Final project report

WANG TZU YI

December 5, 2025

1. Future AI Capability: Human-Level Visual-Abstract General Reasoning

I argue that in the next 20 years, AI systems may achieve a fundamentally new capability:

> **Human-level visual–abstract general reasoning: an AI system capable of inferring concepts and rules from extremely few visual examples (few-shot), and applying them robustly across novel contexts and even across entirely different domains (e.g., from 2D grid patterns to 3D physical object arrangements or robot manipulation).**

Why This Matters

Scientific Discovery and Engineering Given only a few experimental images (e.g., fluid microscopy, astronomical data, material microstructures), the AI can abstract latent geometric or physical rules and predict outcomes under unseen configurations. This dramatically accelerates scientific insight and experimental design.

Culture and Knowledge Transfer The AI can infer structural or stylistic principles from a small number of visual artifacts (e.g., traditional motifs, architectural decorations) and generalize them across cultures or design domains, enabling new forms of creative synthesis.

Robotics A robot endowed with such reasoning ability could observe a single demonstration and infer functional rules such as:

> "Place object X near Y," "Sort items by size," or "Rearrange components according to spatial relations."

It would then generalize these rules to new environments, tasks, and object types.

In other words, this capability is not about performing one task well, but about acquiring:

$$\text{abstraction ability} + \text{generalization} + \text{visual understanding} + \text{rule induction} + \text{transfer learning.}$$

2. Ingredients Required for This Capability

Achieving such general visual reasoning requires several interacting components.

2.1 Data

- **Multi-modal, diverse, few-shot task data**: Each task consists of a small set of input-output examples, testing whether AI can infer the rule and apply it to new inputs.
- **High-quality visual data (pixel- or object-level)**: Clear object boundaries, colors, shapes, spatial layouts, and geometric structures.
- **Structured annotations**: not only bounding boxes but also:
  - object identity,
  - part-whole relations,
  - spatial relations: adjacency, containment, symmetry,
  - hierarchical structure: composite objects,
  - transformation or action descriptions,
  - semantic labels (e.g., "handle", "chair leg", "repeated motif element").

2.2 Tools and Algorithms

- **Neuro-symbolic hybrid systems**: Combining neural perception with symbolic reasoning, structured representations, and program synthesis. Recent work explores Vector Symbolic Architectures (VSAs) for this purpose.
- **Vision–language synergistic reasoning architectures**: Visual modules capture spatial/structural patterns; language modules express rules, constraints, transformations, and verification steps.
- **Program synthesis + search/planning**: The AI must produce executable rules in a domain-specific language (DSL) and verify them.
- **Test-time adaptation**: Because each task is novel, models need few-shot adaptation or rule induction at inference time.

2.3 Hardware and Environment

- **High-performance multimodal compute**: Capable of handling vision, symbolic structures, and program search.
- **Interactive/simulation environments**: Required for tasks involving physical manipulation or feedback-based learning.

2.4 Learning Setup

- Mixed paradigms:
  - self-supervised visual representation learning,
  - meta-learning and few-shot learning,
  - neuro-symbolic learning + program induction + verification,
  - continual learning and cross-domain transfer.
- Reinforcement learning may be used when tasks involve interactions or consequences.

All components depend on one another: without structured data, neural-symbolic models cannot induce general rules; without a symbolic layer, neural models cannot express or manipulate abstract transformations; without environments, robots cannot use such reasoning for real-world tasks.

3. Relevant ML Paradigms

This capability requires a hybrid, multi-stage learning system.

Self-supervised / Unsupervised Learning

Needed for learning object representations, boundaries, spatial relations, symmetry, and other structural properties from large unlabelled corpora of images, 3D scans, or sensor data.

Meta-learning / Few-shot Learning

Crucial because each reasoning task contains only a tiny number of examples. The model must learn how to abstract rules efficiently.

Neuro-symbolic Learning + Program Induction

After representation learning, the system must induce executable symbolic rules. This resembles algorithm learning rather than classical supervised learning.

Reinforcement Learning (optional)

Applied when tasks involve actions, feedback, or physical consequences.

4. Solvable Model Problem: First Step via ARC-AGI

To make progress toward the long-term goal, we formulate a simplified toy problem inspired by the ARC-AGI benchmark.

4.1 Problem Definition

Input Format Each ARC task consists of 2–5 training pairs and one test input. Each grid is a 2D array of integers 0–9 representing colors.

```
{
  "train": [
    {"input": [[0,1,2],[3,4,5]],
     "output": [[5,4,3],[2,1,0]]},
    ...
  ],
  "test": [
    {"input": [[1,2,3],[4,5,6]]}
  ]
}
```

Output Format A predicted grid of identical format, matching the ground truth solution.

4.2 Data Transformation

Grids are converted into textual sequences for a language model:

```
Input:
[[0,1,2],[3,4,5]]
Output:
[[5,4,3],[2,1,0]]
```

Data Augmentation

- shuffle training examples,
- rotate grids (90/180/270°),
- mirror horizontally/vertically,
- swap colors.

4.3 Model Choice: GPT-2 with LoRA

- **Few-shot reasoning**: pre-trained LMs naturally perform sequence pattern inference.
- **Sequence flexibility**: grids of varying sizes become variable-length sequences.
- **Efficiency**: LoRA tunes only $\sim 0.6\%$ of parameters.
- **Interpretability**: the generated output reveals the model's inferred rule.

Why Not CNN, Linear Models, or Symbolic-only Approaches? CNNs require fixed-size inputs; linear models cannot express nonlinear transformations; symbolic systems require manual rule engineering and lack generalization ability.

5. Summary

- A major future milestone for AI is **human-level visual-abstract general reasoning**.
- Achieving it requires multimodal few-shot data, structured annotations, hybrid neuro-symbolic systems, strong hardware, and interactive environments.
- The core challenge is the pipeline:

  $$\text{pixels} \to \text{abstract representation} \to \text{symbolic rules} \to \text{generalization.}$$

- ARC-AGI inspired toy tasks provide a tractable first step for empirical investigation.

6. Experimental Pipeline

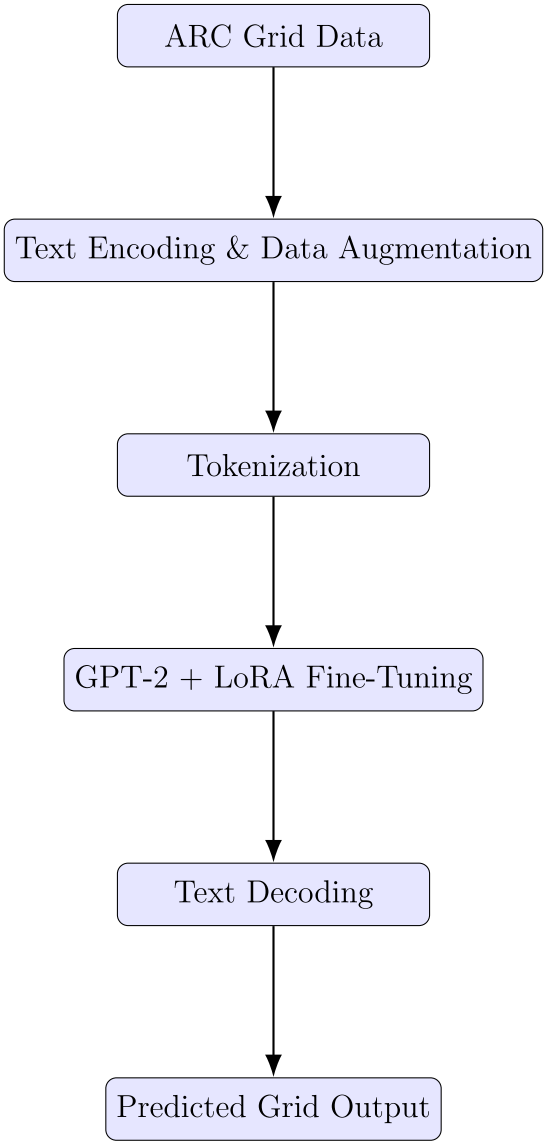

Figure 1: Pipeline for ARC Visual Reasoning Experiment