentiment: -0.42
134. Thus, chain-of-thought prompting
isn’t a guarantee of correctness—expert oversight or external checks remain essential.
→ AI-Generated (99.98%), Perplexity: 161.60, Sentiment: 0.25
135. 8
→ Human-Written (66.27%), Perplexity: 1.00, Sentiment: 0.00