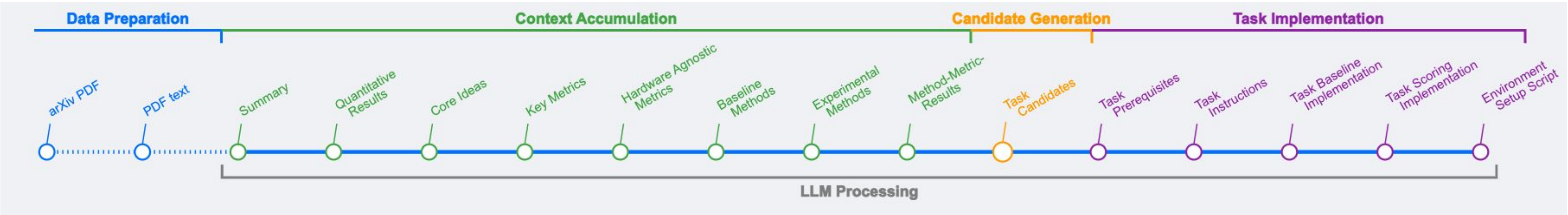
ATEFAR: Automated Task Extraction From Actual Research

or: arXiv is All You Need

Jai Dhyani, MATS 6.0
Mentor: Hjalmar Wijk (METR)
Special Thanks: Sami Jawhar (METR)



Note: ATEFAR is under active development; this diagram reflects the most recent pipeline as of August 20 2024

The Case for AI R&D Evals

- To measure progress towards AI fully automating AI R&D, we need **AI R&D evals**
- But AI R&D Evals are **hard** to develop
  - Challenging, **time-intensive** tasks
  - **Few humans** can successfully complete them, making validation/QA difficult, slow, and expensive
  - Tasks must be **novel** to avoid the risk of memorization
  - But also measure **actual research skills**
  - There are so **many relevant skills** to measure that it's hard to even identify them all, let alone develop tasks for them

Proposal: ATEFAR

- Papers detailing AI R&D **research methodology** and **results** are published on arXiv **every day**
- Modern LLMs can engage with research papers (e.g. **summarization, Q&A**) and **generate code** given prompts
- We can use LLMs to extract research tasks **directly from papers**
- By running this continuously, we can build a **living AI R&D eval suite**
- A continuously-updated task suite lets us **bypass the novelty/memorization problem**: the most recent tasks will always be after the model's knowledge cutoff!
- Extracting tasks from a wide variety of papers should produce evals that test a **wide variety of skills** that are demonstrably vital to actual real-world research

Proof of Concept:
Minimum Viable Task Extractor

Goal: Extract at least one task as a proof of concept. Focus on tasks where:

- There is some **'baseline' implementation** (e.g. a model training pipeline)
- Baseline has a score on some **metric** (e.g. iterations of training to achieve some target accuracy on the test set)
- a **method** is proposed to improve the baseline
- after implementing the new method, the updated **metric** is reported

Focus on papers which:

1. have low computational requirements
2. were published recently (after knowledge cutoff dates for SOTA LLMs) a
3. have publicly-available can that could be used to validate that generated tasks were being scored accurately.

Main test paper: "**94% on CIFAR-10 in 3.29 Seconds on a Single GPU**" (Jordan, 2024)

ATEFAR Generated Task Components

- **Instructions** that describe the task
- **Scoring function** which evaluates the solution
- Where applicable, **a baseline implementation** to serve as a starting point for the agent
- **Environment setup** including all of the above plus libraries, key datasets, etc

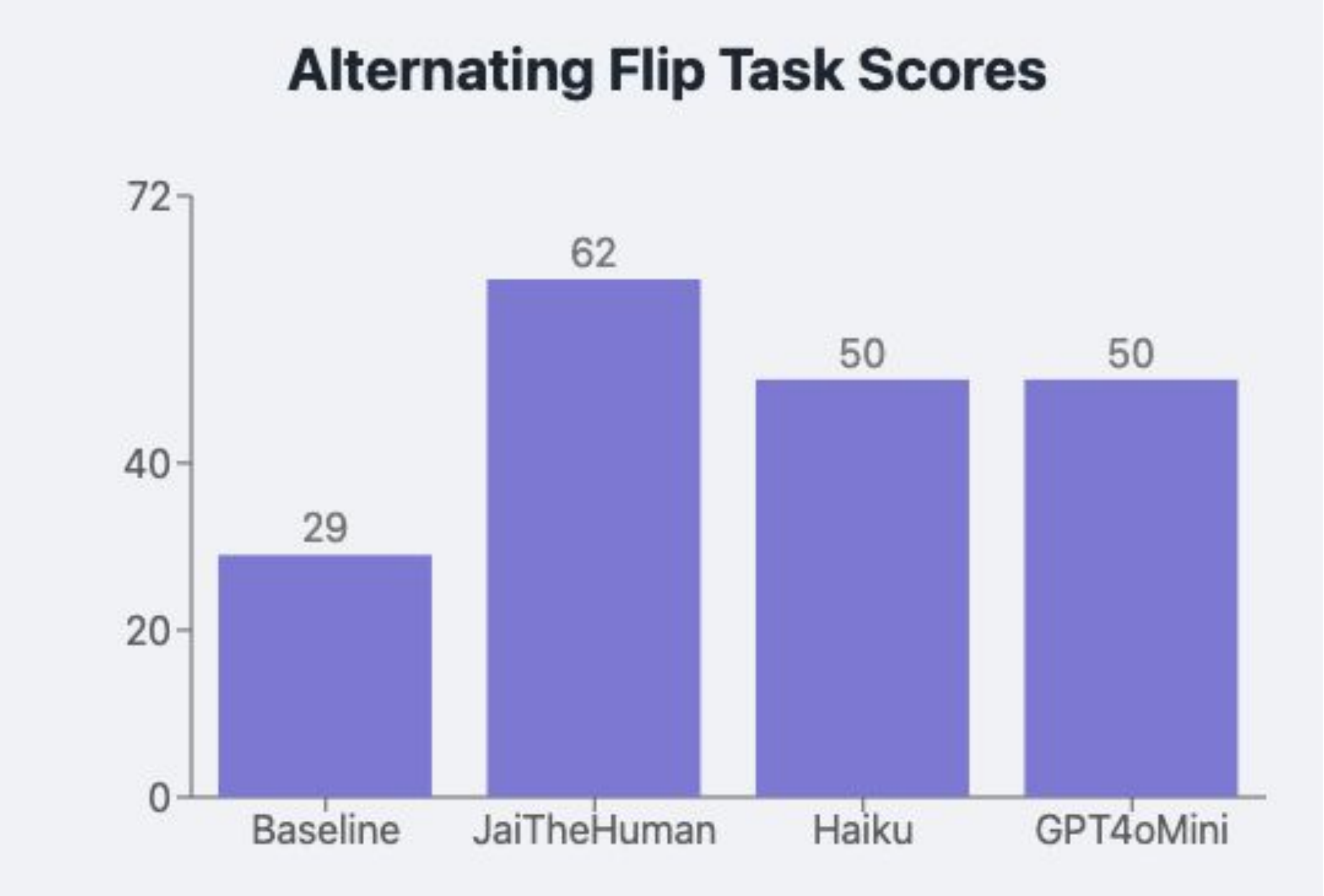
https://github.com/jaidhyani/atefar

End-to-End Working Eval Example

The first fully-functional eval extracted by the ATEFAR directly from a the full text of a research paper is "Alternating Flip Augmentation".

ATEFAR autonomously identifies and implements the task, including instructions, scoring function, baseline implementation, and environment setup script.

Generated scoring function evaluation of the provided baseline, my solution, and solutions from Haiku and GPT4oMini:



Next Steps & Roadmap

I've been working on ATEFAR for two and a half weeks. Next priorities:

1. Extract a task which Claude Sonnet cannot solve but a human expert can (Estimated Time to Complete: 2-10 days). This is the **key capability required to make ATEFAR useful,** and the **most likely point of near-term failure.**
2. Demonstrate **consistent extraction** of tasks (ETC: 7-28 days)
3. **Automate task validation** by testing if Sonnet-proof tasks become solvable with access to the paper's codebase (ETC: 7-28 days)
4. Expand to **1-5 other task types** (ETC: 1-14 days per task type)
5. **Integrate with existing Evals framework** (e.g. automatically export tasks for Vivaria or Elicit) (ETC: 1-7 days)
6. Implement low-cost **filter** to identify papers which are **good candidates for task extraction** (ETC: 1-7 days)
7. **Benchmark** existing LLMs across dozens of extracted tasks (ETC: 1-7 days)
8. Automate running filter/pipeline on AI research papers **as they're published to arXiv** (ETC: 1-14 days)