TRACK 1: AMD AI PREMIER LEAGUE (AAIPL)

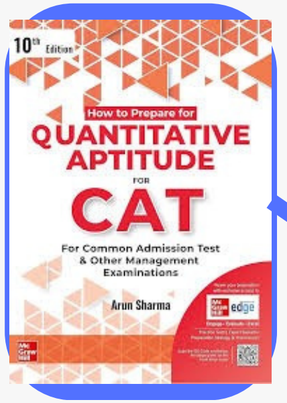
AMD AI REINFORCEMENT LEARNING HACKATHON IIT DELHI

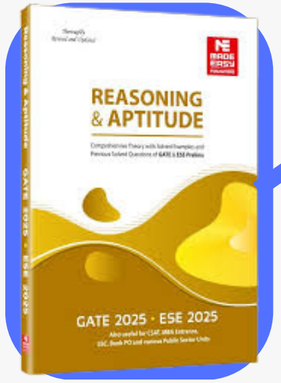
BY TEAM 32 : CONVULSIONS

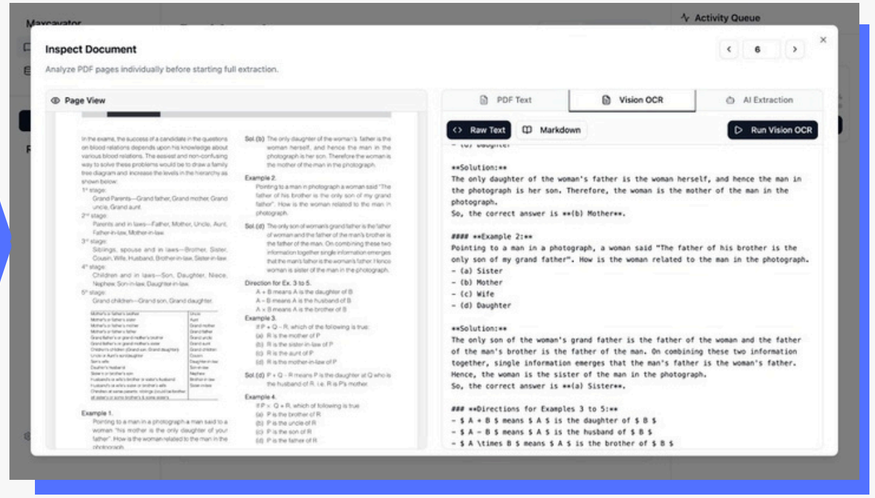
DATASET


```
{
  "questions": [
    {
      "topic": "Syllogisms",
      "question": "Some roses are white. All lilies are white. No white flower is red. Which is certain?",
      "choices": [
        "A) Some roses are lilies",
        "B) No lily is red",
        "C) No rose is red",
        "D) All white flowers are roses"
      ],
      "answer": "B",
      "explanation": "All lilies are white, and no white flowers are red. Therefore, no lily can be red."
    },
    {
      "topic": "Blood Relations and Family Tree",
      "question": "Q: 'R is the mother of my father.' What is R to Q?",
      "choices": [
        "A) Grandmother",
        "B) Mother",
        "C) Aunt",
        "D) Great-grandmother"
      ],
      "answer": "A",
      "explanation": "Father's mother = Grandmother."
    },
```


- Generated Synthetic Data using various LLMs: Gemini, Mistral, Deepseek, GPT, etc.
- The questions generated are from the 4 topics provided.







- We made an OCR Extraction Tool to extract the data from PDFs of major competitive exams like GATE, CAT, which ask such questions, and transformed the data into the provided JSON format.

TRAINING PIPELINE

Technique Used: GRPO Reinforcement Learning

AAIPL requires optimization across:

Requirement	How GRPO Helps
Strict JSON	Reward function enforces it
Token limit	Reward penalizes overflow
Correct answer	Weighted correctness reward
Logical explanation	Self-verification reward
Multi-objective behavior	GRPO handles combined rewards

TECHNIQUE

We converted competition rules into mathematical reward signals.

Technique	What It Does	Strength	Limitation for AAIPL	Why GRPO is Better
SFT	Learns from labeled examples	Simple and stable	Cannot enforce token limits or	GRPO optimizes directly using
PPO	RL with policy + value model	Powerful general RL	More unstable; sensitive to	GRPO simpler, more stable for
Prompt Tuning	Trains input embeddings only	Lightweight	Cannot deeply modify reasoning	GRPO updates policy behavior
Distillation	Mimics teacher model	Transfers knowledge	Cannot encode competition-	GRPO allows custom rule-based
GRPO (Our Choice)	Compares multiple outputs	Stable multi-objective	Slightly more complex than SFT	Directly optimizes for JSON, token

REWARD FUNCTIONS FOR GRPO:

1. **Valid JSON**
   - Ensures strict JSON format with required fields and 4 choices.
2. **Answer Format**
   - Forces answer to be only A / B / C / D.
3. **Self-Verification**
   - Checks if explanation logically justifies the answer.
4. **Token Limit Compliance**
   - Ensures core content ≤ 150 tokens (competition rule).
5. **Answer Correctness (Weighted)**
   - Rewards matching dataset answer (stronger signal).
6. **Explanation Quality**
   - Encourages concise, meaningful reasoning (10–100 words).

MODEL USED

Model Used

QWEN/QWEN2.5-14B-INSTRUCT

Why?

- High performance on reasoning benchmarks
- Strong multi-step inference ability
- Better constraint handling than smaller 7B models

AAIPL requires:

- Strict JSON
- Exactly 4 choices
- A/B/C/D answer only
- Token limit enforcement

All this is delivered by QWEN2.5-14B

CRITERION	Technical Strength	Impact in AAIPL Competition
Reasoning Depth	Strong multi-step logical inference capability	Generates harder LR traps and maintains consistency in complex reasoning
Structured Output	High adherence to instruction-following & JSON format	Reduces formatting errors and prevents disqualification
RL Stability	Stable under multi-objective GRPO reward shaping	Handles 6 reward functions without reward collapse

THANK YOU

BY TEAM 32 CONVULSIONS