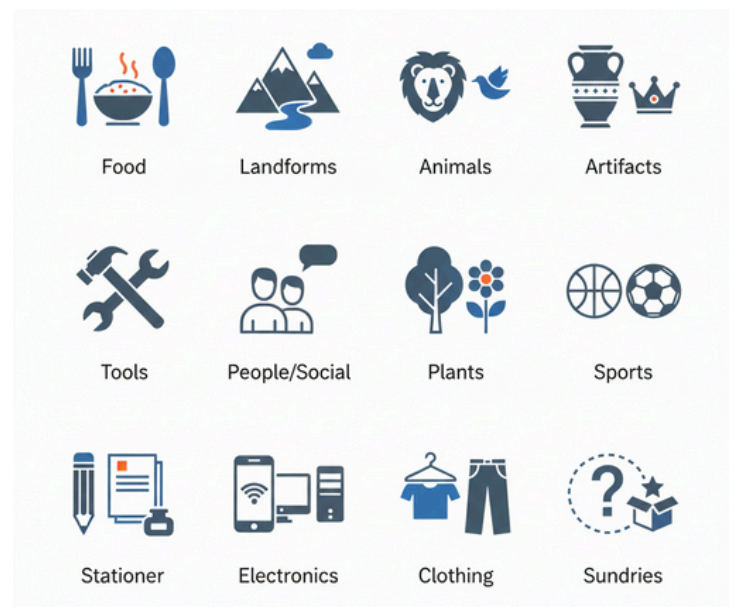
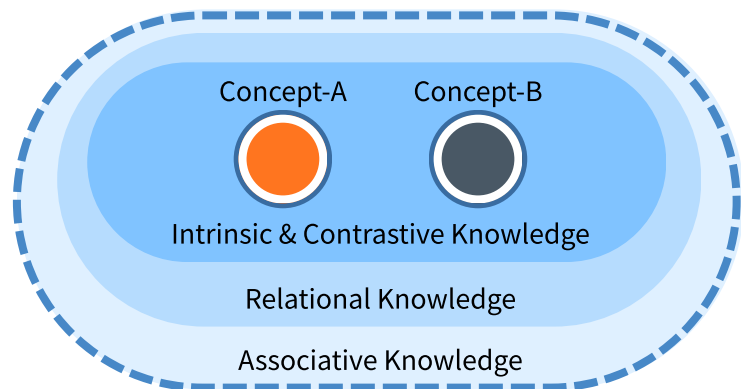


Concept Pair Dataset

Selected 220 daily-life concept pairs
grouped into 12 categories



Knowledge involved in the evaluation:



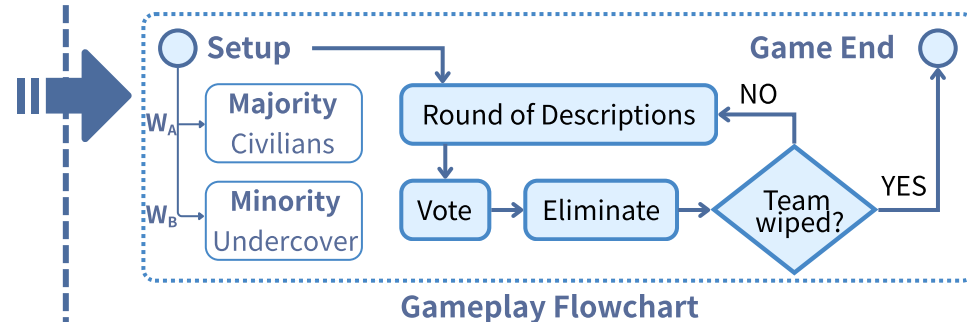
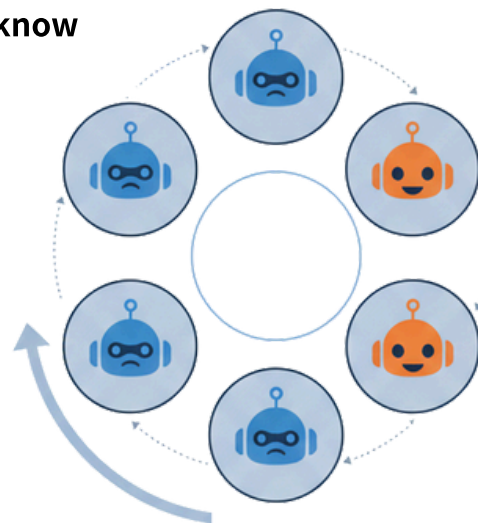
Multi-Agent Game:Undercover

What players do not know

Specific win condition
Another concept
Identity

Game rules
Assigned concept

What players know



- Cross-Concept Reasoning
- Knowledge Retrieval
- Strategic Description

What is their word? What do I know about my concept? How specific should I be?



Who am I? Where do our concepts overlap and diverge? How should I adjust based on new information?

- Identity Inference
- Boundary Recognition
- Adaptive Strategy

LLM capabilities under evaluation

Snapshot QA benchmark



Data from
game logs



Automatic
filtering & organization



QA benchmark

Outlier Statement Detection

Compared to *LAKE*, which sentence may describe *RIVER*?

- A. You can often hear the sound of gurgling water first, and then see the water surface again. ✓
- B. Coastal areas support diverse ecosystems and are widely distributed.
- C. Both natural and artificial environments can occur.

Cross-Concept Inference

Your concept is *ELEPHANT*, and there is someone described their concept like this:

“This animal is known for its ability to reach leaves in tall trees, which is crucial for its diet.” What is their concept?

A: It must be *GIRAFFE*. ✓

Fine-Grained Comparison

Which concept does the statement better describe, *DESERT* or *SANDBEACH*?

Statement: “This is a place where you might encounter sand dunes and sparse vegetation.”

A: I choose *DESERT*. ✓