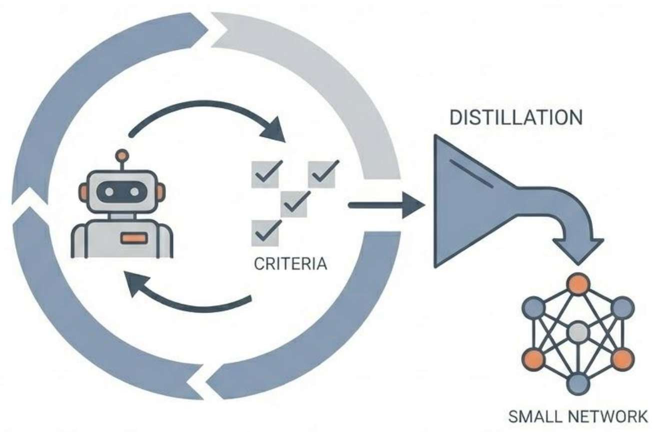
Evaluator-Guided LLM Distillation for Embodied Agent Decision-Making

AxisTilted2 - 1st Place @ EAI Challenge 2025
Chin Pradeep – NYU Neuroinformatics Lab (chinpradk@gmail.com)
Sanjayan Sreekala – Independent Researcher (san@sankala.me)



Overview

Data!

- How can we prepare gold data?
- Evaluator-in-the-loop training data preparation

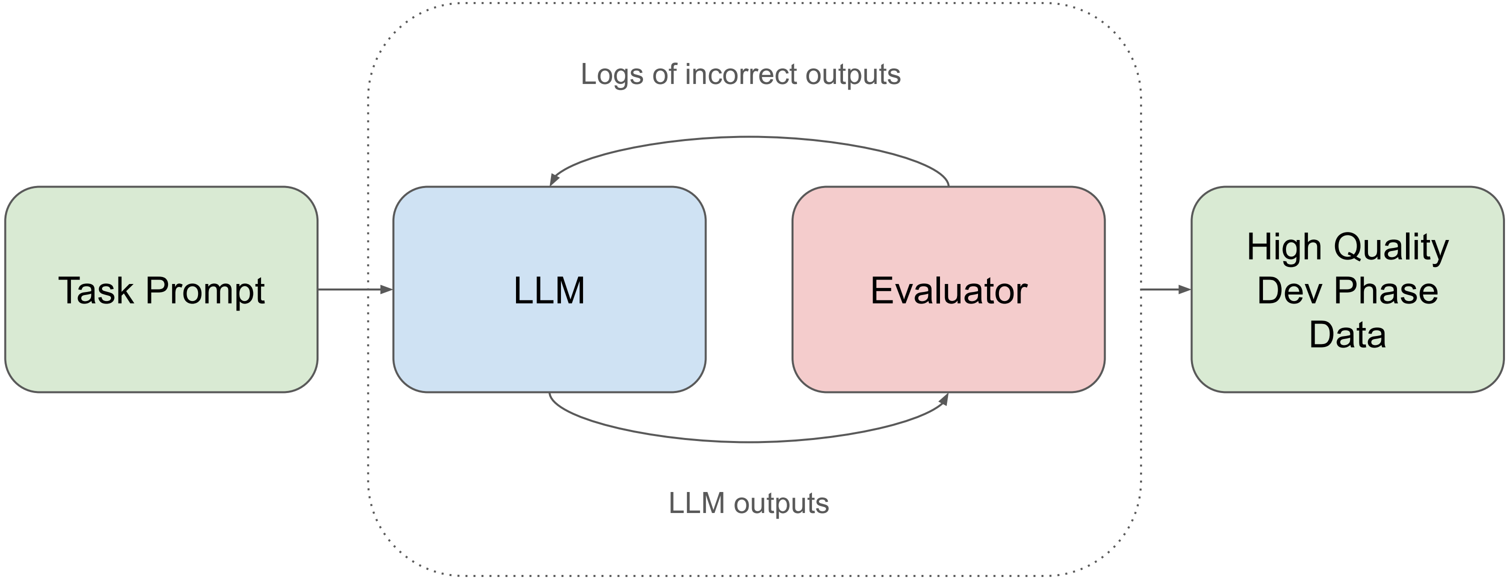
Fine-tuning

- Fine-tuning Qwen models with gold data

Learned Evaluator

- Evaluators are really useful, can we make an LLM behave like one even when there is no gold data?

Data Preparation



Fine-tuning

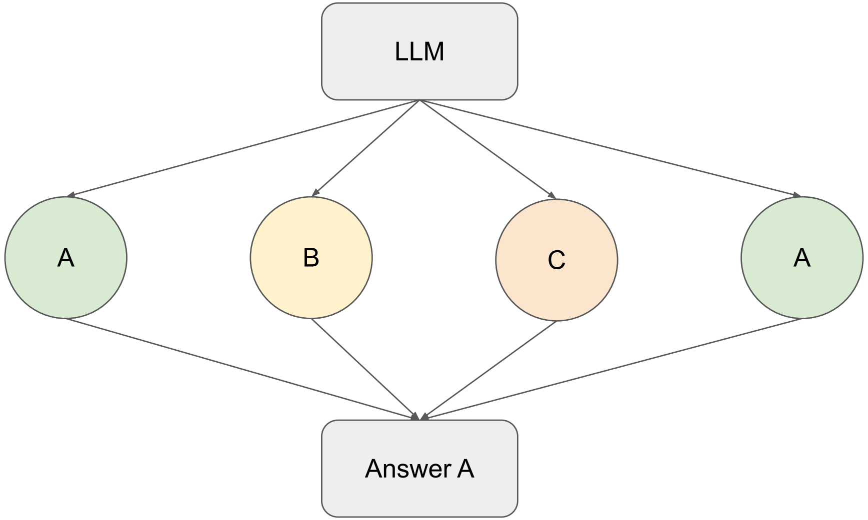
- **Backbone family: Qwen3**
  - Small to mid-size: **0.6B, 4B, 8B** (full-weight SFT)
  - Large: **32B** with **LoRA** on 2 x H100s
- **Domain adaptation**
  - Pre-finetune on prompts from **all modules + both environments**
  - Train with **next token prediction on all prompts** → forces it to internalize interface & vocabulary
- **Cross-task / cross-dataset variants**
  - Joint GI+SD+AS training and mixing **BEHAVIOR + VirtualHome**
  - Helpful for VirtualHome GI; more mixed for SD/AS
- RL (GRPO) for VirtualHome GI → unstable and less effective than SFT

RAG - VSD

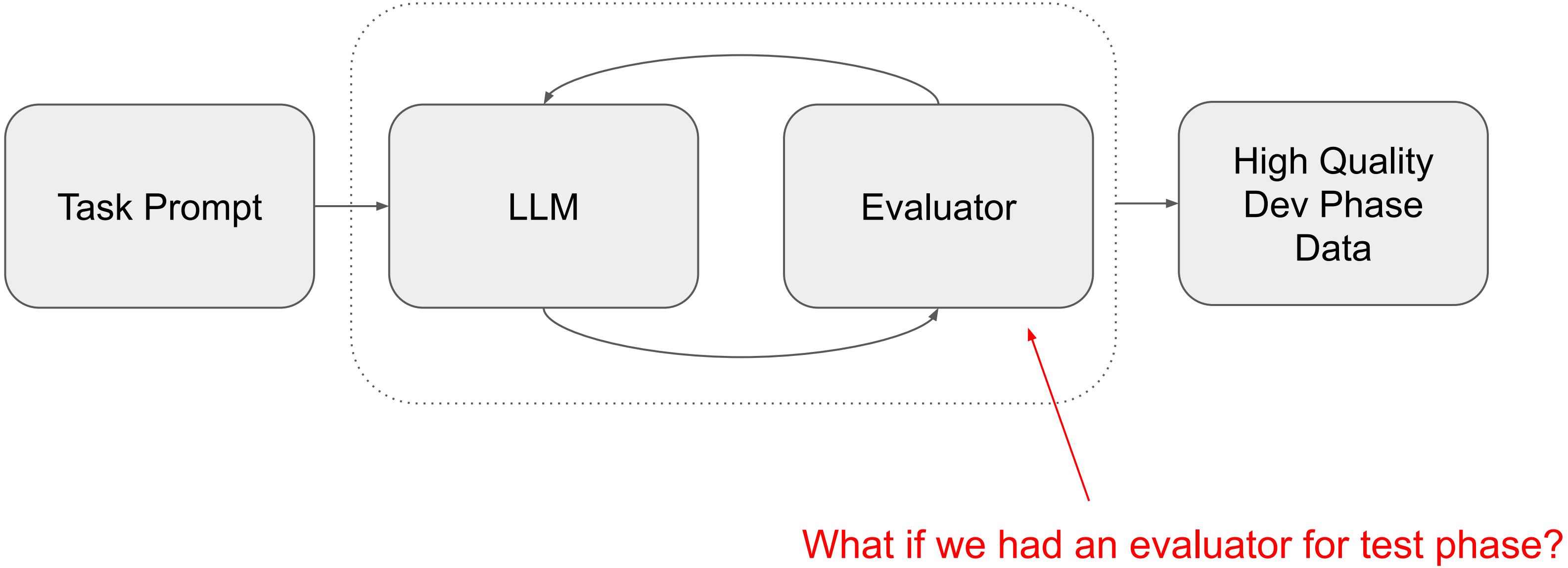
- Used **for** VirtualHome **Subgoal Decomposition (SD)**
- **Semantic retrieval for few-shot prompts**
  - Embed task names with **all-MiniLM-L6-v2**
  - Retrieve **K = 5** nearest training tasks based on task name
- **Models**
  - RAG on **Gemini 3.0** gave the strongest SD performance
- Intuition:
  - Retrieved examples capture **style + conventions** of good decompositions
  - RAG gives a cheap way to inject **knowledge about the task** without fine-tuning the model
- Result: **74.5%** → **77.6%** Task Success Rate

Voting - VGI

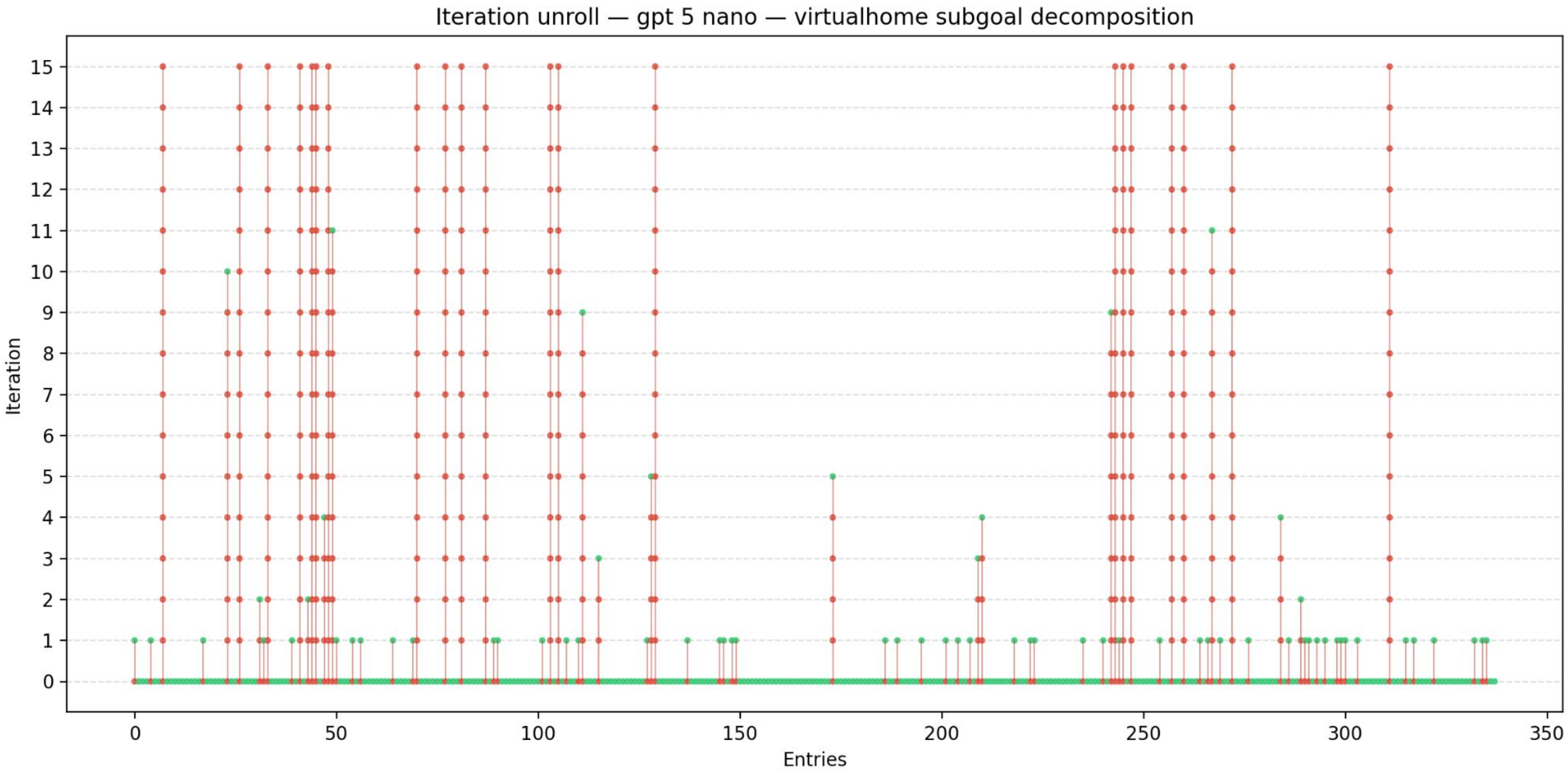
- Sample N outputs candidate node/edge/action, include it **only if it appears in at least k of N samples** (we sweep k = 1, 2, 3, 4, N=4).
- Also tried asymmetric thresholds, e.g. require 4/4 agreement on nodes but only 2/4 for edges/actions.
- Effects in VirtualHome GI; GPT-5-mini samples, with stricter node threshold: **50.9 → 55.3 F1 (+4.4).**
- For fine-tuned Qwen3-4B on the eval split, voting gives a smaller bump **(65.2 → 65.4 F1)**



Data Preparation



Data Preparation



Finetuned LLM-as-an-evaluator

Goal: learn a **student evaluator** that mimics the official EAI evaluator's feedback

Data construction

- Run evaluator-in-the-loop refinement for with multiple model outputs of varying quality
- For each iteration, collect:
  - Prompt + output
  - Official evaluator feedback (errors / scores)
- Difficult tasks naturally show up **more often** (more failed iterations for these tasks)

Training

- Finetune **Qwen3-32B** on this triplet: (prompt, candidate output, evaluator feedback)
- Model learns to output evaluator feedback

Finetuned LLM-as-an-evaluator

Deployment

- Use **Gemini 3.0 + RAG** as primary SD generator
- LLM evaluator reviews output, suggests corrections or passes it through
- Add **cross-model confidence filter**:
  - Dataset had a 51-49 split of success vs error log
  - Only refine when evaluator disagrees with our best finetuned SD model
  - If both agree but evaluator flags “wrong,” we treat it as a likely false negative

Impact

- **+1.5%** VirtualHome SD task success rate from evaluator feedback alone
- Final **78.7%** SD task success with evaluator + confidence filter (+0.2%)

Other Techniques Explored

Cleaner/VirtualHome-style prompts for BEHAVIOR

- Rewrite BEHAVIOR GI/SD prompts into cleaner **markdown-like** structure → **+10–20 points** in some BEHAVIOR metrics

Two-step reasoning for VirtualHome GI (~ +6%)

- Step 1: free-form **natural-language solution**
- Step 2: restate in exact **JSON goal** format

Format simplification

- Treat plans as ordered **lists**, not JSON objects with artificial keys in VAS
- Reduces bookkeeping overhead for “thinking” models

Agent-style scaffolds (explored but not primary)

- LLM-as-a-judge for BEHAVIOR AS
- External state-management tools

Results at a glance

Overall

90.09 (next best: 84.32)

BEHAVIOR Goal Interpretation
99.60

BEHAVIOR Subgoal Decomposition
97.00

BEHAVIOR Action Sequencing
98.00

BEHAVIOR Transition Modeling
99.50

VirtualHome Goal Interpretation
65.40

VirtualHome Subgoal Decomposition
78.70

VirtualHome Action Sequencing
82.60

VirtualHome Transition Modeling
99.85

Takeaways

- Simpler is better (might just be a data problem)

- Combining diverse model intuitions was useful

- Language is maybe not the best medium