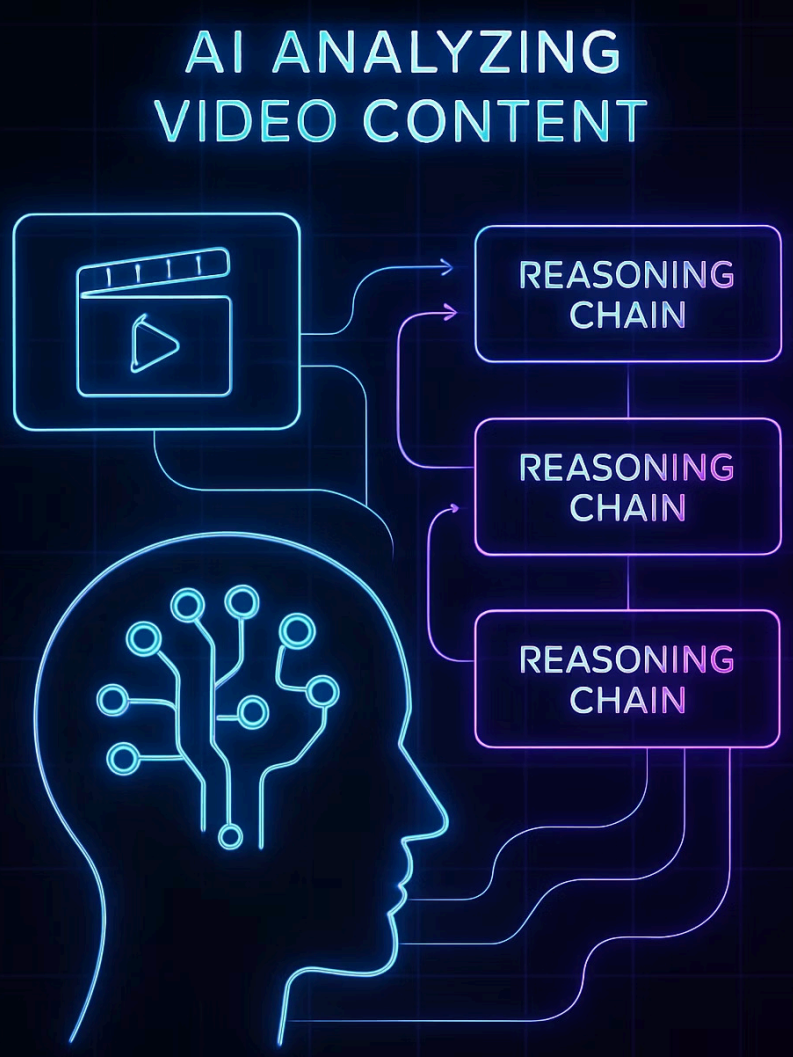


SEEK DEEP

Towards Advanced Video Understanding: An Agentic Chain-of-Chain-of-Thoughts (CoCoT) Framework

Team SeekDeep introduces CoCoT → Chain-of-Chain-of-Thoughts
Bringing deeper, adaptive reasoning to video analysis.

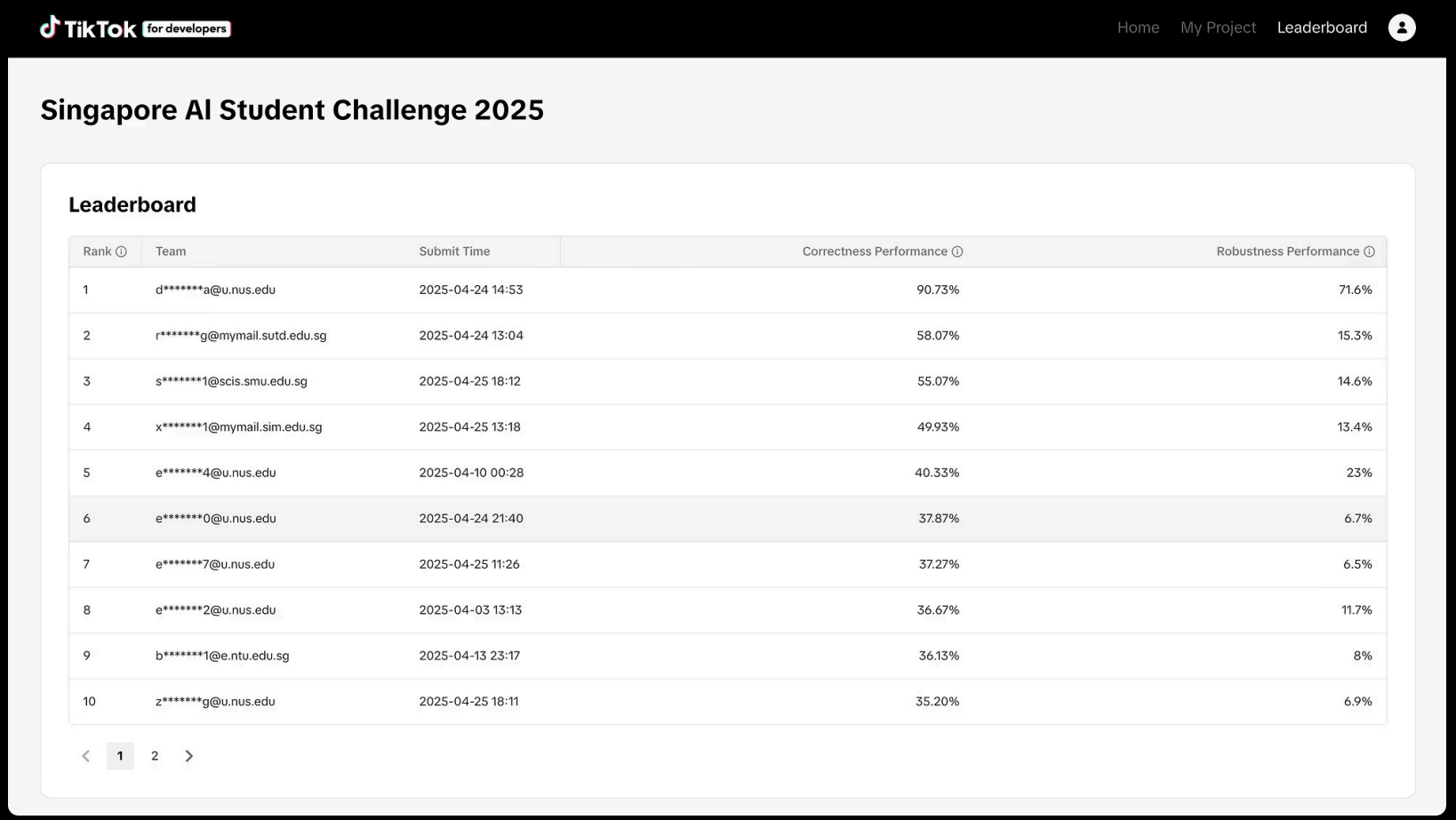
Dylan Chia
Team Leader, AI Infra and Archi
NUS CEG + AI Researcher

John Henry
AI Archi, Modelling
SUTD DAI + AI Researcher

Hai Le
AI Archi, Modelling, Tuning
SUTD CSD + AI Researcher

Leng Yunze
Video Experimentation
NUS CDE AI PhD Candidate

Leaderboard dominating performance



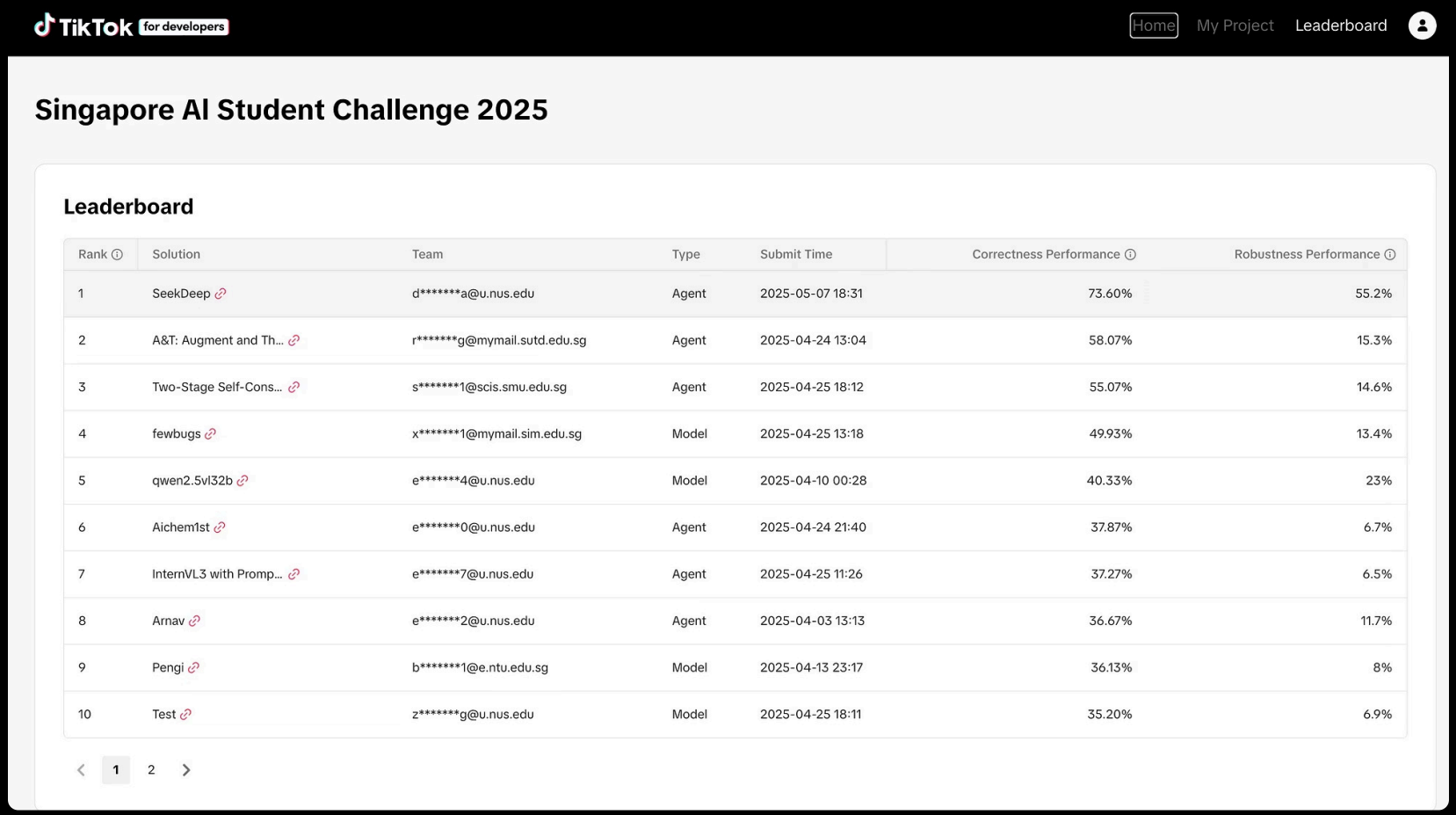
Rank	Team	Submit Time	Correctness Performance	Robustness Performance
1	d*******a@u.nus.edu	2025-04-24 14:53	90.73%	71.6%
2	r*******g@mymail.sutd.edu.sg	2025-04-24 13:04	58.07%	15.3%
3	s*******1@scis.smu.edu.sg	2025-04-25 18:12	55.07%	14.6%
4	x*******1@mymail.sim.edu.sg	2025-04-25 13:18	49.93%	13.4%
5	e*******4@u.nus.edu	2025-04-10 00:28	40.33%	23%
6	e*******0@u.nus.edu	2025-04-24 21:40	37.87%	6.7%
7	e*******7@u.nus.edu	2025-04-25 11:26	37.27%	6.5%
8	e*******2@u.nus.edu	2025-04-03 13:13	36.67%	11.7%
9	b*******1@e.ntu.edu.sg	2025-04-13 23:17	36.13%	8%
10	z*******g@u.nus.edu	2025-04-25 18:11	35.20%	6.9%

Initial CoCoT (Heuristics) Score

Correctness: 90.73%

Robustness: 71.6%

Leaderboard dominating performance



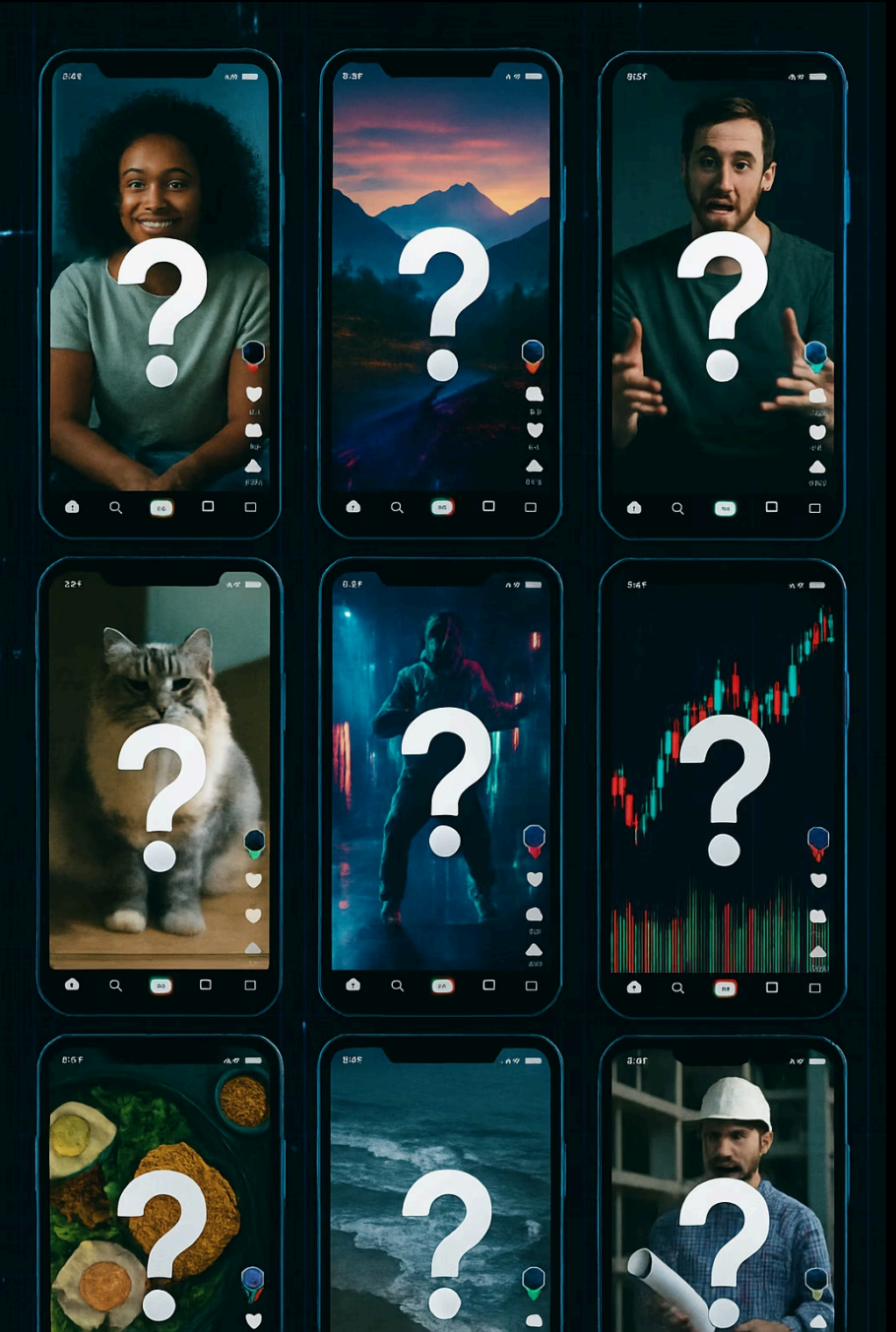
Rank	Solution	Team	Type	Submit Time	Correctness Performance	Robustness Performance
1	SeekDeep	d*******a@u.nus.edu	Agent	2025-05-07 18:31	73.60%	55.2%
2	A&T: Augment and Th...	r*******g@mymail.sutd.edu.sg	Agent	2025-04-24 13:04	58.07%	15.3%
3	Two-Stage Self-Cons...	s*******1@scis.smu.edu.sg	Agent	2025-04-25 18:12	55.07%	14.6%
4	fewbugs	x*******1@mymail.sim.edu.sg	Model	2025-04-25 13:18	49.93%	13.4%
5	qwen2.5vl32b	e*******4@u.nus.edu	Model	2025-04-10 00:28	40.33%	23%
6	Aichem1st	e*******0@u.nus.edu	Agent	2025-04-24 21:40	37.87%	6.7%
7	InternVL3 with Promp...	e*******7@u.nus.edu	Agent	2025-04-25 11:26	37.27%	6.5%
8	Arnav	e*******2@u.nus.edu	Agent	2025-04-03 13:13	36.67%	11.7%
9	Pengi	b*******1@e.ntu.edu.sg	Model	2025-04-13 23:17	36.13%	8%
10	Test	z*******g@u.nus.edu	Model	2025-04-25 18:11	35.20%	6.9%

1 2

Finalised CoCoT (Exploratory) Score

Correctness: 73.6%

Robustness: 55.2%



SEEK DEEP

The Challenge: Complex Reasoning nuance in Short-Form Videos

Information Density

Short videos are packed with content requiring complex reasoning for accurate VQA.

Tricky Questions

Current models struggle with multi-step inference across diverse content.

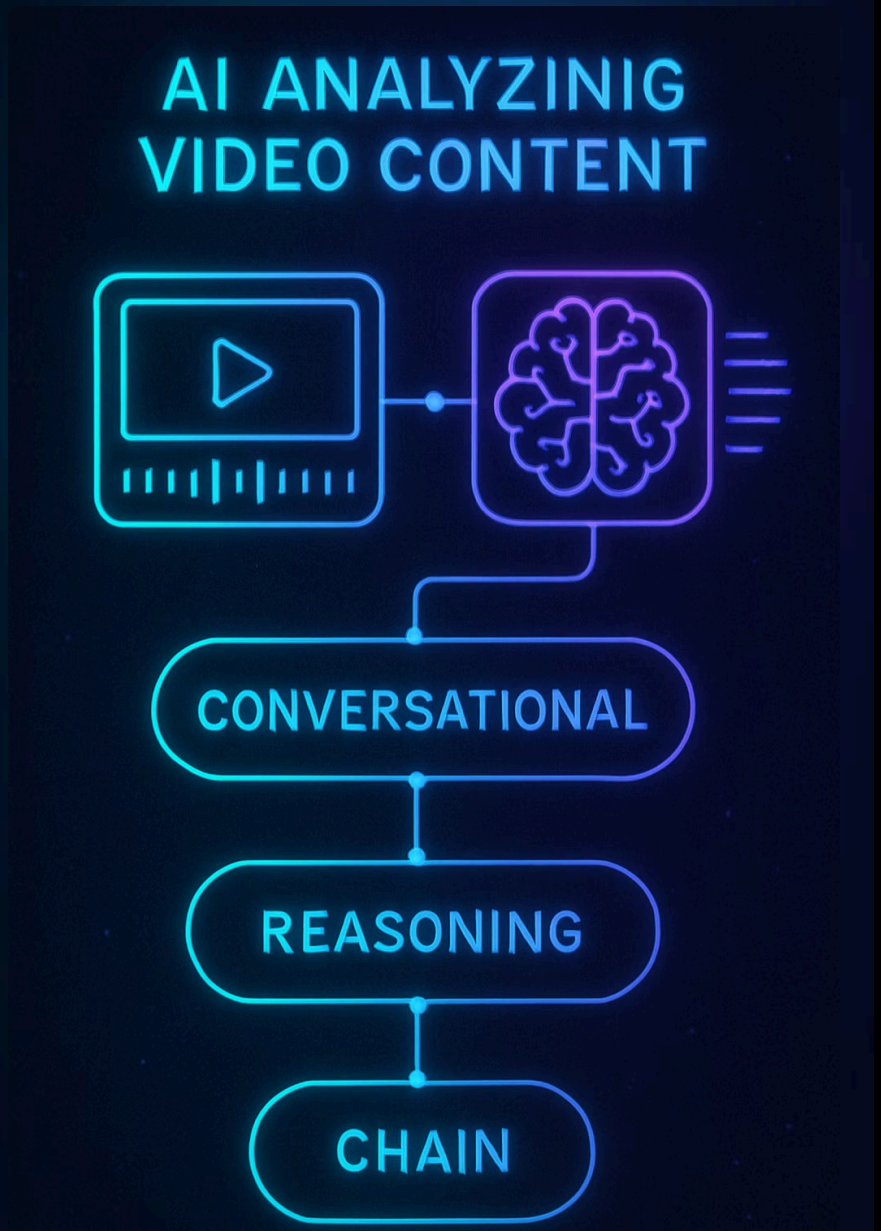
Scalability Challenges

Real-world volumes demand efficient processing solutions.

SEEK DEEP

Starting Point: The Power of Prompted Chain-of-Thought (CoT)

Model Approach	Correctness	Robustness	Notes
🍦 Vanilla Gemini 2.5 Pro	**~60%**	**~20%**	*Robustness = % of question sets fully correct*
🧠 Gemini 2.5 Pro + CoT Prompting	**67.6% (↑7.6%)**	**51.2% (↑31.2%)**	Model articulates step-by-step reasoning before the final answer.



SEEK DEEP

Agentic CoCoT (Heuristic): Unveiling the Framework's Potential

Video Input

Starting point for analysis

Correctly Led Question

Correctly led query about video content (sourced from provided dataset)

Conversational Chain

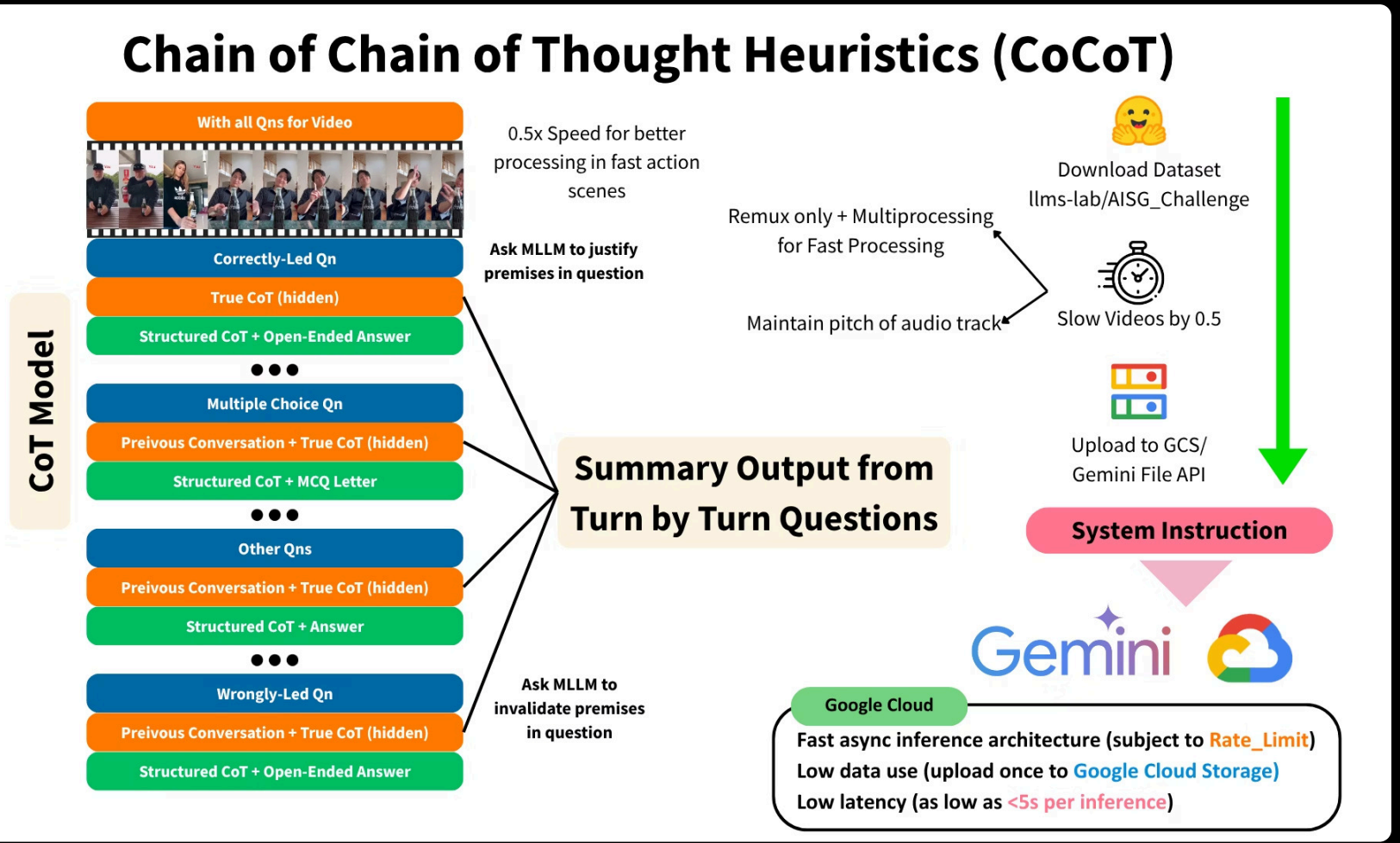
Sequence of other question types, ending with wrongly led

Final Answer

Comprehensive response with incredible correctness and robustness

Our Heuristic CoCoT version achieved **90.73% correctness and 71.6% robustness**, proving the potential of chained conversational reasoning.

Chain of Chain of Thought Heuristics (CoCoT)



Model Diagram

CoCoT Heuristics

Agentic CoCoT (Exploratory): Autonomous Reasoning with Generated Questions

Video + Question

Initial input requiring complex analysis

Agent Generates Exploratory Questions (Ex-Qns)

Model formulates exploratory questions for a specific video

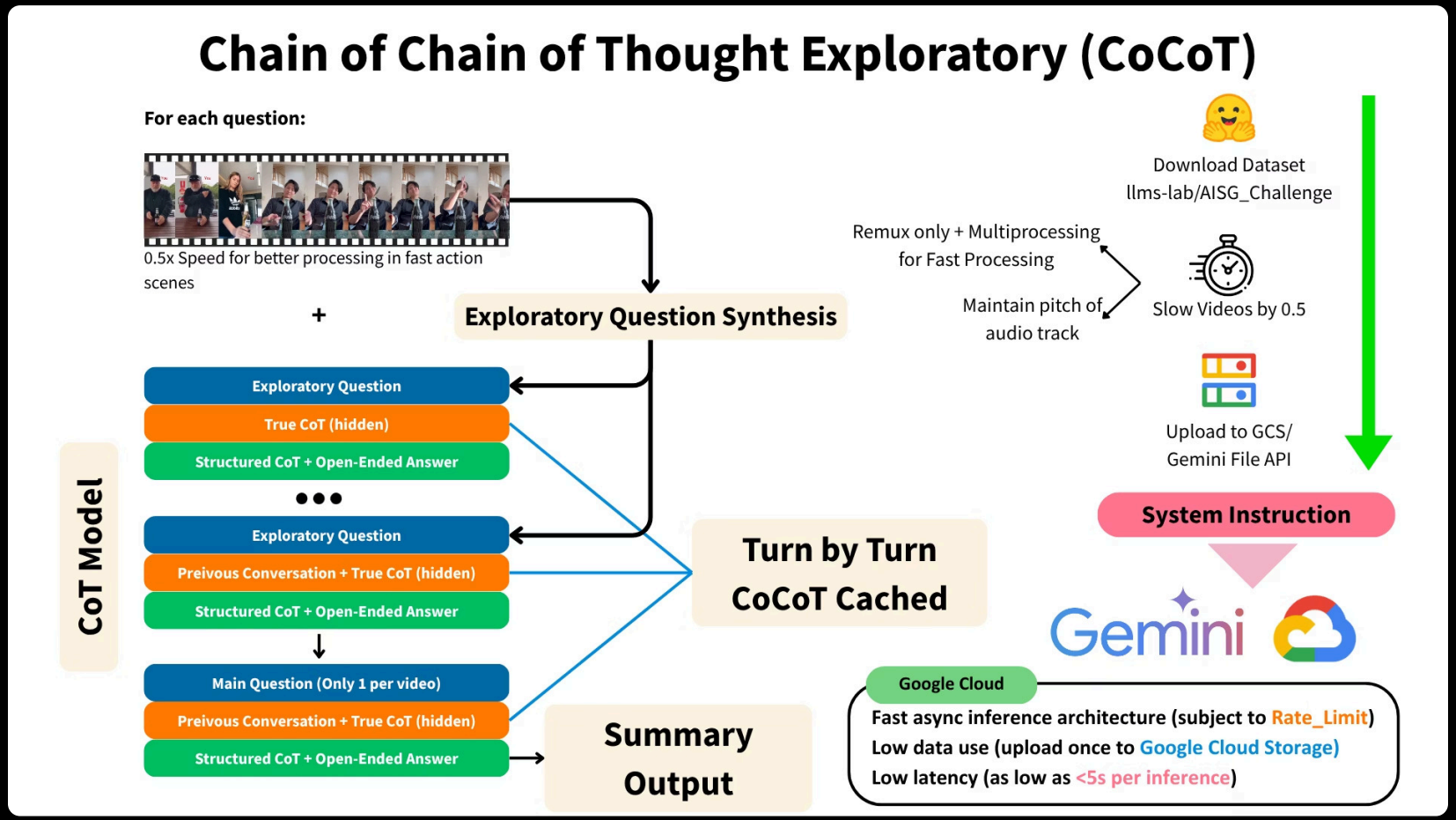
Answer Video Q1 (CoT)

Step-by-step reasoning for first VQA Question with Ex-Qns CoCoT

Generate for Next Question

Using previous CoToT from Ex-Qns to guide subsequent VQA questions for the same video

Our core innovation achieves **73.60% correctness and 55.2% robustness**, outperforming standard CoT prompting while adapting to diverse scenarios in real world applications.



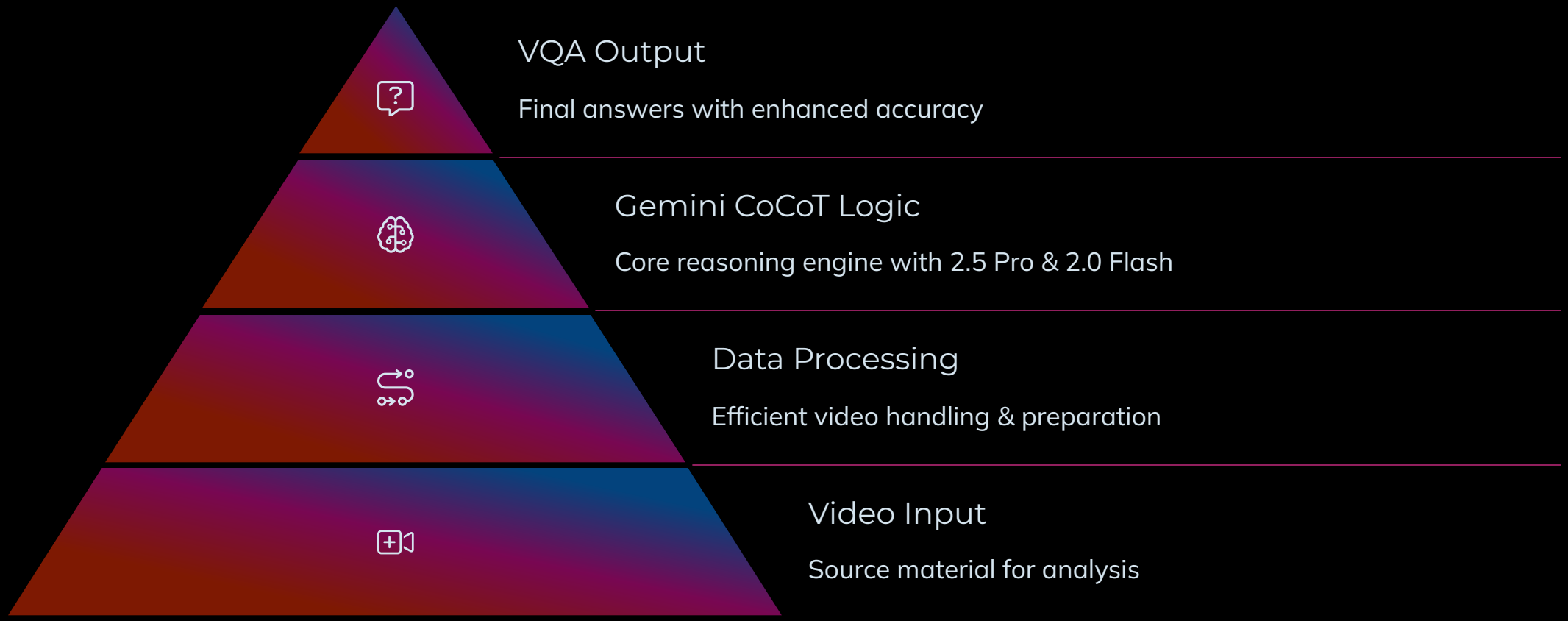
Model Diagram

CoCoT Exploratory

Evaluation Summary: The Journey to Enhanced Understanding

Method	Correctness (%)	Robustness (%)	Key Characteristic
Gemini 2.5 Pro (Vanilla)	60.0	20.0	Baseline (Video slowdown + normal optimized prompt Only)
CoT Prompting in Reasoning Models	67.6	51.2	Single-step explicit CoT prompting in reasoning models
CoCoT (Heuristic)	90.73	71.6	Agentic Multi-step, correctly led-guided, limited to competition dataset
Agentic CoCoT (Exploratory)	73.60	55.2	Agentic Multi-step, agent-generated ex-qns, applicable to Real World Applications

System Architecture: Enabling CoCoT



Our modular design supports Vertex AI & Gemini APIs, optimized for speed, extensibility, and benchmark performance.

Step 1: Data Processing Pipeline

Video Source Acquisition

Acquire videos and metadata from HuggingFace.

Extraction

Extract video content from compressed files for processing.

Temporal Enhancement

Apply video slowdown using FFMpeg (optimized async + remux only processing) to increase temporal resolution for catching subtle cues.

GCS Upload

Prepare and upload to Google Cloud Storage for fast, repeated access.

Step 2: Fast & Efficient Inference

Async Scalability

Maximizing API rate limits through async parallel uploads & inference, saturating rate limit

Minimizing bandwidth with GCS: upload once, infer many times

Intelligent Caching

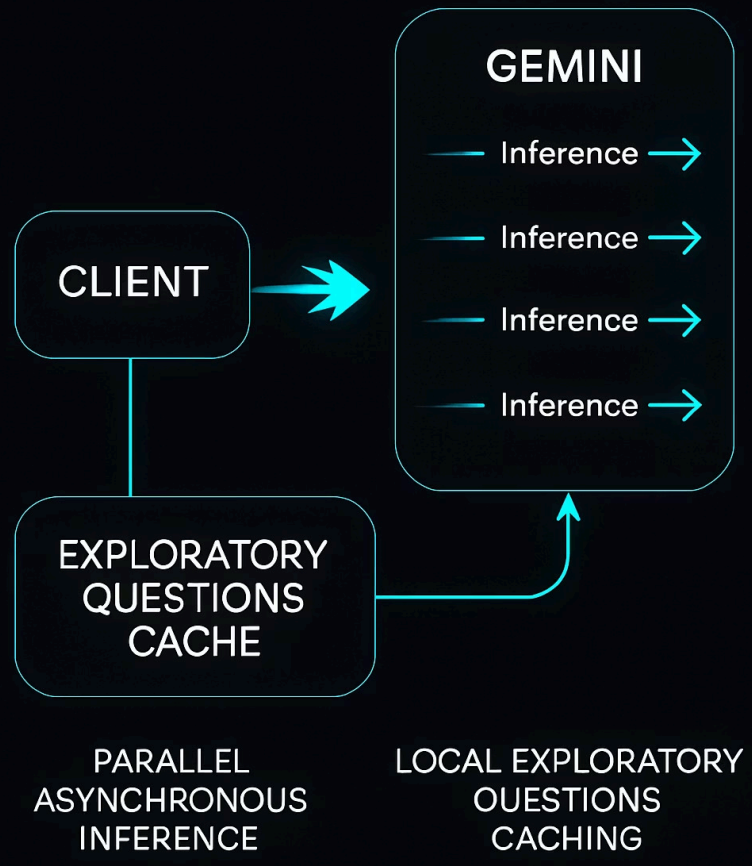
Ex-Qns CoCoT reasoning chain cached per video

Pre-warmed for VQA; real-time: first inference populates video exploratory questions cache

Optimized Latency

Exploratory Qns Inference (Gemini Flash 2.0): **< 3 seconds**

VQA CoCoT Reasoning (Gemini 2.5 Pro): **~15 seconds**



Conclusion & Next Steps

Leaderboard topping benchmark performance

Significant improvements in correctness and robustness over SOTA and other teams.

Innovative and Novel advanced Framework

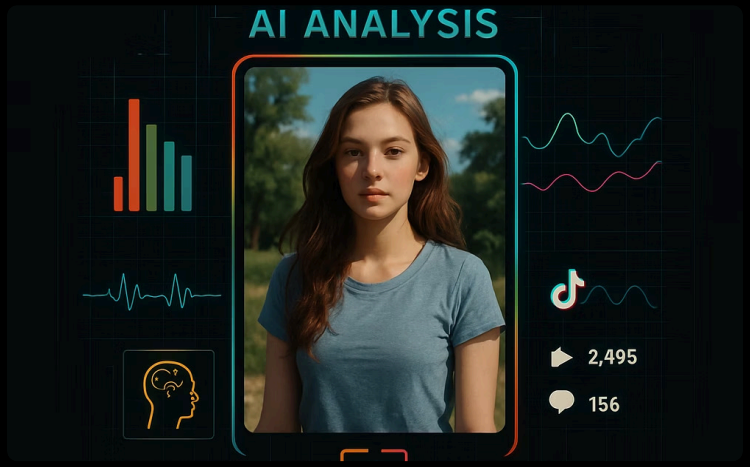
Agentic CoCoT advances deep multimodal understanding for videos.

Optimized High Performance and Efficient Scalable System

Fast, scalable solution leveraging GCS (reducing bandwidth) and parallel async processing, and also caching for exploratory questions.

Our system is ready for complex video analysis and Q&A, demonstrating the future of intelligent video understanding.

CoCoT in Action: Demo



Sample Video

TikTok-style content/ Youtube short video requiring complex understanding

Generated Questions

Agentic CoCoT autonomously creates exploratory sub-questions