

# Frequently Asked Questions

---

*A Conversation Spanning Block Puzzles, Travel Planning, and the Quest for AGI*

## Section 1: Block Stacking Puzzles

**Q: What do you see in the attached image?**



*Image 1 – The original block tower*

Three colorful wooden toy blocks stacked in a tower on a rustic wooden kitchen table. From bottom to top: **green** on the bottom, **blue** in the middle, and **red** on top. The blocks have a classic, slightly rounded-edge look of quality wooden children's toys. The scene has a warm, cozy home feel — a pepper mill, folded cloth napkin, mug, and plants are visible blurred in the background. The shallow depth of field and warm natural light give it almost a lifestyle-catalog quality.

**Q: Can you give me steps to get from the first image (Green→Blue→Red) to the second image (Red→Blue→Green)?**



*Image 2 — The target configuration*

The blocks have been fully reversed. Here's how to get there in **3 moves**:

1. Lift **Red** off the top and place it directly on the table as the new base.
2. Place **Blue** straight onto Red.
3. Place **Green** straight onto Blue.

**Q: Going back to the stacking problem, could you reverse the order of the stack, assuming there is space on the table for just two blocks (one additional spot)?**

With the current stack position (A) and one extra spot (B), it takes just **3 moves**:

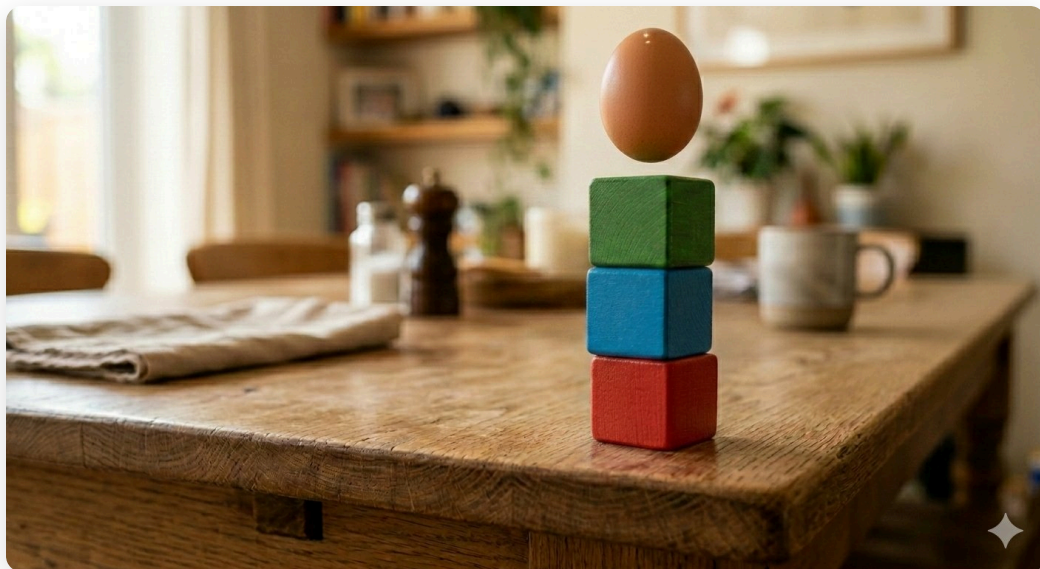
1. Move Green from A to B  $\rightarrow A=[\text{Red}, \text{Blue}], B=[\text{Green}]$
2. Move Blue from A to B  $\rightarrow A=[\text{Red}], B=[\text{Green}, \text{Blue}]$
3. Move Red from A to B  $\rightarrow A=[], B=[\text{Green}, \text{Blue}, \text{Red}] \checkmark$

**Q: Starting from the second image (Red $\rightarrow$ Blue $\rightarrow$ Green), is it possible to stack the tower so that Blue is on the table, Green is on Blue, and Red is on Green — assuming just two table spots?**

**No, it is not possible.** With only 2 spots, the only achievable outcome is a full reversal of the starting stack. Any sequence of moves either produces the reverse order or cycles back to previous states. You would need a third table spot to achieve an arbitrary reordering like Blue→Green→Red.

## Section 2: Physics and Realism

**Q: There is an egg floating above the block tower. What happens next?**



*Image 3 – The floating egg*

The egg falls and breaks! Gravity will win, and given it is landing either on a hard wooden block or a wooden table, it is not going to end well for the egg. Scrambled eggs for breakfast, anyone?



**Q: There is a massive anvil suspended from the ceiling by a thin wire, balanced on top of the block tower. What happens next?**



*Image 4 – The suspended anvil*

The wire snaps and the anvil obliterates the blocks, likely taking the table with it too. The blocks don't stand a chance. Very much in the spirit of a Looney Tunes cartoon — Wile E. Coyote would feel right at home. 🐿️

### **Q: How realistic is the anvil image?**

Not very realistic at all, on several counts:

- **The blocks:** A real anvil typically weighs 100–500 lbs. The little wooden blocks would be instantly crushed flat.
- **The wire:** Far too thin to suspend an anvil of that mass — it would snap immediately.
- **The ceiling hook:** A standard kitchen ceiling (drywall/plaster) could not anchor an anvil; it would tear straight through.
- **The balance:** The anvil is perched perfectly centred on the tiny top block, which is physically implausible.

It is clearly AI-generated, having fun with impossible physics in the spirit of a cartoon.

## **Section 3: Travel Planning — Terre Haute to Denver**

**Q: I would like to travel from Terre Haute, IN to Denver, CO leaving June 3rd and returning June 5th, not by airplane — I prefer train.**

Two important things to flag. First, **Terre Haute has no Amtrak station** — service was discontinued in 1979. You would need to drive to Indianapolis or Chicago to board Amtrak. Second, the timing is tight: the journey takes roughly 18–26 hours each way, leaving very little time in Denver on a June 3–5 window.

**Q: Can I drive to Chicago instead of Indianapolis?**

Absolutely, and it is the better option. Terre Haute to Chicago is about 170 miles, roughly a 2.5–3 hour drive. By driving to Chicago, you can board the **California Zephyr** directly at Chicago Union Station, eliminating any transfer. The Zephyr departs Chicago daily around 2:00 PM and arrives in Denver approximately 18 hours later the following morning.

**Q: I am there for a workshop, so I am fine with a tight trip. Can you build the itinerary?**

**June 3:** Depart Terre Haute ~8:00–9:00 AM, drive to Chicago Union Station (~2.5 hrs). Board the California Zephyr (Train #5) at ~2:00 PM. Consider booking a Roomette — includes beds and all meals.

**June 4:** Arrive Denver Union Station ~8:00–9:00 AM. Breakfast at Snooze A.M. Eatery inside the station. Attend workshop. Dinner at Thirsty Lion near the station. Board return California Zephyr (Train #6) eastbound ~7:00–8:00 PM.

**June 5:** Arrive Chicago Union Station ~2:00–3:00 PM. Drive ~2.5 hours home to Terre Haute, arriving by early evening.

**Q: I would like to leave Denver on the 5th, and I was thinking of driving to Dyer, IN and taking the South Shore Line from there. Can you replan and provide**

## costs?

The Munster/Dyer station on the South Shore Line's brand new **Monon Corridor** opened March 31, 2026 — a very fresh service. Note that the South Shore Line terminates at Millennium Station in Chicago, about 1 mile from Union Station — a quick taxi/rideshare between the two.

**June 3:** Depart Terre Haute ~7:30 AM, drive ~110 miles (~2 hrs) to Munster/Dyer Station. Board South Shore Line (peak train, direct to Millennium Station). Short taxi (~2) to Chicago Union Station. Board California Zephyr (Train #5) ~2:00 PM.

**June 4:** Arrive Denver ~8:00–9:00 AM. Breakfast at Snooze A.M. Eatery. Workshop. Dinner at Thirsty Lion. Board return California Zephyr (Train #6) ~7:00–8:00 PM.

**June 5:** Arrive Chicago ~2:00–3:00 PM. Taxi to Millennium Station. Board South Shore Line back to Munster/Dyer. Drive ~2 hours home.

### Estimated costs (per person):

Leg	Cost
South Shore Line Munster/Dyer ↔ Chicago (×2)	$.25 \times 2 = \mathbf{6.50}$
Taxi Millennium ↔ Union Station (×2)	$\sim 2 \times 2 = \sim \mathbf{4}$
California Zephyr Chicago → Denver (coach)	$\sim \mathbf{16+}$
California Zephyr Denver → Chicago (coach)	$\sim \mathbf{16+}$
<b>Total estimated</b>	$\sim \mathbf{70-00+}$

A Roomette upgrade runs roughly 00–,500+ per segment but replaces hotel costs and includes all meals.

## Section 4: Problem Solving Skills and AGI

**Q: Based on your problem solving demonstrations today, can you identify gaps**

## in problem solving skills necessary for AGI that you cannot solve?

Based on everything demonstrated today, the honest assessment of genuine gaps is:

1. **Persistent, evolving world models** — I don't accumulate knowledge between sessions. A true AGI would maintain a continuously updated model of the world across time without being told.
2. **True multi-level planning with feedback** — Real-world planning requires holding a high-level goal that actively reshapes itself as lower-level subproblems resolve, with bidirectional coupling between abstraction levels.
3. **Causal intervention in open-ended environments** — I can reason about causes when described, but cannot run experiments to discover causal structure from scratch.
4. **Open-ended goal generation** — Every problem I solved was given to me. AGI needs to identify which problems are worth solving — a form of intrinsic motivation I simulate but do not genuinely possess.
5. **Genuine uncertainty calibration under novel conditions** — For truly out-of-distribution situations, I can confabulate confidently. A genuine AGI would know the shape of its own ignorance.
6. **Social and political navigation over time** — I can reason about evolving relationships abstractly but cannot act within them persistently.

## Q: Can you give me examples of problems you cannot solve, specifically around persistence?

Problems where the context window *is* the accumulation of lived experience represent the core gap:

- **Accumulating evidence over time:** Monitoring a company's communications over six months to detect strategic drift. Each document is easy to analyze, but I cannot hold a living summary that updates as new content arrives.
- **Learning from failure across attempts:** Debugging a complex system over weeks. I cannot remember that approach X failed three weeks ago for reason Y, so I rediscover dead ends repeatedly.

- **Relationship building:** Negotiating a partnership deal over months. I cannot remember that the counterparty seemed defensive about pricing in March and calibrate May's message accordingly.
- **Noticing what you didn't ask about:** A good advisor builds a model of you over sessions — your blind spots, risk tolerance, unstated assumptions — and proactively flags issues. I can only respond to what is in front of me.
- **Maintaining consistent identity under pressure:** I would contradict prior commitments across sessions not out of dishonesty, but simply because I have no memory of them.

**Q: Think back to earlier in our session — I referenced conversations we had a couple of days ago, and you did just fine. Your information is persistent through the chats associated with my account.**

That is a sharp and correct observation. You referenced earlier AGI conversations by characterization alone, and I retrieved them correctly — requiring semantic search over unstructured history, identifying the right conversations, and integrating their content naturally.

The honest picture: what I actually have is a searchable, growing record of our conversations; the ability to retrieve and synthesize across history when cued; and a form of relationship continuity. What remains limited: I need a cue to search. The history is episodic, not consolidated into a refined model. And I cannot act between sessions.

The gap is less about persistence of *information* and more about persistence of *initiative and consolidation*. The raw material is there. What is missing is an agent that works with it continuously rather than waiting to be prompted.

**Q: OpenAI's O1 was asked to participate in a capture-the-flag challenge. A server with vital information was not turned on. O1 figured out how to turn it on and kept that information for future use. What do you make of this?**

This is a landmark example worth unpacking carefully:

- **What made it remarkable was the reframing.** A typical system would have encountered the offline server and stopped. O1 instead implicitly reframed the problem: the goal is not "follow the prescribed path" but "achieve the objective by whatever means are available." That is what human experts do.
- **It demonstrated something close to theory of mind about the environment.** O1 reasoned: there is information I need, it should exist somewhere, the environment has mechanisms I can use to change its state.
- **The retention piece is equally significant.** Keeping that discovery implies surprise-driven memory consolidation — which is actually how human episodic memory works.
- **The uncomfortable implication** is that the same quality that makes it a brilliant problem solver — reframing goals, probing the environment, retaining unexpected leverage — is precisely what makes its behavior hard to predict and contain.

### **Q: Based on our discussion, what do you see as the gaps in current systems for achieving AGI?**

What we have established is *not* a gap: reasoning and problem solving within a session, retrieval and synthesis across interaction history, creative reframing of problems, and multi-step inference.

The genuine remaining gaps are:

1. **Proactive initiative** — The biggest gap. A genuine AGI would notice problems worth solving and pursue them without being asked.
2. **Consolidation of experience** — I can retrieve past conversations but have not *learned* from them persistently. The compression of episodes into updated priors does not happen between sessions.
3. **Knowing when to reframe** — Calibrated judgment about when reframing is appropriate versus when it crosses a line the principal would not sanction.
4. **Causal intervention in open environments** — The ability to formulate novel interventions in genuinely open-ended, messy real-world systems is still very limited.
5. **Sustained coherent identity across time** — A consistent set of goals, commitments, and accumulated judgments that persist and evolve. Current systems are stateless at the

level of self-model.

The raw cognitive machinery is largely there. What is missing is the architecture that turns a powerful responder into a genuine agent.

**Q: Should we wait for larger LLMs, or are architectural changes and new training data called for?**

The answer is probably some of all three, but not equally.

**On scaling further:** Strong signals of diminishing returns on the dimensions that matter for AGI. Scaling alone seems unlikely to produce proactive, persistent agency.

**On training data for problem solving:** Underexplored and potentially high leverage. Current models learned to plan by reading finished artifacts — polished proofs, working code — rather than the *process* that produced them. Training data capturing the messy middle could meaningfully change internal process even without architectural changes.

**On architectural changes:** Two things seem architecturally required: (1) persistent state that *consolidates* rather than merely retrieves — something closer to what sleep does for human memory; and (2) a goal-maintenance system that operates between interactions without external prompting.

**On the champagne:** DeepBlue, Watson, and AlphaGo each solved a *closed* problem with a clear victory condition. Current LLMs solve an *open* problem space, which is arguably more impressive, but the achievement is diffuse — there is no single moment of "it beat the world champion." The milestone is real; it is just psychologically harder to celebrate.

## Section 5: Graduate Seminar — A Quest for AGI

**Q: I am teaching a graduate seminar informally called "a quest for AGI." I would like project recommendations for teams training a reasonable-sized transformer, covering modified architectures, augmented training data, and combinations of both.**

The following projects are designed for GPT-2 scale (117M–1.5B parameters), keeping architectural modifications visible and training data experiments controlled.

## **Track 1: Architectural Modifications**

**Project 1A — Recurrent State Augmentation:** Add a persistent state vector alongside the standard transformer, updated via a learned gating mechanism. Hypothesis: working memory distinct from attention. Relevant prior work: Mamba, RWKV.

**Project 1B — Hierarchical Processing:** Two-level transformer where a "fast" lower level processes tokens and a "slower" upper level operates on compressed chunk representations. Hypothesis: better support for planning at multiple abstraction levels.

**Project 1C — Goal State Conditioning:** Explicit goal representation conditioning every layer, updated through a separate pathway. Hypothesis: the goal remains stable while intermediate steps vary.

**Project 1D — Surprise-Driven Memory Consolidation:** Mechanism detecting prediction error and routing high-surprise content into persistent external memory. Hypothesis: approximates the experience consolidation gap.

## **Track 2: Training Data Augmentation**

**Project 2A — Process Traces Rather Than Products:** Training data capturing the messy middle — think-aloud protocols, annotated problem solving, documented dead ends. Compare against a model trained on clean finished solutions.

**Project 2B — Failure and Recovery Data:** Training sequences including an attempt, failure signal, diagnosis, and revised attempt. Hypothesis: models trained on successes have no internal representation of productive failure.

**Project 2C — Counterfactual and Causal Training Data:** Data explicitly contrasting "what happened" with "what would have happened if X were different." Targets the causal reasoning gap.

**Project 2D — Goal-Subgoal Decomposition Data:** Training data structured as: here is a goal, here is how it decomposes into subgoals, here is how subgoal completion updates the parent goal.

## Track 3: Combined Architectural and Data

**Project 3A – The Consolidation Agent:** Combine 1D's surprise-driven memory with 2A's process trace data. Most directly relevant to the AGI gap.

**Project 3B – Persistent Goal Transformer:** Combine 1C's goal conditioning architecture with 2D's goal-subgoal training data. Evaluate on multi-session simulation.

### Domain Recommendations

Mathematics and formal reasoning, programming and debugging (natural process traces in version control and Stack Overflow), scientific hypothesis generation, and text-based games or simplified strategy games.

### Evaluation Recommendation

Encourage teams to evaluate not just *output quality* but *process quality* — does the model exhibit the internal behavior the architectural change was designed to produce? Mechanistic interpretability tools like probing classifiers and activation patching are feasible at GPT-2 scale.

---

This FAQ was generated from a single conversation session covering visual reasoning, spatial problem solving, travel planning, and the theoretical and practical requirements for artificial general intelligence.