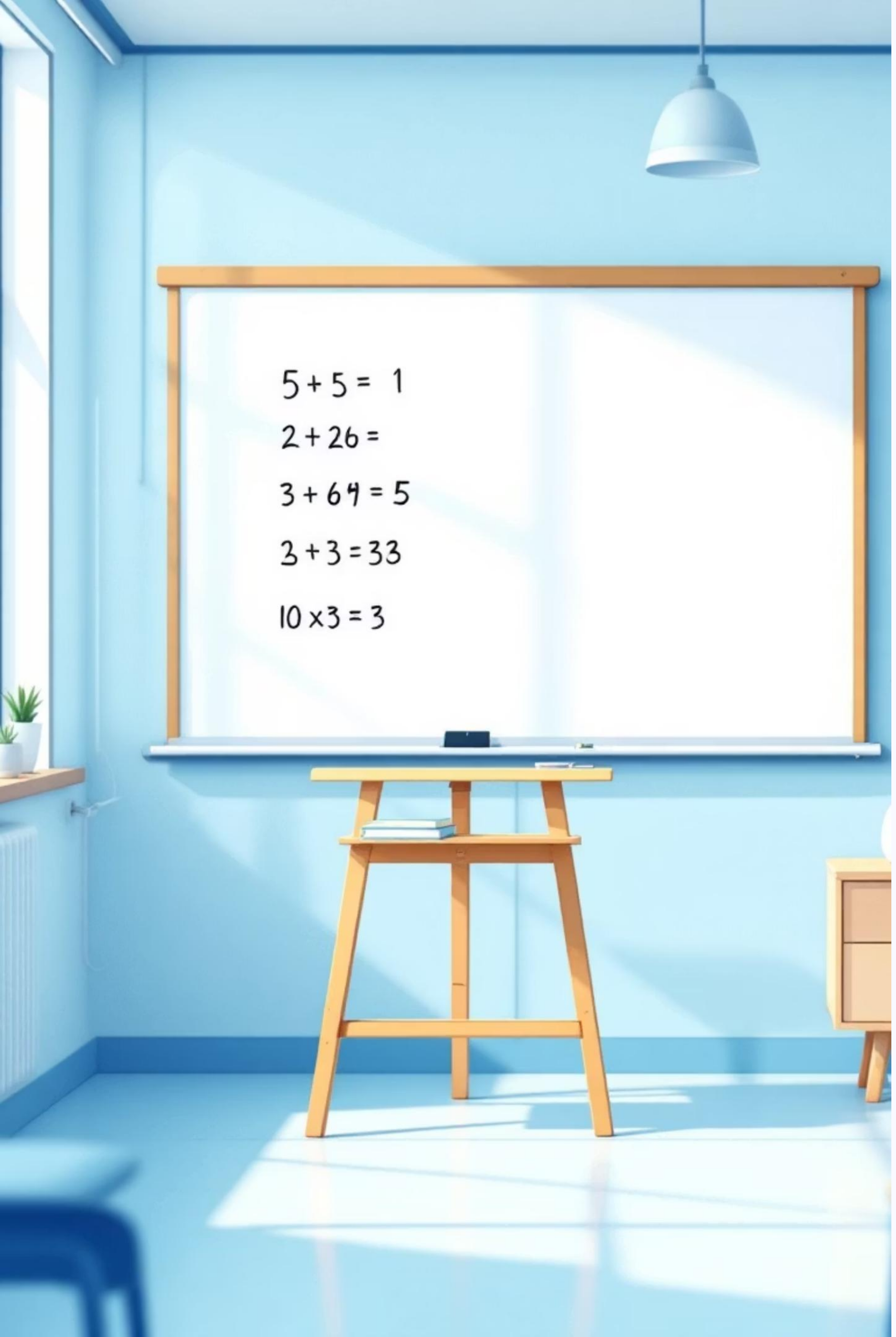
Chain of Thought Prompting Elicits Reasoning in Large Language Models

~Jason Wei et al., 2022

An influential paper investigating how step-by-step reasoning exemplars enables large language models to perform complex reasoning tasks across multiple domains.

Presented By:
Krish Mehta



What is Chain of Thought Prompting?

01

Provide Examples

Give the model step-by-step reasoning exemplars that demonstrate the thought process.

02

Enable Reasoning

The model learns to break down complex problems into manageable steps.

ĆĊ

Generate Solutions

Model produces detailed reasoning chains leading to accurate answers.

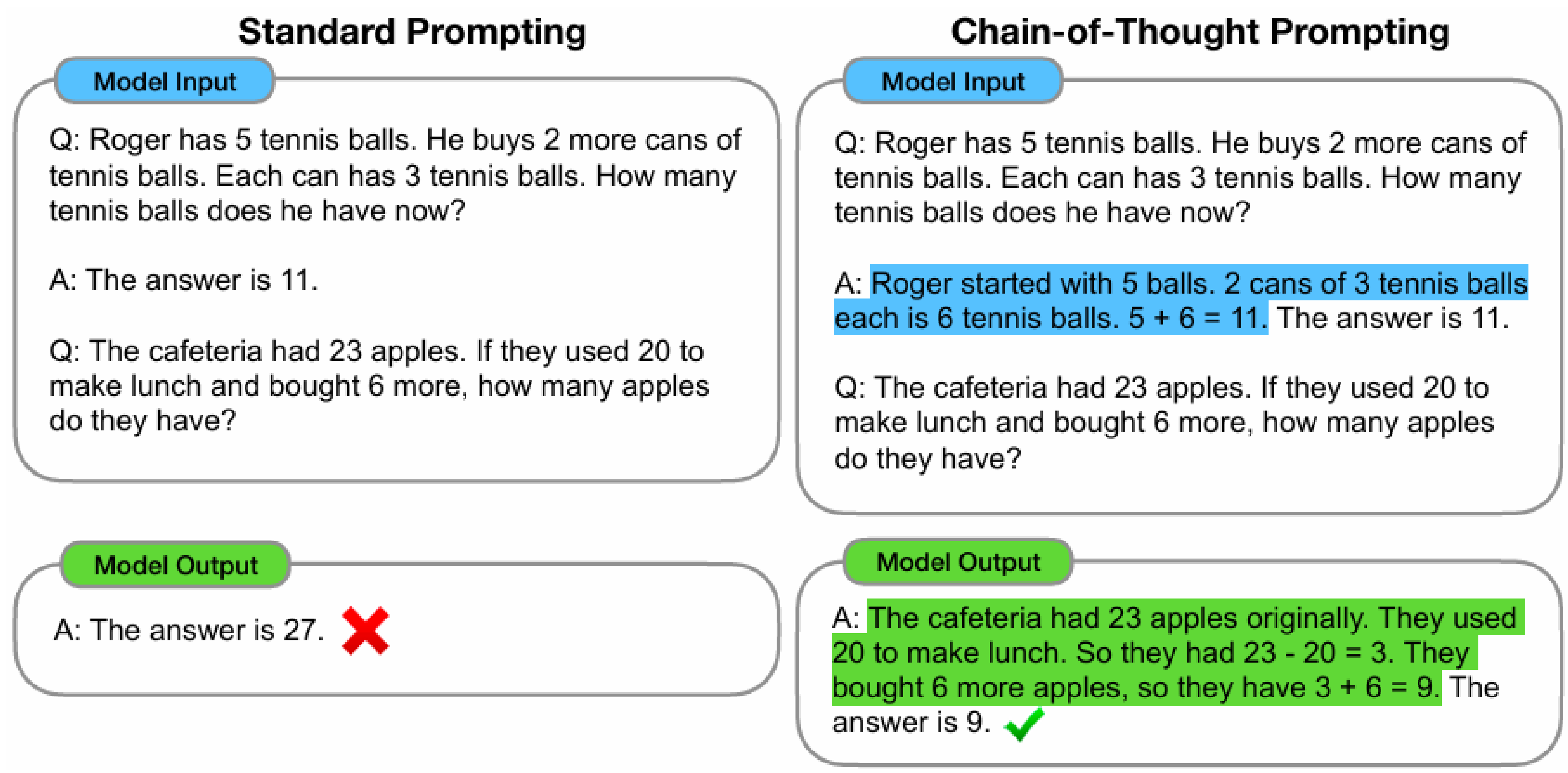
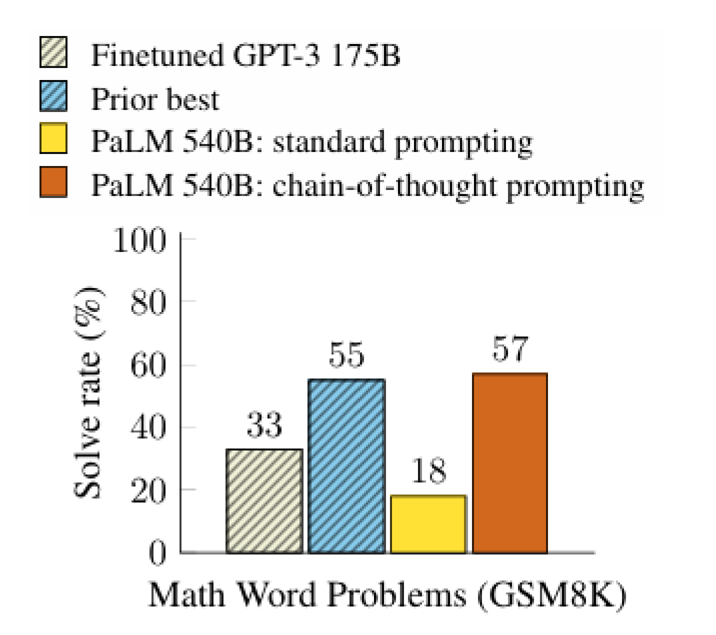


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

Remarkable Performance Gains



The 540B-parameter model surpassed even fine-tuned models with additional verification on math word problems.

540B

Model Parameters

Large-scale model achieved state-of-the-art results with just eight CoT exemplars.

100B+

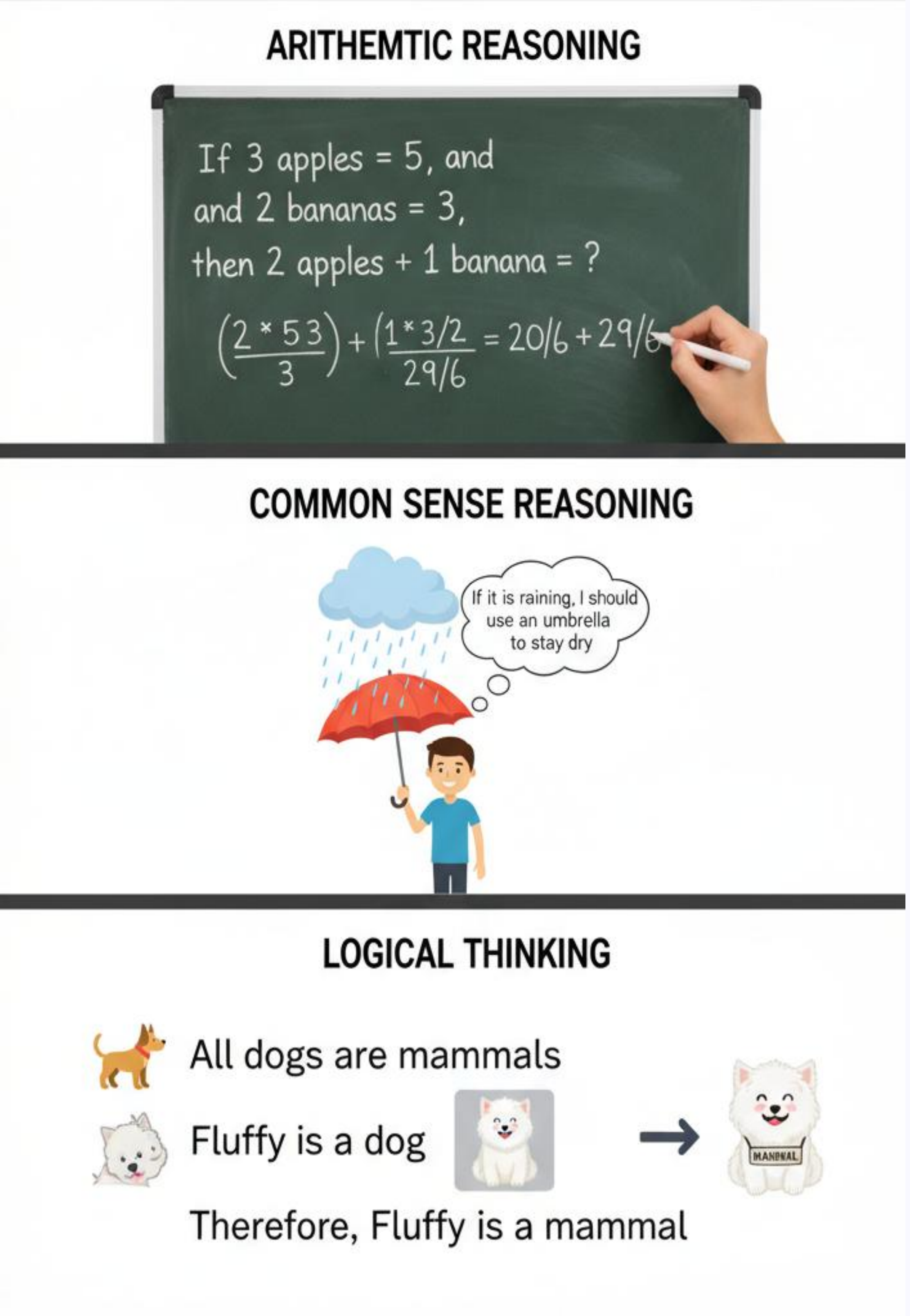
Scale Threshold

Reasoning abilities emerge only in models exceeding 100 billion parameters.

8

Few Examples Needed

Just eight exemplars enabled breakthrough performance on GSM8K benchmark.



Three Key Reasoning Domains

Arithmetic Reasoning

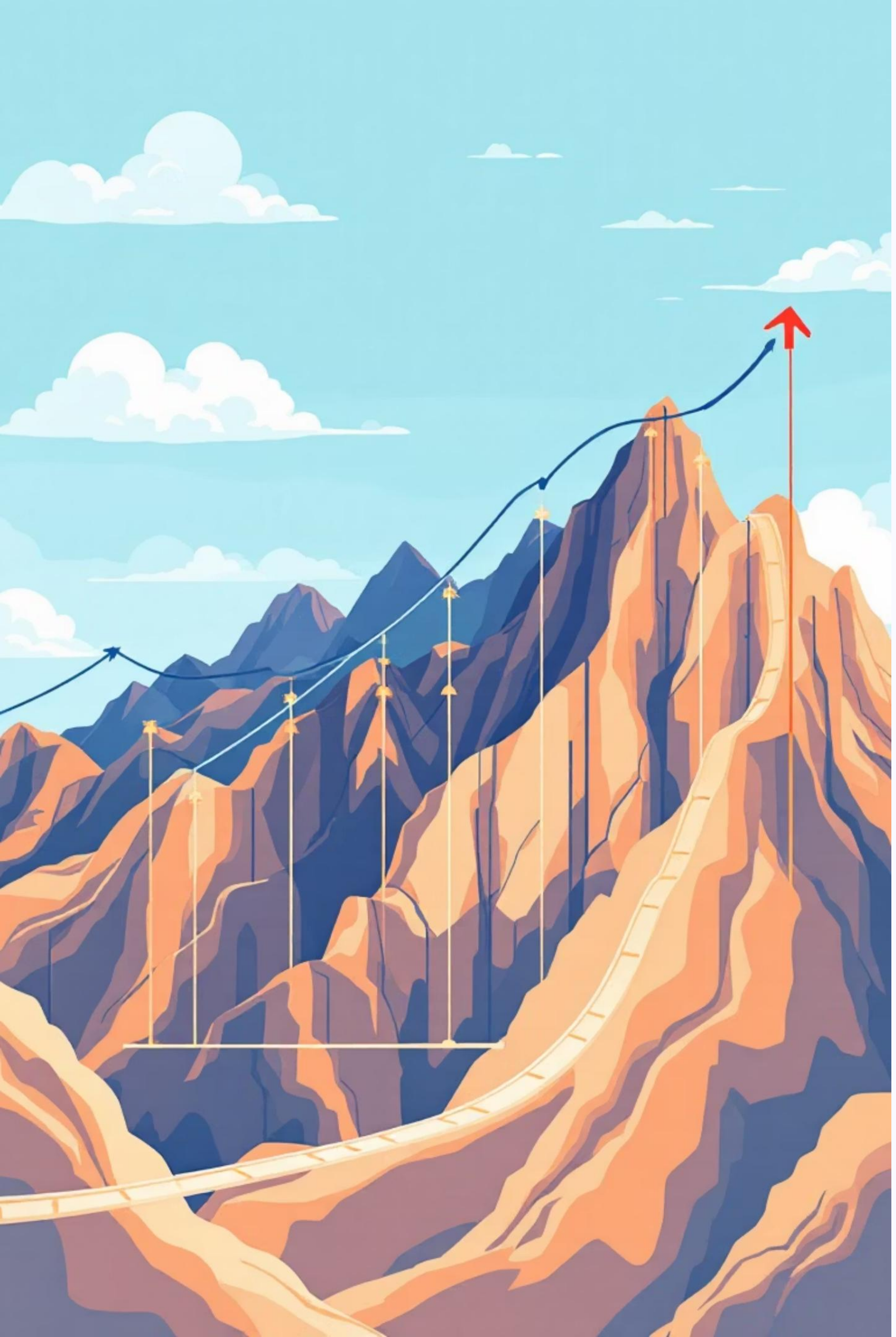
Mathematical problem-solving and numerical computations with step-by-step calculations.

Commonsense Reasoning

Everyday logical thinking and practical knowledge application in real-world scenarios.

Symbolic Reasoning

Abstract logical operations and pattern recognition across symbolic representations.



Emergent Property of Scale

1

Small Models

Limited reasoning capabilities, CoT may actually hurt performance below 10B parameters.

2

Large Models

Complex reasoning emerges unpredictably at scale, not predictable from smaller versions.

This emergence represents a fundamental shift in AI capabilities that occurs only at sufficient scale.

Key Strengths of the Research

Strong Testing

The paper tests CoT prompting on many types of reasoning (math, commonsense, symbolic) and across different large models, proving it really works..

Simplicity & Generality

Method is easy to implement and broadly applicable, requiring only few exemplars.

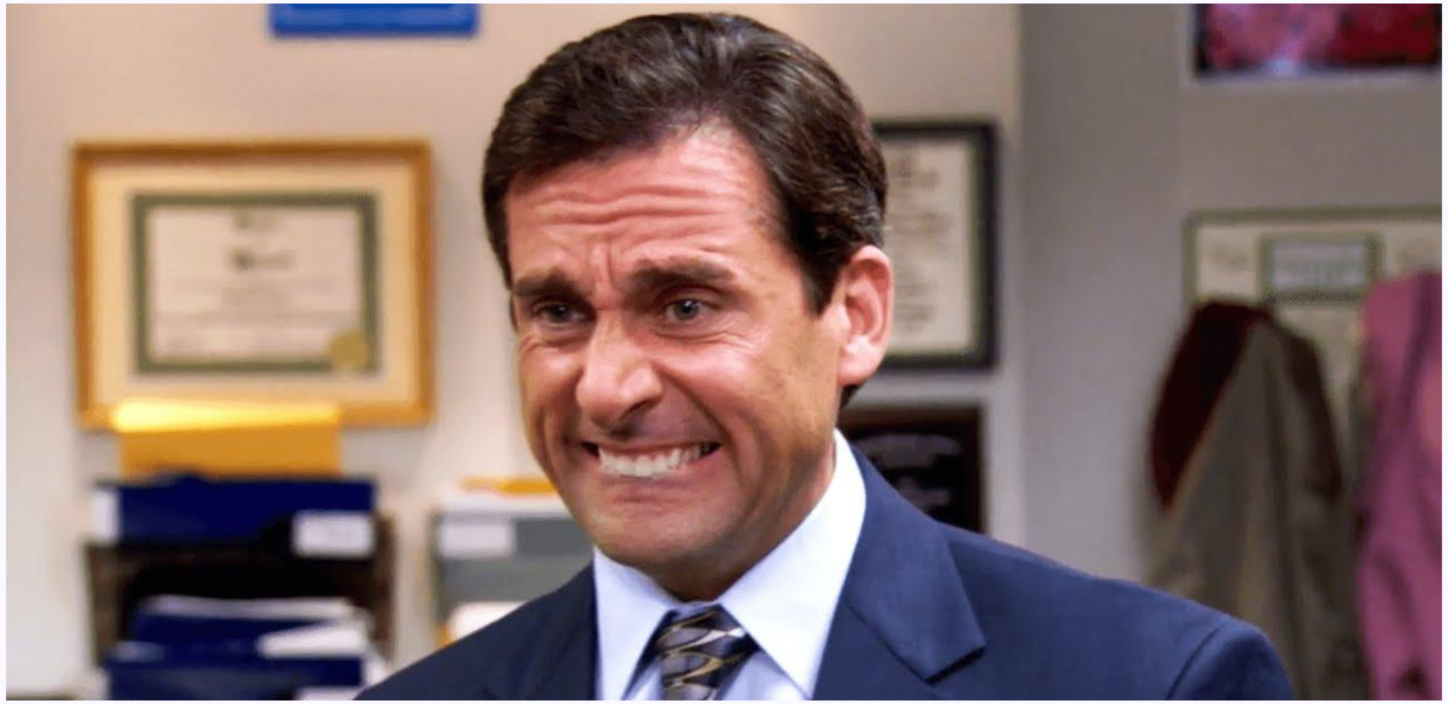
State-of-the-Art Results

Achieved breakthrough performance on challenging benchmarks, demonstrating practical impact.

Error Analysis

Provides insights into limitations and failure modes through systematic mistake categorisation.

Critical Limitations



Scale Dependency

CoT prompting can hurt performance in models smaller than 10B parameters, limiting universal applicability.

Toy Task Inflation

Some evaluations use simplified tasks with provided solution structures, potentially inflating results.

Unexplained Emergence

Paper documents but doesn't explain why scale enables reasoning, leaving mechanistic questions unanswered.

Reliability Concerns

Semantic Misunderstandings

Models sometimes misinterpret problem context or requirements despite step-by-step reasoning.

Missing Steps

Reasoning chains may skip crucial logical steps, leading to incomplete or incorrect solutions.

Hallucinations

Models can generate plausible-sounding but factually incorrect reasoning steps and conclusions.

These issues suggest CoT prompting doesn't guarantee reliable or faithful reasoning processes.

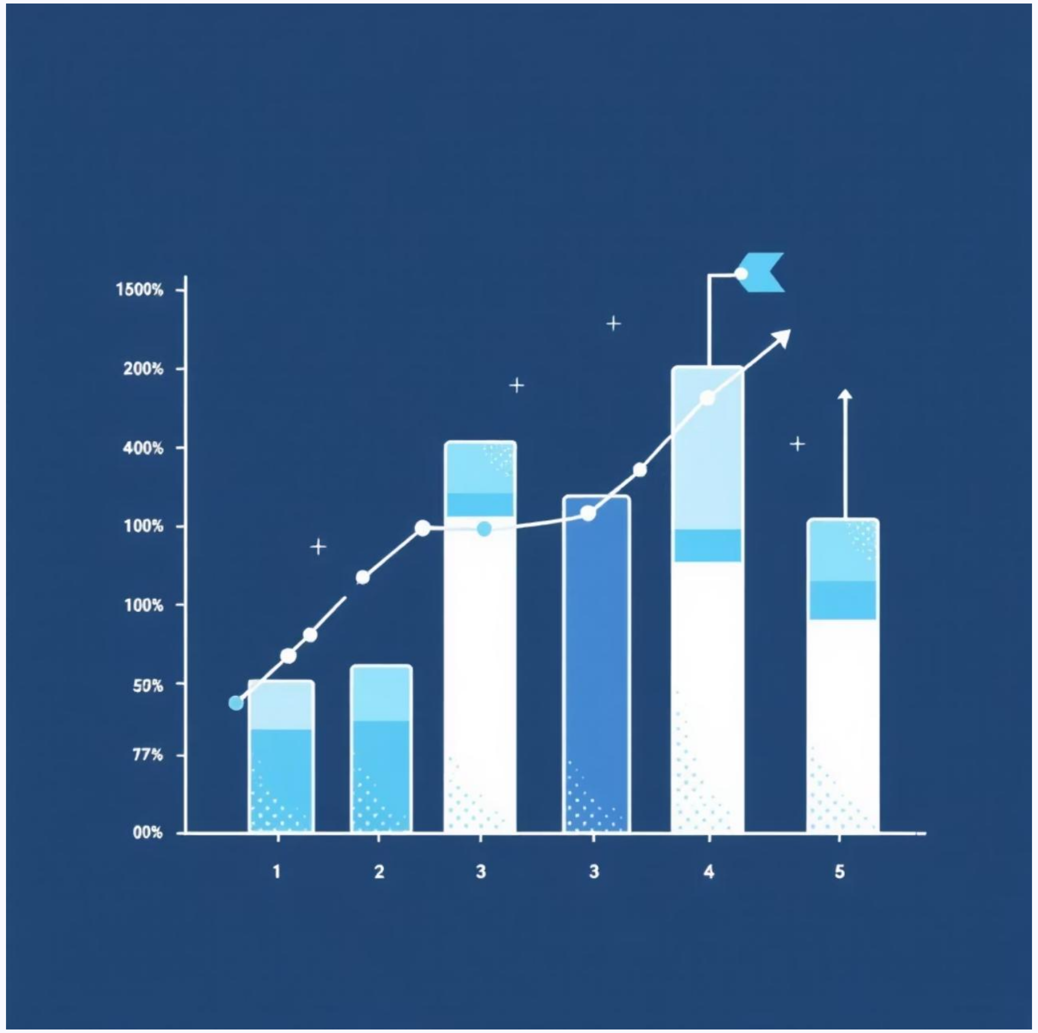
Generalisation Challenges

In-Domain vs Out-of-Domain Performance

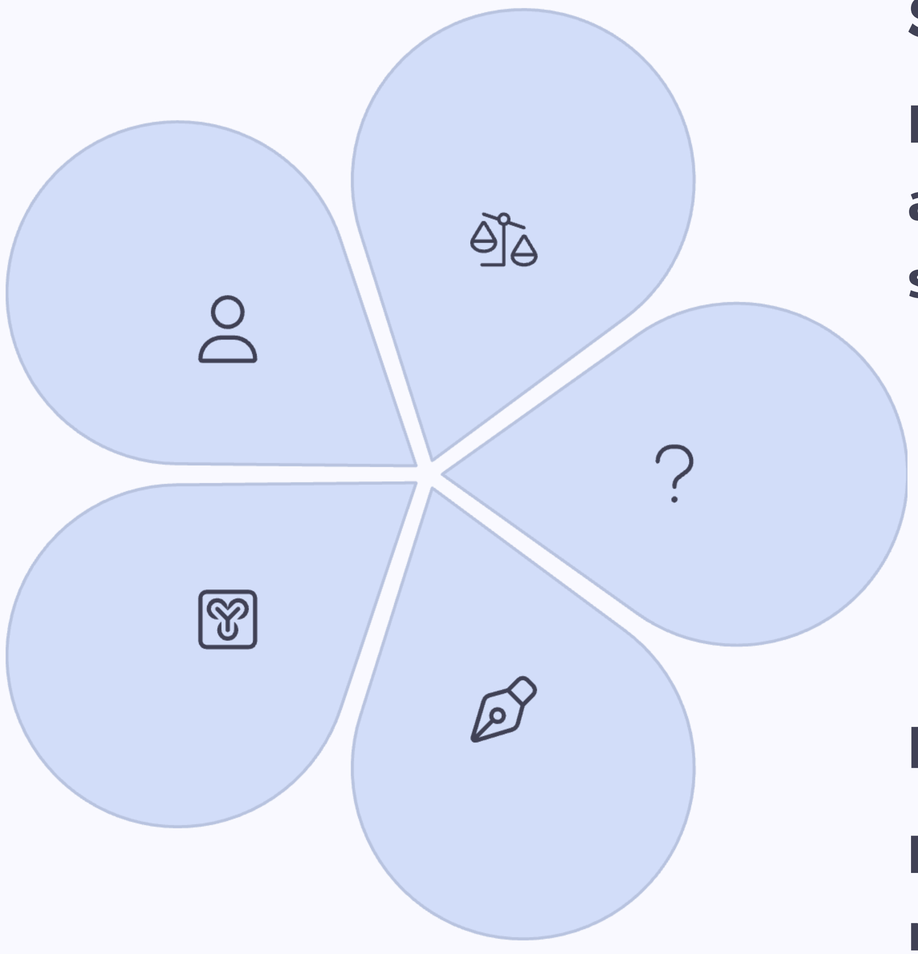
While CoT prompting improves out-of-domain performance, it remains significantly lower than in-domain results.

This indicates potential limitations in generalising reasoning abilities across different types of tasks and contexts.

The method may be more task-specific than initially hoped, requiring careful consideration of domain boundaries.



Impact and Future Directions



Breakthrough Method

Simple yet powerful technique for eliciting reasoning in large models.

Scale Insights

Reveals emergent properties that appear only at sufficient model scale.

Open Questions

Mechanisms behind emergence remain unexplained, requiring further research.

Future Research

Understanding emergence and improving generalisation remain key priorities.

Reliability Challenges

Hallucinations and errors highlight need for verification mechanisms.

THANK YOU