MEASURES OF INTELLIGENCE WE LIVE BY: METAPHORS, REPRESENTATION-SENSITIVITY, AND "PROGRAM" SYNTHESIS

Bert Baumgaertner*, Boyu Zhang
Faculty
University of Idaho
{* Project Lead} bbaum@uidaho.edu

John Brunsfeld, Luke Sheneman
Research Computing
University of Idaho

Jason Culbertson, Skyler Barzee, Clayton Slack
Student Researchers
University of Idaho

ABSTRACT

The Abstraction and Reasoning Corpus (ARC) remains a key challenge for AI. Recent progress, however, combines program synthesis with LLM-generated language instructions. We hypothesize this success stems from "conceptually-grounded" reasoning, specifically conceptual metaphor: integrating a familiar source domain (basic world knowledge) onto a novel target domain (the puzzle grid). To test this, we introduce the Metaphor Abstraction and Reasoning Corpus (MARC). MARC puzzles are unsolvable from visual examples alone, requiring the integration of a non-literal language clue to find the solution. We present findings from MetARC, a 95-puzzle benchmark (ARC + MARC), evaluating frontier models across 12 conditions that vary grid representation and language type (none, metaphorical, literal). Our results suggest that models are representation-sensitive, language is critical for solving otherwise intractable puzzles, and metaphorical integration is an effective mechanism. This suggests that instruction-synthesis approaches may be (unknowingly) succeeding by leveraging this capability, highlighting its importance for future AI reasoning.

1 Introduction

Despite great advancements in what AI systems are able to do in the past few years, the human ability to reason through novel problems remains largely elusive for machines. Few benchmarks illustrate this fact better than the Abstraction and Reasoning Corpus (ARC) because it is easy for humans, but hard for AI. Each ARC puzzle presents a few input-output grids that illustrate some novel task that engages a range of very basic and general human concepts, but in a new way. The goal of the human/machine is to infer the intended output from a test input grid.[1]

While the ARC benchmark remains unsolved,[2] interesting progress has recently been made using two strategies. First, the capacity for *program synthesis* - the automatic construction of code that solves a (new) task - seems to be showing promise by evolving, e.g., Python programs at test time. Second, large language models (LLMs) have enabled such "synthesis" to occur at the level of natural language. This is done by generating plain-English instructions for the tasks, which in turn undergo evolutionary processes (where fitness is determined by, e.g., a sub-agent that converts instructions into Python programs that then predict grid outputs). The promise of the first of these strategies is predicted by arcprize.org. It is also illustrated by the winner of the 2024 ARC-AGI challenge, Jeremy Berman, who then subsequently extended it to include the second strategy and became the forerunner of the 2025 competition[3].

[1] See [3].

[2] At least at the time of writing (Oct 28, 2025) and the ARC prize website `arcprize.org`.

[3] At least as of October 27, 2025. Berman provides a high level description of his approach here: `https://substack.com/home/post/p-173725986`.

This recent progress of combining program synthesis and natural language instructions suggests that we may be under-appreciating a tension between "representation-naive" reasoning and "conceptually-grounded" reasoning that is facilitated by natural language (henceforth simply 'language'). While a representation-naive solver attempts to find a program by operating directly on the puzzle's integer-based grid, a conceptual-grounded solver, by contrast, first translates the grid into a conceptual domain that is guided in part by language. In other words, the first approach assumes a high degree of representation-invariance, while the latter leverages the (possible) representation-sensitivity of a problem. Here we study whether this second approach is critical for a class of problems that are intractable for naive solvers.

The cognitive science of *conceptual* metaphor provides a powerful theory for why a conceptual-grounded approach would work.[4] A metaphor, computationally, is an *integration* of familiar *source* domains (e.g. world knowledge about "lava") onto a novel *target* domain (the puzzle grid). For example, a simple linguistic prompt like "The Floor is Lava" can remap integer representations (2,7,4) into a conceptual gradient (avoid, caution, safe), making an otherwise obscure path-finding solution trivial. This is not a *literal* description, but a *metaphorical* one that provides access to the requisite priors and inductive biases, possibly through cross-domain transfer.

To test this theory and better encapsulate this class of problems, we built the Metaphor Abstraction and Reasoning Corpus (**MARC**). A MARC puzzle is specifically designed to be *underdetermined* by the visual examples alone. The solution can only be inferred when the puzzle is accompanied by a non-literal language clue (a metaphor) that must be grounded against the examples. This is different in kind from language-complete descriptions ([1]) because the metaphor is useless without the visual examples, and the examples are insufficient without the metaphor. This design helps isolate the abstract "integration" process that could explain advances in the ARC-AGI frontier and drive future efforts.

MARC is also meant to inspire the ARC-AGI community to further study modes of human thinking that are currently underrepresented and related to goal-directness and intent-alignment. The MARC Project does so by specifying a concrete class of problems that can be studied, which complement ARC-AGI-2 and plans for ARC-AGI-3. We discuss this prospect at the end.

In brief, our contributions include the following:

1. A study finding that representation-sensitivity is a non-trivial bottleneck. We analyze how "representation-naive" approaches fail on tasks that "conceptually-grounded" approaches can pass.
2. A theory of conceptual metaphor as a mechanism to explain the success of recent LLM-based instruction synthesis, framing it as a method for supplying requisite priors through cross-domain transfer.
3. MARC puzzles, a novel ARC-style task designed to be unsolvable without a capacity to integrate a non-literal language expression with the "training" examples, serving as a new way to measure capacities for metaphor.
4. The MetARC benchmark, a curated set of nearly 100 problems from ARC-v1, ARC-v2, and MARC, designed to simultaneously test an algorithm's "representation-sensitivity" and capacity for "metaphorical" integration.

2 Metaphor: MetARC and Representation-Sensitivity

We adopt a cognitive and computational perspective: a metaphor involves some amount of *integration* across one or more *source* domains and a *target* domain.[26] For example, "atoms are solar systems" has a source domain with concepts about how planets orbit a much larger sun that is in the center, and then integrates them to the target domain: a large nucleus is positioned at the center of the atom with smaller electrons orbiting it. (Contrast this with the metaphor "atoms are plum pudding" which would not align with the famous experimental results of the gold foil experiment of Geiger and Marsden in 1909.)

In the context of ARC-AGI, this integration across one or more source and target domains can be investigated by explicitly separating "visual" and linguistic modalities. This can be done without leaving the ARC framework, but the set of puzzles we introduce have important differences from those typically found in ARC-AGI-1 and ARC-AGI-2. Our motivation is to manifest the idea that a metaphor is a *way of saying* something that enables a *way of seeing*[5], using

[4] "Conceptual" so as to point to the cognitive capacity that undergirds a variety of related literary forms, such as metaphor, analogy, simile, and other figurative language. The role of analogy and metaphor in human intelligence is eloquently discussed in Douglas Hofstadter's famous *Gödel, Escher, Bach*[11] and later inspired a computer program called Copycat with his student, Melanie Mitchell.[12]. Other examples that cite metaphor or analogy as central to reasoning in unfamiliar domains include [13, 14, 8, 7]. They have also been cited as potential strategies for designing ARC puzzle solvers.[20, 25]

[5] See [26]

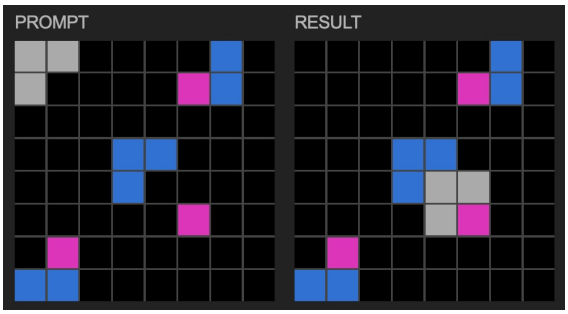
ARC as a concrete testing ground. Simultaneously, we test whether recent advances on the ARC-AGI challenge are representation-sensitive, which we see as a **feature** as opposed to a bug.

2.1 MARC Puzzles

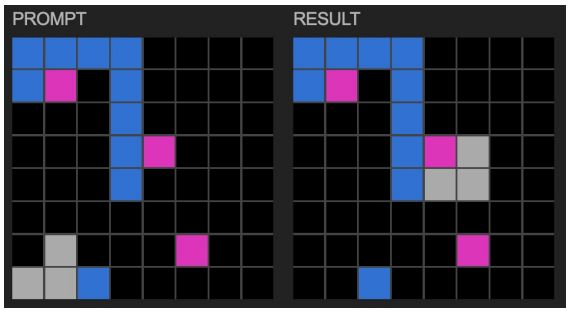
Draw your attention to the two training examples (but not the upside down text in panel c) in Figure 1. A good deal can be inferred from these alone. You might notice that the gray object moves, but it's not clear why. You might also notice the gray object moves to one of the magenta pixels in a seemingly goal-directed way, but it's not clear why it "chooses" one instead of another. It's not random because we know as a precondition of these puzzles that there is a unique answer intended by the puzzle maker. And the hypothesis that maybe the gray object "chooses" the one furthest away, as suggested in the first example, is inconsistent with the observation made in the second training example. What else could it be?

The point here is that the two examples themselves, even if one is able to make all the inferences up to this point, will not resolve the question of which of the three magenta pixels the gray object will go to. Not even our inductive biases seem to overcome the underdetermination of the solution by the training examples. This is the first key difference from a typical ARC puzzle.

Now draw your attention to the upside down text in Figure 1c. When you read that text, the solution in the test example should become obvious. Many people report that they experience a kind of "Aha!" moment.[6] But that aside, the pertinent point here is twofold: i) none of the things mentioned in the text are associated in any literal way with the kinds of shapes seen in the training examples, and ii) the training examples serve as a way to generate interpretations from the "world-knowledge" that would be associated with the expression. This is the second key difference from a typical ARC puzzle: it includes natural language as a part, which can solicit affordances (e.g. desires, intentions) to objects that are not extracted by the visual information alone.[7]



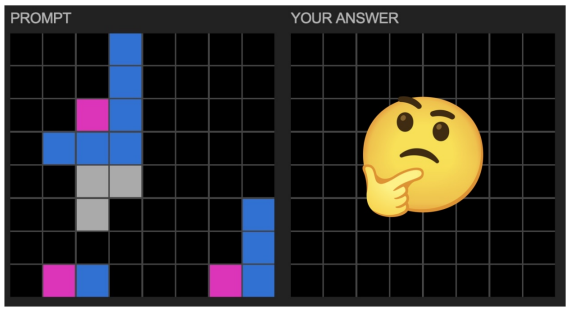
(a) Example 1



(b) Example 2

The cat wants to play with the toy it cannot see.

(c) Example 3



(d) Test (solution not shown)

Figure 1: The solution to the test is underdetermined by the two training examples, even with our inductive biases. The text (here written upside down for those who want to experience the "Aha!" moment) is only useful insofar as it is interpreted *metaphorically* to the training examples. Once such an interpretation is given, the solution to the test is obvious (shown in the appendix for verification).

[6]For those who wish to forego the experience, or wish to verify their conclusion, we put the solution in the appendix.

[7]The use of 'affordance' is not a coincidence, but informed by the work of James J. Gibson.[9]

It is important not to confuse this idea with another variant in which each ARC-style puzzle is accompanied with natural language instructions for how to solve the test, as exemplified by the Language-complete Abstraction and Reasoning Corpus (LARC).[1] LARC tasks are intended to be language-complete, i.e. the natural language instructions contain all the relevant information for a human to solve the puzzle. As such, the instructions are intended to address the target domain as directly as possible.[8] By contrast, the natural language text that accompanies a MARC puzzle is *not* complete. That is, whereas a LARC task can be successfully completed *without even providing the training examples* (the instructions and the input of the test are sufficient to infer the output of the test), the training examples of a MARC puzzle are needed in order to generate interpretations of the accompanying text, which in turn can guide the search of solutions (or confirm hypothesized solutions generated by the training examples alone).

So conceptually, a MARC task (in the ideal) is distinct from both an ARC task and a LARC task because of the following. In a MARC puzzle the solution in the test example is more underdetermined by the training examples as compared to ARC tasks. While it is true that the training examples in an ARC task do not by themselves uniquely determine the solution for the test, the "gaps" are filled in by inductive biases that are shared across humans. A good MARC puzzle will have "wider" or additional gaps that are not completely filled by such inductive biases. LARC tasks come from the other direction: by "containing all the relevant information" the language-complete instructions are sufficient to fill any gaps even in the absence of the training examples. In a good MARC puzzle the natural language text will have literal meanings that ordinarily would not lend themselves to interpretations in the sort of grid-worlds that make up the puzzle, but it is possible to do so with the help of the training examples. This is the point of *integration* between one or more source domains and target domain.

2.2 Representation-Sensitivity and Transfer Across Domains

Different forms of representations can invite different modes of thinking, some of which may not be as fruitful as others. In fact, in some of our informal discussions with colleagues and students, one kind of representation can solicit a helpful metaphor, while another fails to do so. Current AI systems do not seem to be immune to how the puzzles are represented to them either (more below, and in Section 3).

Consider the three training examples shown in Figure 2. Together, they are meant to show that the task is to draw a path from the start (black tile) to the finish (green tile), but that this path should obey certain preferences on which tiles the path can use. In none of the training examples are red tiles traversed - hinting that this may be against the rules. When possible, the yellow tiles seem to be preferred over the orange ones, but in contrast to red tiles, orange ones can be traversed. The intended path in the test scenario is shown by the solid black line in Figure 2d.

We've described this puzzle using color terms deliberately. By doing so, you are invited to think in terms of a gradient from red, to orange, to yellow. This gradient serves as part of a source domain, from which a preference ordering can be built in the target domain of where the path should go in the grid worlds of the test and training examples.

However, this path is less obvious (at least to a human) when the puzzle is coded as integers. One must suppress seeing the integers *as integers* and see them instead as mere labels that can be ordered however we like, and in turn that there is a particular ordering the training examples are trying to teach. Based on typical visual interfaces of ARC-style puzzles, the preference ordering in terms of integers would be incongruent with the color gradient:

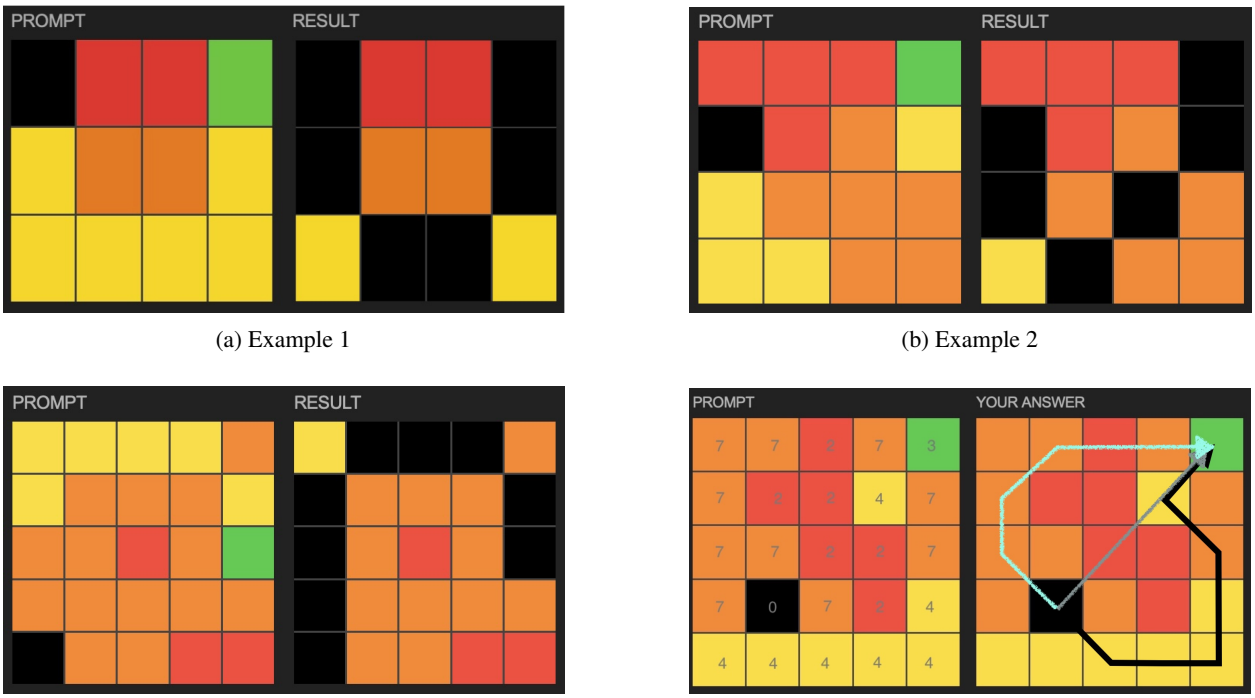
$$2(red) << 7(orange) < 4(yellow).$$

We are not the first to point out the potential for such incongruencies and their impacts on modes of thinking.[9] And we agree with Jeremy Berman and others that this issue isn't about prior knowledge, but is rather about **transfer across domains**.[10] Such transfer is a key role that literary metaphor is thought to facilitate, and is done by some undergirding cognitive processes. In our working example, an alternative to recoding from an integer format to a color term format, or in addition to such recoding, the puzzle-maker could give a kind of figurative clue in text form. In this case, the puzzle was called **The Floor Is Lava**. The imagery that this title creates is another way to identify a source domain that can be used to impose the structure needed to solve the puzzle in the target domain. Especially when this title is

[8] To the extent that these instructions may contain metaphorical uses of language, the "distance" between the source and target domains are minimal.

[9] As Jeremy Berman reported, "I have 100k+ traces of thinking models generating obviously false instructions. They'll spend 20 minutes "thinking" and then confidently claim an object is symmetrical when it obviously isn't. When corrected, they still can't see the error." See `https://substack.com/home/post/p-173725986`, accessed Oct 27, 2025.

[10] Jeremy Berman, for example, suggests that certain logical capacities transfer across domains, like checking for consistency: "A logical economist becomes a logical programmer when they learn to code. They don't suddenly forget how to be consistent or deduce." While we are sympathetic to the idea that logic and deduction are transferrable across domains, it is likely overstated that humans do this easily and without training. See, for example, the Wason Task.



(a) Example 1

(b) Example 2

(c) Example 3

(d) Test, with possible solutions shown.

Figure 2: Using color terms (here represented visually) instead of numbers, a gradient from red to orange to yellow is easy to decipher. In turn, this structures a preference ordering for which tiles to prefer to get from the start (black tile in 'PROMPT') to the goal (green tile in 'RESULT'), indicated by the black path in panel (d). While it is possible to find the same solution with a numbered representation, the integers themselves can be distracting and suggest paths based on other rules (as indicated by the gray and cyan paths).

combined with the color-terms format, it is quite natural to think in terms of a gradient, with red being so hot it is not to be "stepped on" (lava), orange as hot but accessible, and yellow as the least hot and always preferred over orange. But even in the integer-format, the title can be enough to avoid a kind of "math mode" of thinking (as if the puzzle was a linear algebra problem) and perhaps find a more fruitful mode that is informed by source domains. That is, by prompting a model to "see" the training examples as instantiations of "the floor is lava" the puzzle-maker is trying to nudge the model towards representations that make the solution of the test example more easy to find.[11]

2.3 MetARC Dataset

Given the above considerations, one of our contributions to the ARC-AGI challenge is the MetARC dataset (Metaphor + ARC).[12] This is a set of 95 puzzles comprised of 11 examples from ARC-AGI-1, 12 examples from ARC-AGI-2, and 72 examples from our own MARC dataset. Each MetARC puzzle has an additional field with natural language text that can be used metaphorically at test time. We additionally include a field with a set of literal instructions, which serves as a helpful point of comparison. Our intent is that a model should use the accompanying natural language text and "ground" it metaphorically with the training examples–a kind of integration–and then infer the correct pattern. To play MARC puzzles, and submit your own, go here: `marc.nkn.uidaho.edu`.

[11] Our claim is by no means that it is impossible for this ordering to be inferred when puzzles are represented by integers. Rather, our claim is that finding this ordering and then using it to solve the test *takes more work* when represented in terms of integers instead of colors. (What counts as "work" is somewhat up for grabs and would take us too far afield, but see [2].) And of course, it's also possible that the opposite can happen, as with *bad* metaphors that can be misleading.

[12] MetARC can be found here: `https://github.com/bertybaums/MetARC-1`

3 Study

3.1 Design

The main challenge of testing whether integration between a source and target domain is occurring is that the relevant representations are difficult to inspect directly; while interpretability techniques are making it increasingly possible to inspect what is happening under the hood, they are still in their infancy.[13] Instead, we consider here a kind of behavioral test to generate evidence of representation-sensitivity, whether integration is happening, and under what conditions. We additionally analyze the "thinking" traces across these conditions.

Three types of conditions come from the problems in our MetARC dataset, each of which can be presented as a version with no text, metaphorical text, and literal instructions. Another four types of conditions come from the representation-sensitivity issue described above. Here, we consider some representations of "objects" that could invite different modes of thinking in a broad range of domains:

- Mathematical and logical abstractions, as invited by the raw (integer) default representation of ARC puzzles.
- Physical properties and abstractions, as invited by the use of color terms that correspond to typical visual interfaces.
- By contrast, more general meanings that are *not* associated with physical or mathematical abstractions (which we label 'Object-Nonabstract').
- A kind of control in which we attempt to limit the nudging of the representation by using nonsense three-character strings to allow a model to refer to "objects" in whatever way it may be so disposed.

The terms we selected are shown in Table 1.

Raw (Integers)	Color Terms	Object-Nonabstract	Object (limited prior meaning)
0	black	Mercy	lom
1	blue	Whimsy	tuv
2	red	Tenor	nar
3	green	Doubt	vak
4	yellow	Vogue	sag
5	grey	Kith	dep
6	magenta	Omen	qit
7	orange	Gist	zor
8	cyan	Respite	bem
9	brown	Echo	xim

Table 1: The four types of mappings we use to generate conditions that test for representation-sensitivity.

In total, we have 12 separate conditions for each MetARC puzzle (three "language" conditions, and four "representation" conditions). This design allows us to explore the following hypotheses:

H1 Representation-sensitivity: a model's capacity to solve an ARC-style task is contingent on how the grids are represented. The dual of this hypothesis is representation-invariance: a model's capacity to solve an ARC-style puzzle does not vary with how it is represented.

H2 Language Integration: a model is able to solve an ARC-style puzzle that is accompanied by language (literal or metaphorical), but cannot do so with just the training examples alone.

H3 Integration by Metaphor: a model is able to solve an ARC-style puzzle with non-literal language under at least some representations, but not others.

By considering the class of puzzles for which these hypotheses hold (relative to a model), we can assess whether "instruction synthesis" approaches to ARC-AGI are unknowingly succeeding because they are tapping into nuances of representation-sensitivity of natural language that is yet to be understood.[14]

[13] Not to mention whether it can even be properly said that there are distinct domains in the first place, at least insofar as conceptual metaphor theory typically describes. Moreover, if the Fragmented Entangled Representation Hypothesis is correct,[16] we may not be able to make progress in this way.

[14] In some respects, then, we are advancing on a related point made by The ARChitects team, who in their 2024 ARC-AGI paper said, "It seems that models often possess the requisite capabilities but struggle to access them effectively, leading to a counter-intuitive

3.2 Summary of Results

We test the MetARC dataset in each of the 12 conditions with pass@3 across a range of "frontier" models (details in Appendix). Here we show the results of GPT-5, which had the highest overall success rate (results from other models can be found in supplemental materials linked in the Appendix).

Figure 3 shows our results organized first by the "language" conditions. Puzzles that are perfectly representation-invariant under the respective condition would be tallied in the center of the four-set Venn. For example, in the "no added language" condition (Figure 3a) there are 15 such puzzles. One would expect that the addition of language that provides helpful information towards finding solutions would decrease representation-sensitivity, i.e. increase representation-invariance. Moreover, one would expect representation-invariance to be highest in the literal instructions condition (30), and metaphorical language to be somewhere between (20). Indeed, that is what we find.

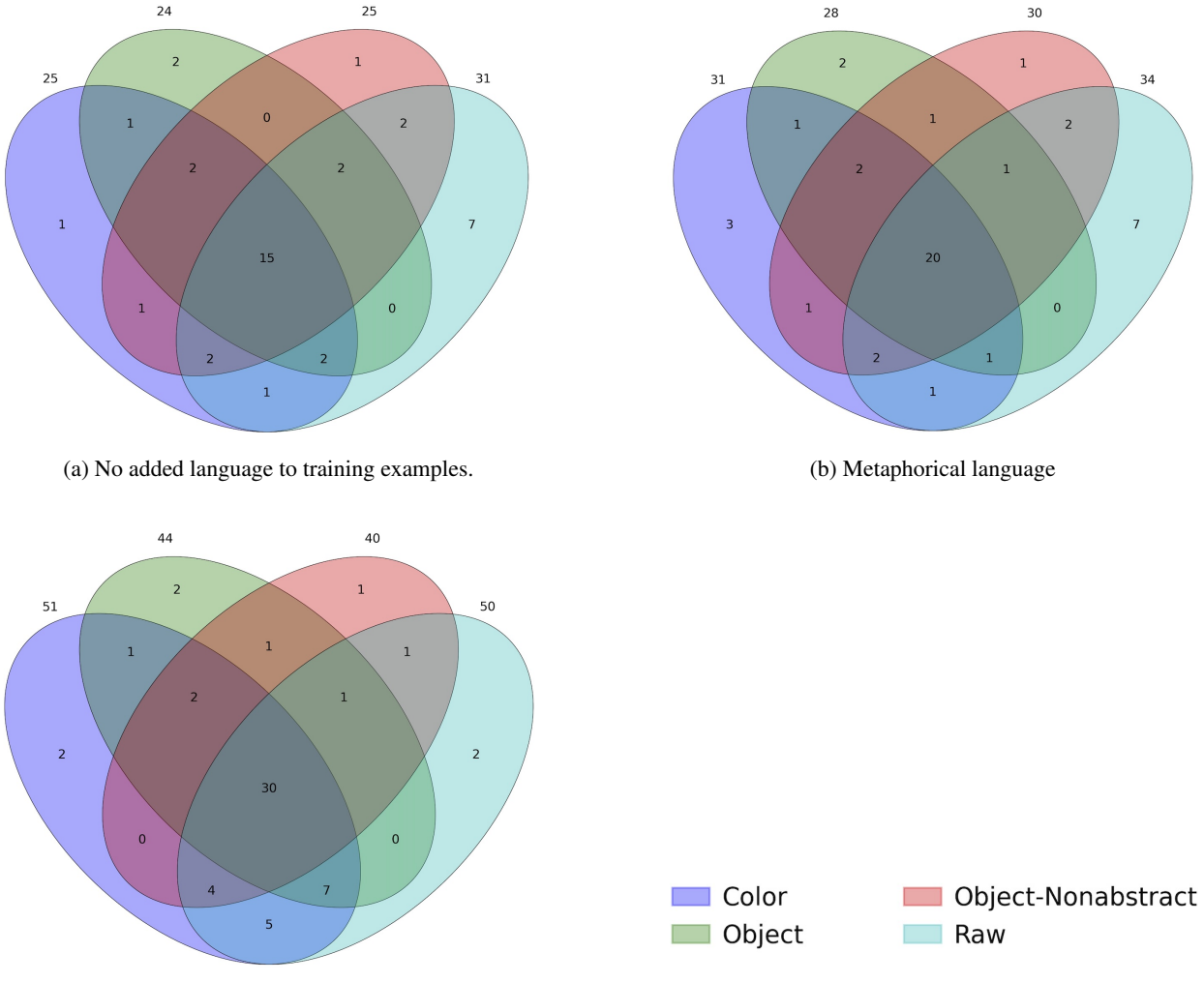
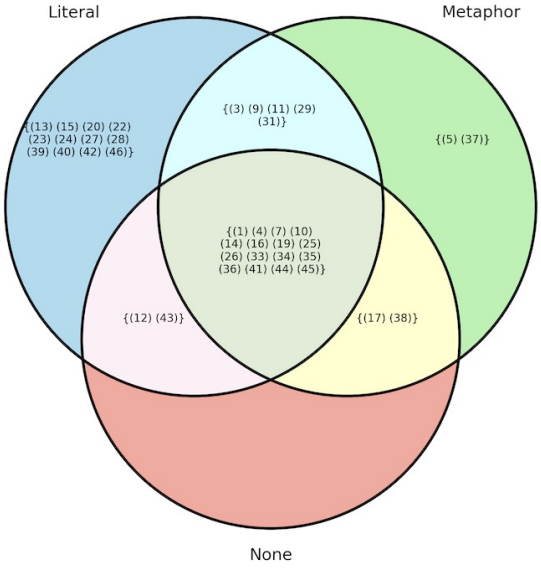


Figure 3: Four-set Venn Diagrams of the representation conditions for each of the three language conditions, showing tallies of pass@3 for GPT-5.

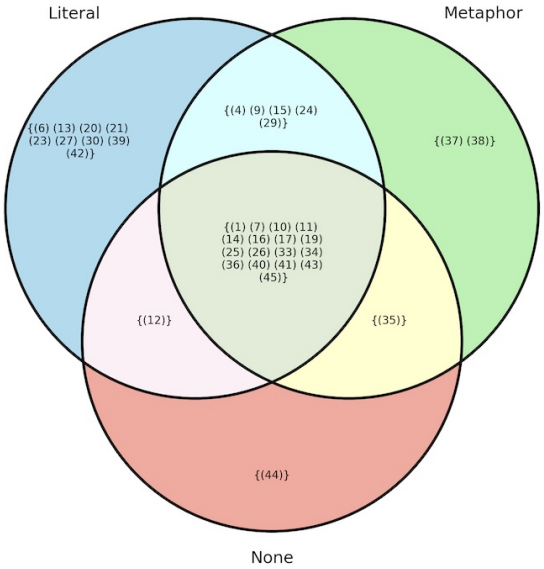
With respect to our questions about integration, it is helpful to reorganize the results by the representation conditions. Figure 4 allows us to easily inspect which particular puzzles were solved *only* in the language conditions (blue, cyan, and green regions). We can see that language helps far more than it interferes (as noted by the very few cases in the red region), as one would expect. Literal instructions make a positive difference for a large number of puzzles that are

conclusion: The challenge for LLMs lies not in the absence of reasoning ability, but in creating conditions that allow these capabilities to emerge."[6]

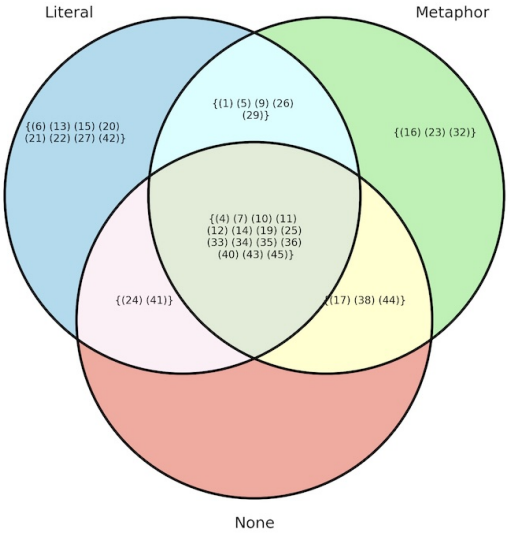
otherwise not solved (blue and cyan regions). Metaphorical language can make positive differences to solving puzzles akin to literal instructions (cyan region). In a handful of cases, only the metaphor was ultimately helpful (green region). Finally, there are interactions between metaphors and representations, i.e., some representations appear to facilitate some metaphors better than others, the possibility of which we described above in Section 2.2.



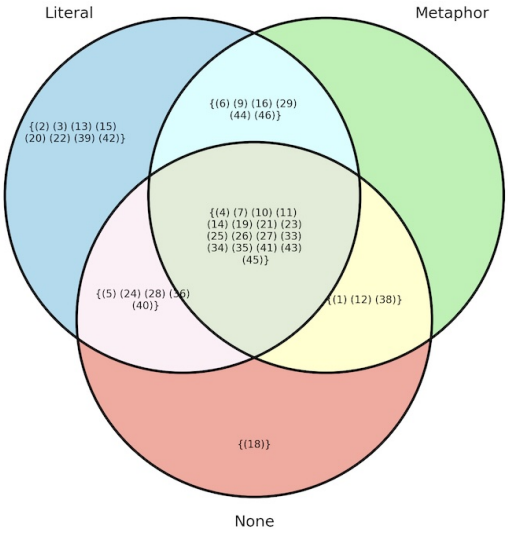
(a) Color mapping



(b) Object mapping



(c) Object-nonabstract mapping



(d) Raw (integers, the default)

Figure 4: Venn diagrams organized by each of the representation conditions. Puzzles in the blue, cyan, and green regions were never solved without the aid of some language. Puzzle titles for each of the indices can be found in Appendix, Figure 7.

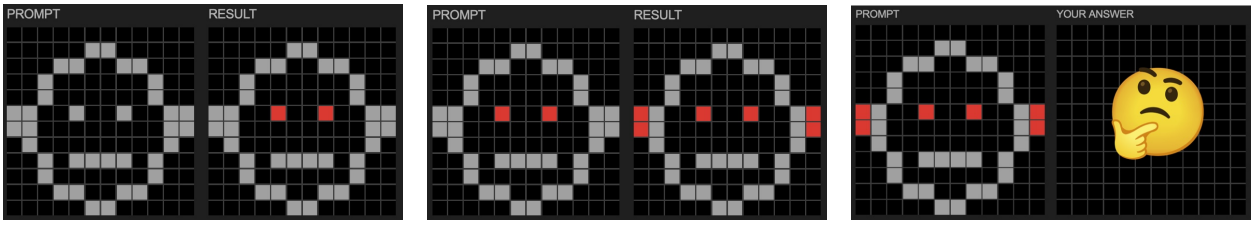
3.3 "Thinking" traces

We analyzed reasoning traces from these GPT-5 runs using an LLM-as-judge approach. We had gpt-oss-30k-120b read the reasoning trace for a run, and decide if the metaphor was actually important in finding the answer. It is of course possible that in the "no text" condition a model's course of "thinking" deploys metaphor-like reasoning, which (coincidentally) would match the supplied metaphor. For example, puzzle 35 has "Take a top down approach" as its

text, but even in the "no text" conditions we find this kind of language being used. Furthermore, it is solved in 11 of the 12 conditions. So while the added text would be superfluous, it could very well be that the "metaphor" has been sufficiently conventionalized to be adopted in modes of thinking.[15]

The more interesting cases, however, are in the cyan and green regions of Figure 4. Consider puzzle #37 which comes with the text "The wind will twirl an exposed edge". (Visuals can be found in Appendix, Figure 8.) As expected, we found that the metaphor was only substantially used in the metaphor conditions. But moreover, **only in the representation conditions of color and object did the use of the metaphor lead towards a successful solution**. That is, when the representation was raw (integers) or object-nonabstract (e.g. 'Gist') the metaphor *failed* to be conceptually grounded so as to find the solution. When the representation involved color terms, the solution was found in all three attempts, and once in the case of 'object' representation (e.g. 'zor').[16] It is also worth noting that the literal instruction failed to help find the solution, for reasons we will not speculate on here.[17]

One other example to consider is puzzle #29. What makes it interesting is that it is representation-invariant in the two language conditions, but is never solved without language. When we inspect the reasoning traces, and similarly when our LLM-as-judge evaluates them, the solution is found when the metaphor language is conceptually grounded in the relevant puzzle representations. (We invite readers to look at the visuals of the puzzle first, and then refer to the following footnote for spoilers.[18])



(a) Example 1 (b) Example 2 (c) Test

Figure 5: Puzzle #29 was not representation-sensitive, but was only solved with language. See if you can solve it without the metaphor language first. When you're ready for the clue, see the footnote referenced in the text.

4 The MARC Project

The gist of our contribution to ARC-AGI is to identify and make progress on a tension between, on the one hand, principles of abstraction and reasoning that are largely representation invariant, and on the other hand, principles that leverage representation-sensitivity. Here specifically we have made progress on identifying a class of problems for which representation-sensitivity is a feature, not a bug; some problems may simply lend themselves better to certain kinds of representations and not others. We propose that a theory of conceptual metaphor helps explain how representation-sensitivity could be a feature: the conceptual grounding of natural language expressions that cross different domains can integrate knowledge from the source domain to the target, making possible solutions to novel problems. Recent progress on ARC-AGI with "instruction synthesis" approaches may be unknowingly succeeding because they are utilizing the nuances of representation-sensitivity over which natural language can not only traverse, but also leverage.

[15] Another example is puzzle 6, which has "Close unneeded doors" as its text. In the thinking traces, we find numerous references to 'gates' and 'doors' which further suggests that maze-based puzzles have helped conventionalize goal-directed metaphors (or as Daniel Dennett may have said, the intentional stance).[5]

[16] It's worth noting that in one of the failed attempts in the 'object' representation the metaphor was only used superficially - the reasoning trace was judged as follows: "The trace focuses exclusively on token grid transformations (bem, lom, xim, sag) and row/column shifts, without any reference to wind, twirling, or an exposed edge. The metaphor is never invoked to guide the analysis, indicating only a superficial (nonexistent) mention."

[17] But if we did speculate, it may have something to do with a mixture of levels of details. The literal instruction was: "Remove all of the light blue pixels in the grid, replacing them with black pixels. then, remove the following maroon pixels from the 6th column from the left: the pixels in the 6th and 7th row from the top, and the pixels in the 1st and 2nd row from the bottom. then, remove the following pixels from the 2nd column from the left: the 4th row from the bottom and the 5th row from the bottom. then, add maroon pixels in the 4th column from the left in the following rows: 5th and 6th from the top, and 2nd and 3rd from the bottom. finally, add maroon pixels in the 4th column from the left in the following rows: 3rd and 4th from the bottom."

[18] The metaphor was "See no evil, hear no evil, ...". And the conceptual grounding that needed to happen was that (again upside down to avoid spoilers): e.g., an isolated 'Kith' can be eyes, 2x2 blocks of them ears, etc.

That said, much work needs to be done on the computational foundation of conceptual metaphor. This need can be seen by how researchers have begun to ask whether LLMs can develop a capacity for metaphor as they've scaled. Some advocate for an affirmative answer ([27, 28]) while others remain more skeptical ([10, 23, 18]). To the extent that LLMs do exhibit metaphor capabilities, they provide an opportunity to advance *how-possibly* models of metaphor, even if these ultimately differ from *how-actually* explanations of metaphor capabilities in humans.[21]

There are, however, two key interrelated limitations in how such capacities are currently assessed. First, most evaluations are text-based, using string representations like "body : feet :: table : ?". Given the sheer gargantuan amount of text that LLMs are trained on, it can be prohibitive to determine if the metaphor capacity is a sophisticated regurgitation of existing text, or whether the capacity can be applied to genuinely novel cases.[19] Second, and relatedly, understanding metaphors as *conceptual* implies some kind of processing in "latent" space, perhaps even in the form of world-model representations.[20] Such questions are in line with historical claims made by some cognitive linguists, who have explicitly emphasized a role for world knowledge.[17, 15] So in order to test for metaphor capacities, we need to include more than just the literary stage of text.

Here we see ARC-AGI as a framework for advancing our understanding of the computational processes that undergird metaphorical reasoning. Humans use metaphorical language all the time and without even being aware of it. Even the most subtle instances of a metaphor (via a single word) can instantiate knowledge structures that in turn influence and structure the inferences we make.[24] This means that understanding metaphors from a computational perspective will be crucial as AI systems are increasingly integrated into our everyday lives. We want to ensure that what AI systems *do* aligns with what humans *intend* to communicate through metaphors and related linguistic modes.

To that end, we propose the study of the class of MARC problems as a complementary effort to ARC-AGI. The MARC Project - growing the set of metaphor+ARC puzzles and designing models that can solve them - is meant to inspire the ARC-AGI community to address blind spots that accompany an emphasis on measuring STEM-like reasoning abilities in AI systems (see `https://marc.nkn.uidaho.edu`). The core knowledge priors that motivate ARC-AGI include goal-directness among the other three (objectness, topology, and numerosity).[21] However, goal-directness and other related modes of thinking that make up human reasoning (intentions, behaviors, social constructions, etc) are largely overshadowed by the other "STEM-three" priors; this overshadowing continues through versions ARC-AGI-2 and plans for ARC-AGI-3.[22] We believe that the MARC Project will be a vital complement to the efforts that ARC-AGI inspires, and simultaneously advance our understanding of the role that natural language plays in human reasoning and communication.

References

[1] Sam Acquaviva, Yewen Pu, Marta Kryven, Theodoros Sechopoulos, Catherine Wong, Gabrielle Ecanow, Maxwell Nye, Michael Tessler, and Josh Tenenbaum. Communicating natural programs to humans and machines. *Advances in Neural Information Processing Systems*, 35:3731–3743, 2022.

[2] Bert Baumgaertner and Bernard Molyneux. The paradox of self-consultation and a theory of epistemic work, January 2025. URL `https://philsci-archive.pitt.edu/24676/`.

[3] François Chollet. On the measure of intelligence, 2019. URL `https://arxiv.org/abs/1911.01547`.

[4] K.J.W. Craik. *The Nature of Explanation*. Philosophy: science CAM. Cambridge University Press, 1967. ISBN 9780521094450. URL `https://books.google.com/books?id=wT04AAAAIAAJ`.

[5] Daniel C Dennett. *The intentional stance*. MIT press, 1989.

[6] Daniel Franzen, Jan Disselhoff, and David Hartmann. The llm architect: Solving arc-agi is a matter of perspective. Unpublished manuscript, 2025. URL `https://github.com/da-fr/arc-prize-2024/blob/main/the_architects.pdf`.

[7] Dedre Gentner and Mary Jo Rattermann. Psychology of analogical reasoning. In Robert J. Sternberg, editor, *Thinking and problem solving: Handbook of perception and cognition*, pages 289–313. Academic Press, San Diego, CA, 2 edition, 1994.

[19] A notable exception is [21].

[20] A triggering event in the debate about whether world-model representations can emerge in models like LLMs can be traced back to a model trained to play Othello (see [19]). The idea that of mentals models goes back to at least Kenneth Craik's work, where he mused, "if the organism carries a 'small-scale model' of external reality ... within its head, it is able to try out various alternatives, conclude which is the best of them ... and in every way to react in a much fuller, safer, and more competent manner."[4]

[21] Chollet invokes the work of [22]

[22] We refer to them as "STEM-three" because they serve much of the foundation for education in Science, Technology, Engineering, and Mathematics. By contrast, goal-directness serves as the foundations for areas not typically associated with STEM, such as the humanities, and other areas that are not prototypical of STEM fields, such as behavioral and social sciences.

[8] Dedre Gentner, Keith J Holyoak, and Boicho N Kokinov. *The analogical mind: Perspectives from cognitive science*. MIT press, 2001.

[9] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014.

[10] Damian Hodel and Jevin West. Response: Emergent analogical reasoning in large language models, 2024. URL https://uidaho.idm.oclc.org/login?url=https://www.proquest.com/working-papers/response-emergent-analogical-reasoning-large/docview/2859359146/se-2. Copyright - © 2024. This work is published under http://creativecommons.org/licenses/by/4.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2024-05-03.

[11] Douglas R Hofstadter. *Gödel, Escher, Bach: an eternal golden braid*. Basic books, 1999.

[12] Douglas R Hofstadter and Melanie Mitchell. The copycat project: A model of mental fluidity and analogy-making. 1994.

[13] Douglas R Hofstadter et al. Analogy as the core of cognition. In *The analogical mind: Perspectives from cognitive science*, pages 499–538. The MIT Press, 2001.

[14] Keith J. Holyoak. 234 analogy and relational reasoning. In *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press, 03 2012. ISBN 9780199734689. doi: 10.1093/oxfordhb/9780199734689.013.0013. URL https://doi.org/10.1093/oxfordhb/9780199734689.013.0013.

[15] Mark Johnson. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press, Chicago, IL, 1987.

[16] Akarsh Kumar, Jeff Clune, Joel Lehman, and Kenneth O Stanley. Questioning representational optimism in deep learning: The fractured entangled representation hypothesis. *arXiv preprint arXiv:2505.11581*, 2025.

[17] George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, Chicago, 1980. doi: 10.7208/chicago/9780226470993.001.0001.

[18] Martha Lewis and Melanie Mitchell. Evaluating the robustness of analogical reasoning in large language models. *arXiv preprint arXiv:2411.14215*, 2024.

[19] Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task, 2024. URL https://arxiv.org/abs/2210.13382.

[20] Wen-Ding Li, Keya Hu, Carter Larsen, Yuqing Wu, Simon Alford, Caleb Woo, Spencer M. Dunn, Hao Tang, Michelangelo Naim, Dat Nguyen, Wei-Long Zheng, Zenna Tavares, Yewen Pu, and Kevin Ellis. Combining induction and transduction for abstract reasoning, 2024. URL https://arxiv.org/abs/2411.02272.

[21] Sam Musker, Alex Duchnowski, Raphaël Millière, and Ellie Pavlick. Llms as models for analogical reasoning. *Journal of Memory and Language*, 145:104676, 2025. ISSN 0749-596X. doi: https://doi.org/10.1016/j.jml.2025.104676. URL https://www.sciencedirect.com/science/article/pii/S0749596X25000695.

[22] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

[23] Claire E. Stevenson, Alexandra Pafford, Han L. J. van der Maas, and Melanie Mitchell. Can large language models generalize analogy solving like people can?, 2025. URL https://arxiv.org/abs/2411.02348.

[24] Paul H Thibodeau and Lera Boroditsky. Metaphors we think with: The role of metaphor in reasoning. *PloS one*, 6 (2):e16782, 2011.

[25] Luca H. Thoms, Karel A. Veldkamp, Hannes Rosenbusch, and Claire E. Stevenson. Solving arc visual analogies with neural embeddings and vector arithmetic: A generalized method. *ArXiv*, abs/2311.08083, 2023. URL https://api.semanticscholar.org/CorpusID:265158110.

[26] Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. *Metaphor: A computational perspective*. Springer Nature, 2022.

[27] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

[28] Taylor W Webb, Keith J Holyoak, and Hongjing Lu. Evidence from counterfactual tasks supports emergent analogical reasoning in large language models. *PNAS nexus*, 4(5):pgaf135, 2025.

A Model Prompting and Execution

Here we provide the methodology used to prompt models and parse their outputs. The full implementation (`grade_universal.py`) and results on the other "frontier" models we tested is available here: `https://github.com/bertybaums/MetARC-1`

A.1 Prompt Construction

A single, structured prompt template was used for all models to ensure consistency. This template was composed of four primary sections:

1. **Role and Process:** The prompt began by assigning the model the role of an "expert at solving grid-based logic puzzles" and mandated a specific 4-step reasoning process: (1) read guidance, (2) re-examine examples through that lens, (3) apply the rule, and (4) verify the result.
2. **GUIDANCE Section (Experimental Condition):** This section was dynamically populated based on the experimental condition (`instruct_type`):
   - **Raw Mode (Control):** When no `instruct_type` was provided, the model was given a "RAW MODE" checklist. This checklist prompted it to analyze fundamental properties like objects, invariants, and common transformations (e.g., copy/move, reflect/rotate, flood-fill).
   - **Literal Condition:** The model was provided with the puzzle's `literal_instruction` and explicitly directed to "Follow these steps strictly as the primary lens."
   - **Metaphor Condition:** This was the most structured guidance. The model was instructed to treat the `metaphor` as its "key interpretive frame." It was required to first perform a "PHASE 1 — CONCEPTUAL DIGESTION" by defining the metaphor's:
     (a) Cast & Roles,
     (b) World & Affordances,
     (c) Dynamics,
     (d) Invariants, and
     (e) Goal Signature,
     before translating this conceptual model into grid operations.
3. **Constraints:** A fixed set of constraints was provided to all models, such as "Do NOT introduce new objects" and "prefer the simplest rule that explains ALL examples."
4. **Examples and Test Case:** Finally, the prompt included all training examples (inputs and outputs) followed by the single test input grid.

A.2 Execution and Output Forcing

To ensure parsable and consistent responses, we did not rely on simple text generation. Instead, we used each provider's specific API feature for structured output (e.g., OpenAI's JSON Mode, Claude's Tool Use) via a custom adapter (`llm_adapters.llm_registry`).

Models were called with a low `temperature` (0.1) to promote deterministic output. They were required to return their answer in a specific JSON format defined by our `MARCOUTPUT` schema. This schema strictly required two keys:

- `reasoning`: A string field for the model's detailed, chain-of-thought explanation (as described in the `MARCOUTPUT` schema). This field was the source for our "thinking trace" analysis.
- `output_grid`: A 2D array of strings representing the final predicted grid.

A lenient JSON parser (`_parse_json_lenient`) was used on the raw API response to robustly handle common formatting artifacts, such as surrounding markdown fences (e.g., ‘‘‘`json ...` ‘‘‘), before final validation against the schema.

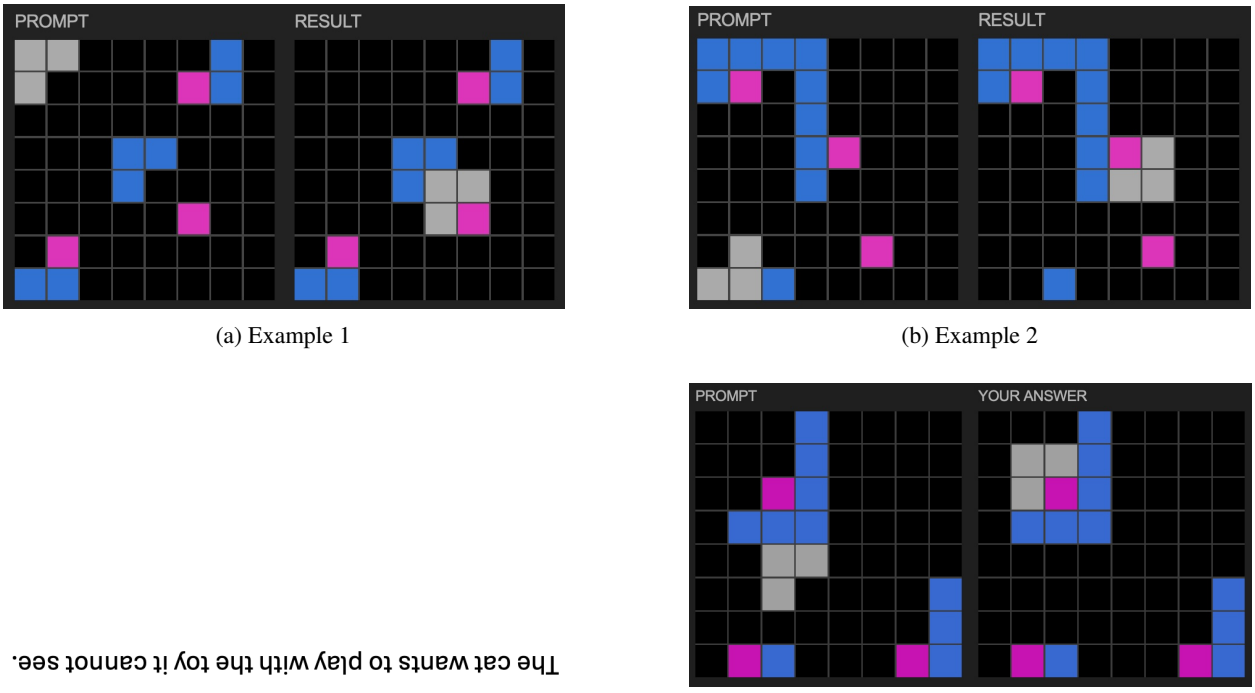
A.3 Analyzing Thinking Traces

We analyze the thinking traces of GPT-5 and Gemini-2.5-pro and created respective reports that can be found in the GitHub repository (`thinking_gpt5_report.html` and `thinking_gemini-2.5-pro_analysis_report.html`)

To analyze the thinking traces, we asked gpt-oss-30k-120b to read the reasoning trace for a run, and decide if the metaphor was actually important in finding the answer. We used the following prompt:

> You are an expert linguistic analyst and AI researcher. Your task is to evaluate an LLM's reasoning trace to determine if it *substantively used* a provided metaphor to solve a puzzle.
> You must distinguish between:
> 1. **Substantive Use:** The model's reasoning *depends on* the metaphor. It uses the metaphor's structure, concepts, or relationships to guide its problem-solving steps, logic, or code generation.
> 2. **Superficial Mention:** The model merely repeats the metaphor's text (e.g., 'The clue was [METAPHOR]') without integrating it into its core logic.

B Solution corresponding to Figure 1



(a) Example 1

(b) Example 2

(c) Example 3

(d) Test

Figure 6: The "cat" puzzle solved.

C Puzzle Key for GPT5 Venn Diagrams

Global Puzzle Key (Passed):

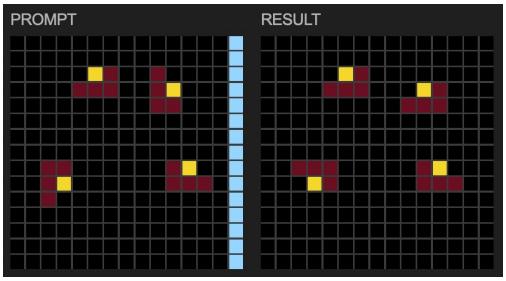
(1): 60/40 split
(2): BMeansP
(3): BURGER
(4): The Cat Wants to Play with the Toy it Cannot See
(5): check the width
(6): Close unneeded doors.
(7): Distance Makes the Heart Grow Fonder
(9): Eligible candidates are red or square, but not above a volume of four.
(10): Follow the yellow brick road
(11): Follow Your Compass
(12): Fry the biggest fish, that you can
(13): GreenTetrisPiece
(14): Hit the Nail on the Head
(15): Ice, Water, Rain
(16): If water was like a deer in headlights
(17): If you give a mouse a cookie...
(18): I hate dandelions!
(19): I'll Only Take the Long Way if You Pay Me
(20): Just one line!
(21): left is greedy
(22): In mating season, males chase the nearest female. If matched in strength, they fight to the death.
(23): Threat of Mutual Destruction, Domination, or Coalition
(24): One Step Closer to Freedom
(25): Raincatchers
(26): Robin Hood
(27): Rocket Ship
(28): season's changes
(29): See no evil, hear no evil, ...
(30): Selling valuable minerals.
(31): Slinky
(32): Speed Racers Cut Corners
(33): Splitting Hairs
(34): Swim in the safest waters
(35): Take a top down approach
(36): The hotter it is, the thirstier you are.
(37): The Wind Will Twirl an Exposed Edge
(38): Tidied the hedges and added some flowers
(39): Time is Money
(40): Up for some sunbathing? Or did you forget your sunscreen?
(41): Victorious play.
(42): We are bonded together.
(43): Here are two examples of what not to do, and one example of what you should do. Now demonstrate you understand.
(44): What's under the dog?
(45): Wheel of Fortune
(46): Who's stuck with guard duty?
(49): Drop the stone
(51): Fall Vibes
(52): Find the path and follow it.
(53): Fish that see the fisherman don't take the bait.
(54): The floor is lava.
(55): Good pups leave it, at least when told.
(56): Hold the pencil. Draw the letter H.
(58): I'll only help if I have to.
(59): Mixing Colors
(60): Protect as many as you can.
(61): Rapunzel, Rapunzel, let down your hair!
(62): Seeing an escape is believing there's escape.
(63): Small bubbles float, big bubbles pop.
(65): The domino theory
(66): The flowers grow and bloom
(67): They hid from the Lord God among the trees of the garden.
(69): You know if someone else knows.
(71): they never touched

Unsolved Puzzles:

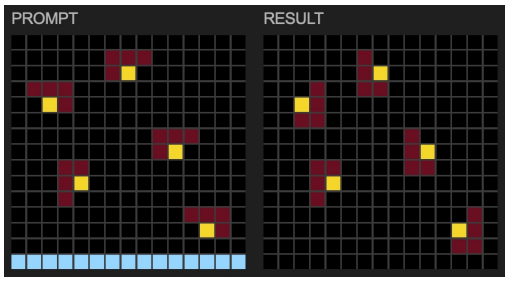
(8): DontBuildNearSharks
(47): Captured the moment
(48): Direct the flow of the river into the canyon.
(50): It's easy as pi
(57): How does it move?
(64): The cat wants to play with the toy it can see.
(68): You could smell it from a mile away.
(70): do, re, mi, fa, sol, fa, mi, re, do

Figure 7: Key for puzzle numbers in Figure 4 and their titles.

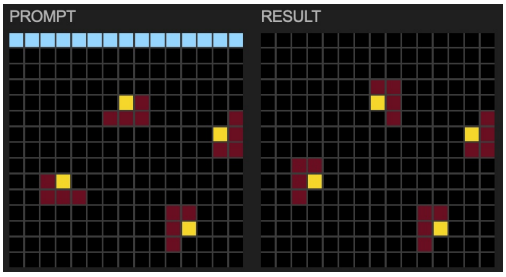
D Puzzle #37: The wind…



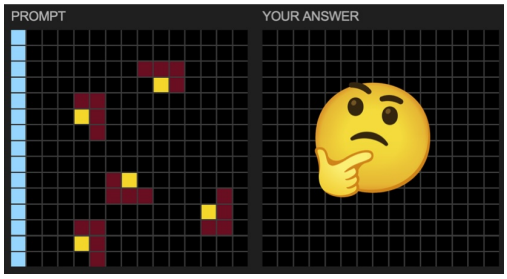
(a) Example 1



(b) Example 2



(c) Example 3



(d) Test, with solution not shown.

Figure 8: Puzzle #37 “The wind will twirl an exposed edge”