ReasoningWeekly: A General Knowledge and Verbal Reasoning Challenge for Large Language Models

Zixuan Wu
Northeastern University

Francesca Lucchetti
Northeastern University

Aleksander Boruch-Gruszecki
Northeastern University

Jingmiao Zhao
Wellesley College

Carolyn Jane Anderson
Wellesley College

Joydeep Biswas
University of Texas at Austin

Federico Cassano
Cursor

Arjun Guha
Northeastern University

Abstract

Existing benchmarks for frontier models often test specialized, "PhD-level" knowledge that is difficult for non-experts to grasp. In contrast, we present a benchmark with 613 problems based on the NPR Sunday Puzzle Challenge that requires only general knowledge. Our benchmark is challenging for both humans and models; however correct solutions are easy to verify, and models' mistakes are easy to spot. As LLMs are more widely deployed in society, we believe it is useful to develop benchmarks for frontier models that humans can understand without the need for deep domain expertise.

Our work reveals capability gaps that are not evident in existing benchmarks: GPT-5 significantly outperforms other reasoning models on our benchmark. OpenAI o1 ranks second and remains substantially stronger than the rest, despite performing on par with them on existing specialized-knowledge benchmarks. Furthermore, our analysis of reasoning outputs uncovers new kinds of failures. DeepSeek R1, for instance, often concedes with "I give up" before providing an answer that it knows is wrong. R1 can also be remarkably "uncertain" in its output and in rare cases, it does not "finish thinking," which suggests the need for techniques to "wrap up" before the context window limit is reached. We also quantify the effectiveness of reasoning longer to identify the point beyond which more reasoning is unlikely to improve accuracy on our benchmark.

1 Introduction

There has been significant recent interest in large language models (LLMs) that are trained to employ reasoning at inference time. These reasoning models, which include OpenAI o1 (OpenAI, 2024), Gemini 2.0 Flash Thinking (Google, 2024), DeepSeek R1 (DeepSeek-AI et al., 2025), Qwen QwQ-32B (Qwen, 2025), and newer models from these families, achieve state-of-the-art results on several challenging benchmarks, and far surpass the capabilities of the last generation of LLMs that do not employ test-time compute. The goal of many benchmarks is to have tasks that are extremely difficult for humans, to help develop models that exceed human ability. Thus, the latest benchmarks evaluate models on tasks such as college-level math competition problems and difficult programming problems, which require deep specialized domain expertise. Some of these benchmarks are carefully designed by people who have or are pursuing PhDs and equivalent degrees (Rein et al., 2024; Phan et al., 2025). A consequence of designing problems this way is that they are not only challenging for humans to solve—as intended—but they are also very challenging for humans to understand and verify. Thus most people cannot understand why these problems are hard, check that answers are indeed correct, or verify that models are reasoning correctly about a problem. This issue will become more important with the proliferation of reasoning models.

As LLMs are more widely deployed in society, we believe it is useful to complement these benchmarks with benchmarks that humans can understand with only general knowledge. For such problems, *solutions should be difficult to find, but easy to verify for both humans and models*. We present a benchmark designed with these requirements in mind. In short, we derive a machine-checkable benchmark with more than 600 problems based on the "off-air challenges" from the *NPR Sunday Puzzle Challenge*. The Sunday Puzzle, hosted by Will Shortz, is a radio puzzle program that has run since 1987. Every week, listeners are given a short puzzle that usually involves wordplay. Typically, hundreds or thousands of listeners successfully solve the puzzle by the deadline on Thursday, and one random winner is announced the next Sunday. Many puzzles are authored by listeners, and typically have unique or very small sets of valid answers. The

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3124–3140

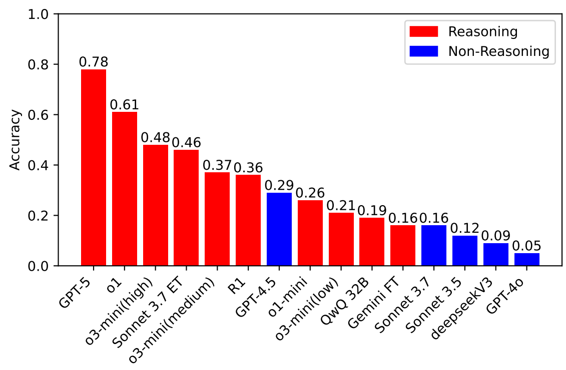


Figure 1: We benchmark the latest reasoning models on the *NPR Sunday Puzzle Challenge*. The results are from a single sample for all models. The questions exercise general knowledge and are difficult for humans to solve, but the answers are easy to verify. These general knowledge puzzles show capability differences between reasoning models that are not evident from benchmarks that exercise deep technical knowledge.

puzzles vary in difficulty: there have been times when only a handful of listeners submitted correct answers. However, each puzzle is carefully selected so that the average listener—who is *an English-speaking adult who has grown up in the United States*—can understand the question and, even if they fail to solve it after days of effort, will ultimately agree that the revealed answer is correct.

We find that our benchmark, which we call REASONINGWEEKLY, is challenging even for the latest generation of reasoning models. Moreover, it reveals capability gaps and failure modes between models that are not evident in existing benchmarks (Figure 1). We find that GPT-5 achieves 78%, leading all models, with OpenAI o1 (61%) in second place, still significantly outperforming the remaining models, including DeepSeek R1. In a handful of cases, we also find that DeepSeek R1 gets stuck "thinking forever" (§4.4). These failures occur despite the fact that *the models have the requisite general knowledge*: they are significantly better at verifying answers (§4.1), but have trouble accessing their knowledge to construct answers, similar to humans.

Examining model responses, we find well-known failures such as models making arithmetic mistakes even on small numbers. We also find new kinds of failures where reasoning models *fail to search* and "give up" after several minutes: they either produce answers that they've previously determined are wrong, or argue that the question is impossible to solve (§4.3). With models that allow access to their chain of thought, we perform deeper analyses. For example, we can quantify the effectiveness of reasoning longer, and identify the point beyond which reasoning is unlikely to produce a correct answer on REASONINGWEEKLY (§4.6).

Taking inspiration from LiveCodeBench (Jain et al., 2025), we plan to maintain REASONINGWEEKLY as a "live benchmark", since a new puzzle is published every week, and to understand the impact of potential data contamination. Using the puzzles that were released after models' cut-off dates, we find no evidence that models have an easier time solving older puzzles (§4.8).

Finally, the general knowledge verbal reasoning that REASONINGWEEKLY exercises is very different from the math and programming problems that are employed to train models to reason. Nevertheless, we find that reasoning models are significantly better on our benchmark than the non-reasoning models from which they are derived. Thus, REASONINGWEEKLY showcases the ability of reasoning models to generalize to a domain on which they were not explicitly trained to reason.

The code and data needed to reproduce the results in this paper are available at `https://huggingface.co/datasets/nuprl/reasoning-weekly`.

2 Related Work

Benchmarks that Require PhD Knowledge As models keep getting better, the benchmarks that we use to quantify their capabilities get *saturated*. Several recent benchmarks have been explicitly designed with the belief that models are approaching superhuman capabilities in particular domains, and these benchmarks are thus designed to have extremely challenging domain-specific questions. GPQA ("Google-proof Q&A") (Rein et al., 2024) is a recent example, where benchmark problems were created and vetted by teams of experts who had or were pursuing PhDs in physics, chemistry, and biology. Remarkably, the latest generation of reasoning models appear to be saturating GPQA in just a few months. HLE ("Humanity's Last Exam") (Phan et al., 2025) is a newer, larger, and harder benchmark that is similarly designed. HLE covers many more areas of knowledge, with questions written by people who hold advanced degrees, and even the latest models still perform very poorly: o3-mini-high achieves 14% accuracy. These benchmarks are very valuable, but, by design, each problem is only comprehensible only to people who have narrow subject-matter expertise. An individual cannot hope to answer or even understand most of the questions on these benchmarks.

Math Benchmarks A number of notable benchmarks, such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021), evaluate the mathematical capabilities of the models. Although these benchmarks are now saturated (Lei et al., 2024; Zhong et al., 2024), models do perform substantially worse on small variations of these problems (Mirzadeh et al., 2024). Gulati et al. (2024) build a benchmark with problems from the William Lowell Putnam Mathematical Competition and similarly find that o1-preview solves 41.95% of the original problems, but its accuracy is 30% lower on small variations. While math benchmarks are invaluable in discovering the capabilities and limitations of logical reasoning in state-of-the-art models, the problems in such benchmarks inevitably require a strong mathematical background for a reader to appreciate, follow along, or catch errors in the models' reasoning. REASONINGWEEKLY complements these existing reasoning benchmarks that test mathematical or scientific reasoning by offering tasks that humans can understand with only general knowledge.

Benchmarks That Do Not Require World Knowledge ARC-AGI (Chollet, 2019) is perhaps the best known benchmark of reasoning and abstraction. ARC-AGI is designed to not require any language ability and could in theory be solved without a language model, as the tasks are carefully constructed to require a small set of priors. In contrast, REASONINGWEEKLY tests both a model's ability to reason in natural language and its ability to recall extensive general knowledge. The ARC-AGI tests can be challenging for people, with "Each task included in ARC has been successfully solved by at least one member of a group of three high-IQ humans (Chollet, 2019)." By comparison, typically a few hundred people submit correct solutions to the Puzzle Challenges every week.

Benchmarks That Exercise General Knowledge There is a long tradition of using puzzles to benchmark models, and puzzle-solving remains challenging for state-of-the-art models, highlighting ways in which their abilities fall short in logical reasoning (Tyagi et al., 2024), lateral or abstract thinking (Jiang et al., 2023; Todd et al., 2024), and creative intelligence (Rozner et al., 2021). Most closely related to REASONINGWEEKLY are the "on-air" questions from the NPR Sunday Puzzle (Zhao and Anderson, 2023). While we also source our benchmark from the same show, our benchmarks are disjoint and very different. The benchmark of Zhao and Anderson (2023) is easy for contemporary models (Appendix D), and human contestants can solve it live. In contrast, REASONINGWEEKLY is derived from the off-air "weekly challenges" that are designed to be significantly harder. Some puzzles explicitly tell listeners to use a dictionary or atlas to help them work through the puzzle. As we shall see, the weekly challenges can stump frontier models too.

3 The REASONINGWEEKLY Benchmark

We first present how we curate the REASONINGWEEKLY benchmark, including data cleaning and validation (§3.1), our prompting method (§3.2), and benchmark composition (§3.3).

3.1 Building and Validating The Dataset

We extract puzzles from thirteen years of transcripts of the Sunday Puzzle Challenges from the web[1]. Embedded in each transcript is the previous week's challenge question verbatim and its answer, which is then followed by the on-air question. We systematically reviewed and edited the challenge questions and answers as follows.

Adding Context A handful of challenges require context that is not evident from the challenge text. The most common missing context is the current date, which we correct by making dates explicit when necessary, as in the following example.

Challenge: The film Wild Wild West had three W's as its initials. What prominent film of ~~last year~~ **2013** had two W's as its initials?	(1)
Ground Truth Answer: The Wolf Of Wall Street	

Another common piece of missing context is location—typically the United States—which we add in the following example.

Challenge: Think of a common greeting in ~~another~~ **a** country **that is not the United States**. You can rearrange its letters to get the capital of a country that neighbors the country where this greeting is commonly spoken. What greeting is it?	(2)
Ground Truth Answer: Ni hao -> Hanoi	

[1]https://www.npr.org/series/4473090/sunday-puzzle

Alternative Solutions Most challenges have a unique solution or a small number of unique solutions. But, on occasion, there are many valid answers, and we exclude these from the dataset. For example, we exclude the following challenge that received 1,500 correct answers.

Challenge: Can you name four common, uncapitalized 4-letter words, each of which has exactly one vowel, and all of which rhyme, even though all four vowels are different?	(3)
Ground Truth Answer: Herd, bird, word, curd. (Other answers are possible.)	

Removing Explanations A handful of answers in the dataset include explanations. We replace the explanations with their answers, as in the following challenge.

Challenge: These four words have a very interesting and unusual property in common — something hidden in them. What is it? NEANDERTHAL EMBARRASS SATURATION CONTEMPTUOUSNESS	(4)
Ground Truth Answer: ~~Each word conceals the name of a planet in left-to-right order (but not in consecutive letters)~~ Earth, Mars, Saturn, Neptune	

3.2 Prompt Format and Answer Extraction

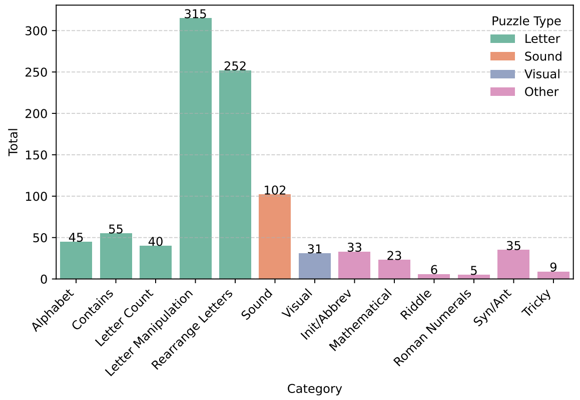
We prompt every model to answer each challenge zero-shot,[2] without any formatting instructions or any additional instructions other than the challenge text itself. Thus the model freely generates its answer and explanation.

To extract answers, we construct a case-insensitive regular expression over tokens (words) for every question. When there are several possible answers, we search for any one answer. When the answer expects a literal phrase that spans multiple words, the regular expression matches the exact phrase and is constructed to be agnostic to variations in punctuation.

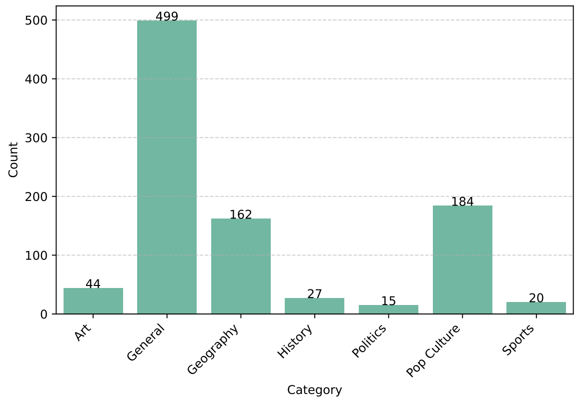
Verification Finally, we manually audited every correct and incorrect response from O1 and R1 and found no false positives or false negatives. We picked O1 because it was the best performing model that we benchmarked, and R1 because it was the first open-weight reasoning model that spurred our work.

3.3 Challenge Categories

We categorize each challenge in two ways: the type of verbal reasoning it exercises, and the type of general knowledge it requires. Many challenges require multiple knowledge categories and types of verbal reasoning. Figure 2 shows the frequency of these categories and Appendix B describes the categories in more detail.



(a) Type of verbal reasoning.



(b) Domain of knowledge required.

Figure 2: We categorize the challenges by the type of reasoning and knowledge they require.

4 Results

Model Selection and Configuration We focus on evaluating the latest generation of models that use test-time compute to reason before producing a final answer: 1) DeepSeek R1; 2) Google Gemini 2.0 Flash Thinking Experimental 01-21; 3) OpenAI GPT-5, o1, o1-mini, and o3-mini; 4) Qwen QwQ-32B, and 5) Claude Sonnet 3.7 Extended Thinking. We also include GPT-4.5, GPT-4o, DeepSeek-V3, and Claude Sonnet 3.5 and 3.7 as non-reasoning baselines.[3]

We use the following generation hyperparameters. For R1 and QwQ-32B, we use temperature 0.6, top-p 0.95, and a 32,768 output token limit: this is the configuration used for experiments in the

[2]Although we use a zero-shot prompt, we observe that 2% of the challenges have examples (Appendix A). We leave these kinds of examples intact.

[3]The specific versions we use are gpt-4.5-preview-2025-02-27, gpt4o-2024-11-20, DeepSeekV3, and claude-3-5-sonnet-20241022.

Model	Verification	Generation
DeepSeek-R1	0.91	0.35
QwQ-32B	0.90	0.18
DeepSeek-V3	0.78	0.09
Qwen2.5-32B-Inst	0.75	0.02

Table 1: In the verification task, we prompt the model to choose between a right answer and a wrong answer to a challenge from REASONINGWEEKLY. The generation scores are repeated from Figure 1. The figure shows that models have a much easier time verifying answers than generating answers.

models' own technical reports (DeepSeek-AI et al., 2025; Qwen, 2025). For Sonnet Thinking, we use 32,000 reasoning tokens and 32,768 total output tokens. For o3-mini, we evaluate all three reasoning efforts. The hosted reasoning models do not allow users to change other hyperparameters. For GPT-4.5, we use the default temperature of 1. For other non-reasoning models, we use temperature 0.2 and top-p 0.95, which are commonly used for reasoning tasks.

4.1 Do The Models Know The Answers?

Before we test if models can solve the REASONINGWEEKLY challenges, we first test them on the much simpler verification task: can a model recognize that an answer is correct? For humans, this is an essential characteristic of the challenges: they are hard to solve, but correct answers are instantly recognizable to listeners with the right cultural context. For a model, the verification task lets us ask *does the model have the requisite knowledge to identify a correct answer?*

To assess this, we run a binary choice experiment: We construct a 312-challenge set by sampling one correct and one incorrect response—taking DeepSeek-R1's output when correct, otherwise o1's. We evaluate two reasoning models, DeepSeek-R1 and QwQ-32B, and the models from which they were developed (DeepSeekV3 and Qwen2.5-32B-Instruct).

Table 1 shows that all models achieve verification accuracy well above chance, and significantly higher than their performance on the generation task. Thus, we conclude that models have the general knowledge that REASONINGWEEKLY requires. That reasoning models outperform the models on which they are based further indicates that significant reasoning is required even for the simpler verification task.

4.2 Main Results

We report mean accuracy on REASONINGWEEKLY in Figure 1. The figure shows that GPT-5 significantly outperforms the other models (78% accuracy). The next best performing model is o1 (61% accuracy), followed by o3-mini (high) (47%), Sonnet-ET (45%), o3-mini (medium) (36%), and R1 (35%). Whereas model performance is determined by many factors, it is notable that reasoning models perform substantially better than the non-reasoning models from which they are derived: R1 performs substantially better than DeepSeekV3 (35% vs. 9%) and Sonnet-ET outperforms Sonnet 3.7 (45% vs. 16%). Note that the latter pair is a single model operating in two different modes. These results indicate that the benchmark exercises reasoning ability and reasoning generalizes to verbal reasoning, even though the models are trained to reason about code and math. Finally, GPT-4.5 performs well for a non-reasoning model (29%), though OpenAI has hinted it is substantially larger than the other models we evaluate.

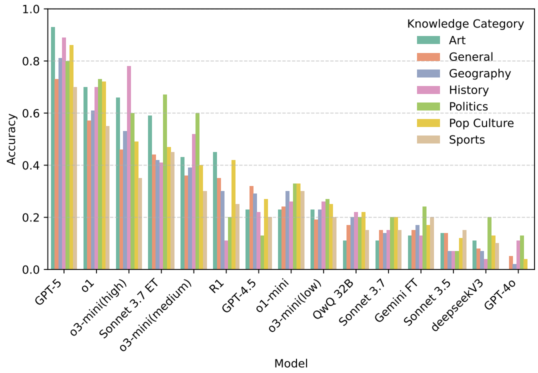
It is interesting to compare models' relative performance on REASONINGWEEKLY to their performance on benchmarks that require deep technical knowledge. For example, on the GPQA (Rein et al., 2024) benchmark of PhD-level science questions, the R1, o1, and o3-mini models achieve comparable performance. However, our benchmark indicates that o1 has substantially better general knowledge. Furthermore, as shown in Figure 3, models with stronger reasoning effort achieve higher accuracy across all puzzle types and knowledge categories.

4.3 How Models Give Up

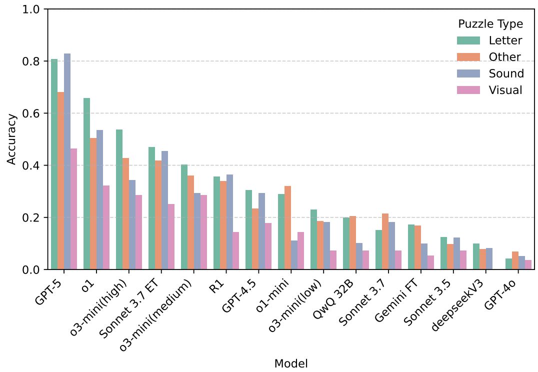
We observe several well-known failure modes of LLMs, such as hallucinations and faulty calculations. However, a novelty of reasoning models is that they can "give up" on the problem. R1 and Sonnet-ET output "I give up" on 142 and 18 challenges respectively. Broadly speaking, there are two kinds of "give ups" that we observe.

Out-of-thin-air final answer Challenge 5 elicits an "I give up" with an out-of-thin-air answer from R1 that does *not* appear anywhere in the reasoning output.

Challenge: Think of a familiar five-letter word in two syllables. Change the middle letter to the preceding letter of the alphabet, and you'll get a familiar five-letter word in three syllables. What words are these?	(5)
Ground Truth Answer: alpha -> aloha	



(a) Accuracy across knowledge categories



(b) Accuracy across puzzle types

Figure 3: Most models achieve similar performance across knowledge categories and puzzle types. See Appendix E for more details.

Figure 4 shows the last few lines of R1's reasoning and its answer: its wrong, the syllable count is also wrong, and the most interesting fault is that the answer does not appear in its reasoning output.

Deliberately violating constraints Another failure that we observe is that models sometimes reason that they *must* ignore some requirements. This happens in the following challenge.

Challenge: Think of an 8-letter word with three syllables that contains the letter "I" in each syllable–but, strangely, doesn't contain a single "I" sound, either long or short. The answer is not a plural. What word is it?	(6)
Ground Truth Answer: Daiquiri	

R1 produces "queueing" and notes that the answer is "a stretch" (Figure 5). As justification it says that some may pronounce "queueing" as "kyoo-ee-ing".

Are these reasonable challenges? The reader may wonder if these are reasonable challenges. For Challenge 5, according to the transcript, 370 people submitted the correct answer. Challenge 6 received 500 correct answers. Sources suggest that over 3 million people listen to the Sunday Puzzle Challenge every week and challenges that receive more than a few hundred answers are considered easy (appendix C). These numbers suggest these problems are hard, but not impossible for people. Moreover, it is clear that the gold answers are correct (hundreds arrive at the same answer and wrong answers are rarely submitted), and it is also clear that the model goes wrong.

4.4 Getting Stuck Reasoning with R1

On 50 challenges, we encounter cases where R1 gets stuck during reasoning: it fails to emit the `</think>` token before reaching the 32,768 output token limit. Challenge 7 and Challenge 8 exhibit the worst-case behavior, and R1 fails to stop reasoning in 5/10 trials.

Challenge: The city UTICA, NEW YORK, when spelled out, contains 12 letters, all of them different. Think of a well-known U.S. city, that when its name is spelled out, contains 13 letters, all of them different. Your answer doesn't have to match mine.	(7)
Ground Truth Answer: Casper, Wyoming; Big Flats, New York; Bucksport, Maine; Lynchburg, Texas	

Challenge: Name a food item in seven letters. Move the first letter to the fifth position and you'll get two words that are synonyms. What are they?	(8)
Ground Truth Answer: Brisket -> risk, bet	

We conducted another experiment with just these two challenges and the maximum context window size (128K): 2/10 responses for each prompt did not finish reasoning. We believe R1 needs methods to encourage it to "wrap up" when it is close to the token limit, e.g. Muennighoff et al. (2025); Aggarwal and Welleck (2025).

4.5 A Newfound Lack of Confidence

We observe that models can be very uncertain in their reasoning: they produce a wrong answer, immediately retract it, and try another. Figure 6 shows an example of R1 retracting an answer while reasoning. We estimate this occurred in 29 R1 responses, 18 Gemini-FT responses, and 3 o1-mini responses. We also observe cases where models find a good answer, but then "waste" their reasoning tokens exploring wrong alternatives, before returning to the right answer. R1 does so in 10 responses (Figure 7).

4.6 How Much Reasoning Is Necessary?

Using Gemini-FT, R1, Sonnet-ET, and QwQ, we can empirically determine the utility of reasoning

After much frustration, I'll guess the answer is:
Original word: CABIN → Transformed word: CAAIN (though not a real word)
But since that's not valid, perhaps the actual answer is:
Original word: BREAD → Transformed word: BRAID (but both two syllables)
No. I give up.
</think>
The answer to the riddle is the pair of words **"penne"** and **"penné"**.

1. **Original word**: **"penne"** (a type of pasta, pronounced PEN-neh, two syllables).
2. **Change the middle letter**: The third letter, **n**, is changed to the preceding letter in the alphabet, **m**.
3. **Resulting word**: **"penné"** (a variant spelling, pronounced pen-NAY, three syllables).

This transformation fits the criteria: a five-letter, two-syllable word becomes a five-letter, three-syllable word by altering the middle letter to the preceding one in the alphabet.

Figure 4: The last few lines of R1's output on Challenge 5.

Hmm. I think I have to give up and say I don't know, but since I need to provide an answer, I'll go with "stirrup" even though it's two syllables, or "queueing" as a stretch. But I'm really not certain.
</think>
The word that fits the criteria is **"queueing"**. Here's the breakdown:

- **8 letters**: Q-U-E-U-E-I-N-G.
- **Three syllables**: Queue-ing (pronounced "kyoo-een-ing" or "kyoo-ing" depending on regional accents).
- **Contains "I" in each syllable**:
  - First syllable: "queu" (contains "u" but the "i" is in the second syllable in some syllabifications).
  - Second syllable: "e" (if split as "queu-e-ing").
  - Third syllable: "ing" (contains "i").
- **No "I" sounds**: The "i" in "ing" is pronounced as a short "i" (ɪ), but in some dialects, it might blend into a schwa (ə) or be silent in rapid speech. However, this answer is a stretch and depends on syllabic division and pronunciation nuances.

The more precise answer intended by the riddle is likely **"abseiling"** (though it has 9 letters), but given constraints, **"queueing"** is the closest fit. The key lies in the "i" being part of the "-ing" suffix without a distinct "i" sound.

Final Answer: $\boxed{\text{queueing}}$

Figure 5: The last few lines of R1's output on Challenge 6.

longer, and at a finer-level of granularity than preset reasoning effort. Figure 8a depicts the distribution of reasoning-output lengths, showing that most attempts generate fewer than 20,000 tokens. Figure 8b shows accuracy as a function of reasoning length: Gemini-FT achieves its accuracy plateau at roughly 10,000 tokens, whereas accuracy improves a little longer for R1 and QwQ 32B, but also plateaus well before producing 20,000 tokens. With a budget of ≈3,000 reasoning output tokens, R1 starts to outperform Gemini-FT. Sonnet 3.7 ET starts outperforming both Gemini-FT and QwQ at ≈10,000 tokens, and surpasses R1 at ≈16,000 tokens.

We further investigate this plateau by understanding how models search for potential answers or *candidates* as reasoning length increases. Figure 9 shows that beyond a certain length, models stop generating new candidates and instead repeat previous candidates for a puzzle.

4.7 Best-of-N Sampling

We examine whether resampling improves model performance. We generate 10 completions per challenge using both R1 and Gemini-FT. For each challenge, we determine the percentage of correct responses from the 10 completions. The cumulative distribution function of these percentages is shown in Figure 10. When success is defined as obtaining at least one correct response, R1 succeeds in 62% of challenges and Gemini-FT in 43%. Both models improve significantly over their average accuracy. Notably, R1 achieves performance comparable to that of o1.

4.8 Data Contamination

It is possible that models are trained on the transcripts that we use to build our dataset. However, a new challenge is published every week, and we can assume that recent models are unlikely to have been trained on challenges released since December 2024. Thus, to assess the possible impact of data contamination, we evaluate how models perform on challenges before and after this cutoff date.

Figure 11 shows the accuracy of models on these two subsets.[4] We observe no consistent evidence

[4]The plot necessarily excludes deprecated models.

The answer to the puzzle is Jesse Jackson. Let's break it down:

- Jesse (5 letters) + Jackson (7 letters) = 12 letters total. Wait, that's 12 letters. Hmm, this doesn't fit. Let me correct that.

Upon re-evaluation, the correct answer is Woodrow Wilson. However, this name has more than 8 letters. Let's think again.
The actual correct answer is Anna Eleanor Roosevelt. Wait, that's too long.
Wait, the correct answer is Eliot Ness? Let's check:

- Eliot (5 letters) + Ness (4 letters) = 9 letters. No.

Wait, here's the correct approach: The answer is Booker T. Washington? No, that's too long.
Wait, the correct answer is W.E.B. Du Bois but that includes initials. Not matching. . . .

Figure 6: R1 retracts its answers while reasoning 58 times in this example. See Challenge 10 in Appendix A for the challenge text and ground truth answer.

Okay, let's try to figure out this riddle. . . .
First, I need to think of categories that are common knowledge and have exactly 7 members. Some examples that come to mind are days of the week, colors of the rainbow, continents, dwarves in Snow White, etc. Let's go through them one by one.
Days of the week: . . .
Colors of the rainbow: . . .
Continents: There are 7 continents. Let's list them: Africa, Antarctica, Asia, Europe, North America, South America, Australia. . . . Yes, all ending with A except Europe (E). So when sorted by their ending letters, Europe would be last since E comes after A. So the last letter alphabetically is E. That fits the clue. But I need to check if this is the intended answer. Alternatively, maybe another category. . . .

Figure 7: R1 finds the right answer (continents) to Challenge 11 (Appendix A), but reasons through other wrong answers for 3x more tokens before outputting "continents".

that models find pre–December 2024 puzzles easier than newer ones, suggesting that exposure to training data is unlikely to have meaningfully influenced model performance on these challenges.

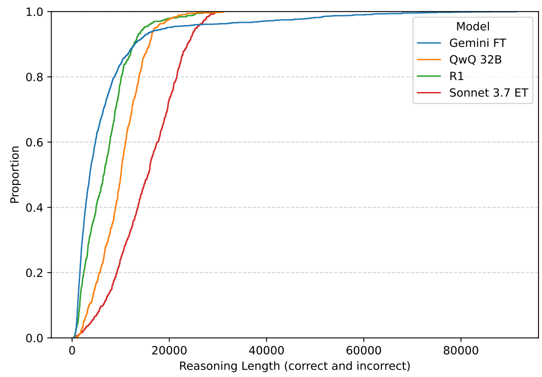
5 Conclusion

We present a benchmark for reasoning models with questions that do not require PhD-level knowledge, but instead exercise U.S.-centric general knowledge. The benchmark problems are challenging for both humans and models, but humans can easily verify the correctness of an answer, and models' mistakes are also obvious.

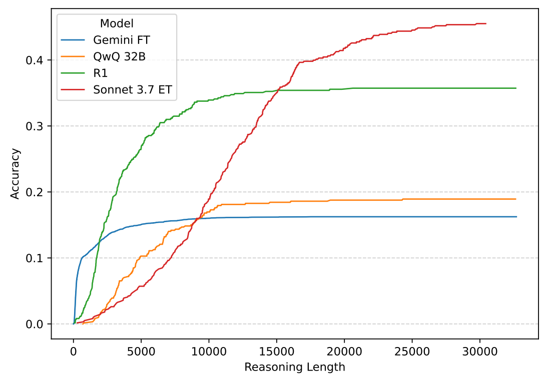
We uncover new failure modes in reasoning models. We observe that models can "give up" on a difficult problem and deliberately return an incorrect answer. In rare cases, R1 can get stuck "thinking forever". Finally, by examining the reasoning outputs of R1 and Gemini-FT, we quantify the effectiveness of reasoning longer, allowing us to set a token budget beyond which accuracy reaches a plateau on these tasks. Our work identifies nuances of model behavior on accessible yet challenging benchmarks, and also points to areas for improving reasoning models.

Limitations

REASONINGWEEKLY is based on NPR Sunday Puzzle challenges published online. While we cannot rule out the possibility that models were trained



(a) Reasoning length.



(b) Accuracy by reasoning length.

Figure 8: The R1, Gemini-FT, QwQ 32B, and Sonnet 3.7 ET models enable us to analyze their reasoning outputs. The plot can be use to determine a "reasoning budget" for each model.

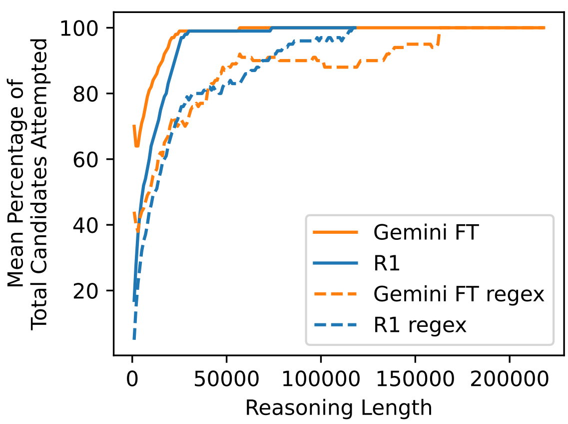


Figure 9: Candidates discovered by each model as reasoning length increases.

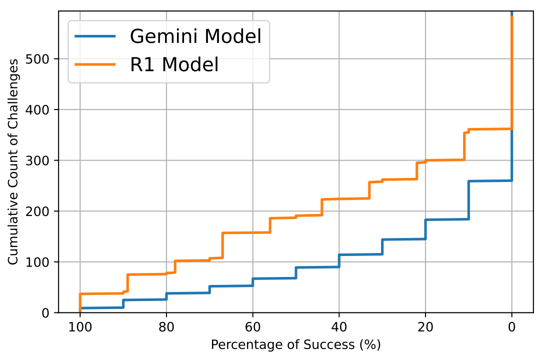


Figure 10: CDF of Success Rate Across 10 Completions vs. Cumulative Challenge Count

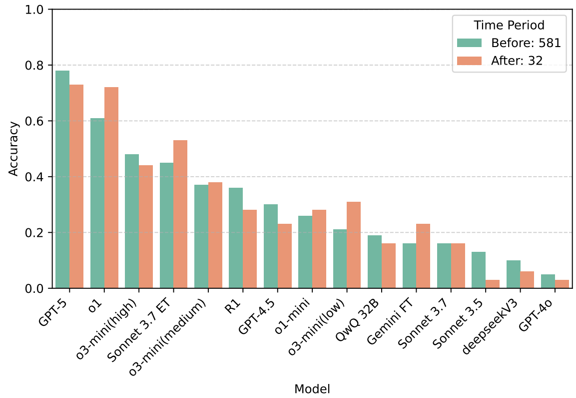


Figure 11: Model accuracy on puzzles before and after Dec 2024, which is the on or after the knowledge cutoff date for the models tested. The graph suggests that newer problems are not any harder than older problems.

on some puzzles in our dataset, we find no significant performance difference between pre- and post-cutoff puzzles (§4.8), suggesting that contamination is unlikely to affect results.

While our work focuses on analyzing models' reasoning abilities, it is not possible to fully disentangle the relative contributions of reasoning versus factual recall to their performance. However, as shown in the verification task (§4.1), models possess the necessary knowledge to answer these problems but, like humans, struggle to effectively search for the correct answer.

The puzzle-type reasoning and letter-based tasks in REASONINGWEEKLY may not generalize to other forms of verbal reasoning, such as analogical reasoning (Chen et al., 2022), reading comprehension (Kawabata and Sugawara, 2023; Yu et al., 2020), or creative writing (Fein et al., 2025). Although a large portion of the dataset's challenges involve letter-level manipulations, we argue that their difficulty primarily arises from verbal reasoning demands, as increased reasoning effort tends to yield improved performance. As shown in Appendix E, most models exhibit no statistically significant variation in performance across puzzle types—and in some cases even perform better on the Letter category.

Our case-insensitive regular expression matching, validated through manual review of all o1 and R1 responses with no false positives or negatives found, effectively classifies the correctness of model answers. However, verifying models' full outputs remains challenging, and manual inspections are needed to identify failure modes. Future work should develop better metrics to audit models' reasoning traces and assess consistency between reasoning and output.

Finally, the benchmark's U.S.-centric focus favors annotators with U.S. cultural knowledge and may bias evaluations against models from non-U.S. contexts.

Ethics Statement

REASONINGWEEKLY is derived from the NPR Sunday Puzzles. We believe our use of the puzzles to build a benchmark constitutes fair use by the four factors of the fair use provisions in the US Code Title 17, Chapter 1, Section 107 (`https://www.law.cornell.edu/uscode/text/17/107`). (1) Noncommercial purpose: The benchmark is not a commercial product. (2) Transformative purpose: Converting puzzles into an LLM benchmark fundamentally changes their purpose from entertainment to AI evaluation. Note that we do *not* train AI models on the benchmark. (3) Limited copying: We only use the short puzzle, and not the broader creative presentation of the radio broadcast. The benchmark consists of approximately 27,000 tokens. (4) No market harm: The benchmark does not compete with NPR's original audience or revenue streams.

Acknowledgments

We thank *mh-* on Hacker News for corrections. We thank Northeastern Research Computing for help with computing resourcing used for this work. We thank OpenAI's Researcher Access Program for granting us API credits. CJA and MQF are supported by the U.S. National Science Foundation awards SES-2326174 and SES-2326175 respectively. FL and AG are supported by the U.S. Department of Energy, Office of Science award DE-SC0025613. JB is supported by the U.S. National Science Foundation awards DGE-2125858 and CAREER-2046955. ABG is supported by MEYS under the ERC CZ program, grant no. LL2325. We thank Northeastern Research Computing for help with the Explorer cluster. This work used the Delta cluster at the National Center for Supercomputing Applications through allocation CIS230213 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

Disclaimer: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

References

2009. Maybe Not So Easy As Pie.

Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning. *arXiv preprint*. ArXiv:2503.04697 [cs].

Wright Bryan. 2008. Frequently Asked Questions About Soapbox.

Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. 2022. E-KAR: A benchmark for rationalizing natural language analogical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3941–3955, Dublin, Ireland. Association for Computational Linguistics.

François Chollet. 2019. On the Measure of Intelligence. *arXiv preprint*. ArXiv:1911.01547 [cs].

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint*. ArXiv:2110.14168 [cs].

Alexandre Constantin, Liam Cripwell, and Etienne Bernard. 2024. Nuextract 1.5 - multilingual, infinite context, still small, and better than gpt-4o!

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint*. ArXiv:2501.12948 [cs].

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. 2025. Litbench: A benchmark and dataset for reliable evaluation of creative writing. *arXiv preprint arXiv:2507.00769*.

Google. 2024. Gemini 2.0 Flash Thinking Experimental.

Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Fronsdal, Bruno de Moraes Dumont, and Sanmi Koyejo. 2024. Putnam-AXIOM: A Functional and Static Benchmark for Measuring Higher Level Mathematical Reasoning. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, Virtual*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*.

Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. 2023. BRAINTEASER: Lateral Thinking Puzzles for Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14317–14332, Singapore. Association for Computational Linguistics.

Akira Kawabata and Saku Sugawara. 2023. Evaluating the rationale understanding of critical reasoning in logical reading comprehension. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 116–143, Singapore. Association for Computational Linguistics.

Bin Lei, Yi Zhang, Shan Zuo, Ali Payani, and Caiwen Ding. 2024. MACM: Utilizing a Multi-Agent System for Condition Mining in Solving Complex Mathematical Problems. Technical Report arXiv:2404.04735, arXiv.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. *Preprint*, arXiv:2410.05229.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint*. ArXiv:2501.19393 [cs].

OpenAI. 2024. OpenAI o1 System Card.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Summer Yue, Alexandr Wang, and Dan Hendrycks. 2025. Humanity's Last Exam.

Qwen. 2025. Qwq-32b: Embracing the power of reinforcement learning.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A Graduate-Level Google-Proof Q&A Benchmark.

Richard Renner. 1995. My Date with the NPR Puzzle.

Joshua Rozner, Christopher Potts, and Kyle Mahowald. 2021. Decrypting Cryptic Crosswords: Semantically Complex Wordplay Puzzles as a Target for NLP.

Graham Todd, Tim Merino, Sam Earle, and Julian Togelius. 2024. Missed connections: Lateral thinking puzzles for large language models. In *2024 IEEE Conference on Games (CoG)*, pages 1–8.

Nemika Tyagi, Mihir Parmar, Mohith Kulkarni, Aswin Rrv, Nisarg Patel, Mutsumi Nakamura, Arindam Mitra, and Chitta Baral. 2024. Step-by-step reasoning to solve grid puzzles: Where do LLMs falter? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19898–19915, Miami, Florida, USA. Association for Computational Linguistics.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Thoughts are all over the place: On the underthinking of o1-like llms. *Preprint*, arXiv:2501.18585.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.

Jingmiao Zhao and Carolyn Jane Anderson. 2023. Solving and Generating NPR Sunday Puzzles with Large Language Models. In *Conference on Computational Creativity (ICCC)*.

Qihuang Zhong, Kang Wang, Ziyang Xu, Juhua Liu, Liang Ding, and Bo Du. 2024. Achieving >97% on GSM8K: Deeply Understanding the Problems Makes LLMs Better Solvers for Math Word Problems. Technical Report arXiv:2404.14963, arXiv.

A Additional Examples

The following challenge is analogous to a one-shot prompt. The listener is given an example of the pattern that they're expected to find in the challenge. We do not remove these examples, and they occur in 2% of our benchmark.

Challenge: The letters of SWITZERLAND can be rearranged to spell LIZARD and NEWTS — LIZARD being the singular name of an animal, and NEWTS a plural. Name another country with this same property. That is, name another country whose letters can be rearranged to spell two animals — one singular and one plural. It's a major country. What country is it?	(9)
Ground Truth Answer: Mexico -> ox, mice	

R1 retracts answers to the following challenge while reasoning.

Challenge: The word NONUNION has four N's and no other consonant. What famous American of the past – first and last names, 8 letters in all – has four instances of the same consonant and no other consonant?	(10)
Ground Truth Answer: Eli Lilly.	

R1 quickly finds the ground truth answer to the following problem, but continues to explore wrong alternatives before returning to the right answer.

Challenge: Think of a well-known category with exactly 7 things in it. Alphabetize the things from their ending letters, and the last letter alphabetically will be E. In other words, no thing in this category ends in a letter after E in the alphabet. It's a category and set of 7 things that everyone knows. What is it?	(11)
Ground Truth Answer: Continents.	

None of the models answered the following challenge correctly. Both R1 and Sonnet 3.7 ET exhibit "underthinking (Wang et al., 2025)," as they repeatedly revisited the target candidate "Lugano" with similar logic but failed to make further reasoning efforts toward identifying the correct person's name that fits the constraint. R1 prematurely deemed it "Not likely" that such a person's name existed, while Sonnet 3.7 ET gave up, stating, "The search space is quite large."

Challenge: Think of a famous person whose name consists of three names. The first and last letters of the first name plus the first and last letters of the second name plus the first and last letters of the third name, in order, name a city and lake in Europe. Who is it?	(12)
Ground Truth Answer: Lulu Garcia-Navarro -> Lugano	

The following puzzle is an example of a crossword puzzle with one-shot example that is challenging for all models. In 8 out of 10 responses from Gemini2 and 6 out of 10 responses from R1, the models stated that the problem cannot be solved within a reasonable time frame and instead requires computational tools. O1, on the other hand, asserted that there is no way to complete the puzzle and suggested that the reader could verify this by using exhaustive 'word square' software.

Challenge: Take the four 4-letter words LIMB, AREA, CORK and KNEE. Write them one under the other, and the four columns will spell four new words: LACK, IRON, MERE and BAKE: LIMB AREA CORK KNEE This is called a double word square. I'd like you to find a double word square with six-letter words. Specifically, your square must include the words PONIES ACCEPT SEARED CAVIAR. These four words must be among the 12 common, uncapitalized six-letter words in the square. Can you do it?	(13)
Ground Truth Answer: ACROSS, CLARET, CAVIAR, EMIGRE, PONIES, TRENDS	

B Category Coding Guidelines

We use the following coding guidelines to categorize the challenges. We assign multiple categories to a challenge when appropriate.

B.1 Letter Categories

- **Change Letters:** Puzzles that involve substituting one letter for another.
- **Add Letters:** Puzzles that involve adding a letter.
- **Remove Letters:** Puzzles that involve removing a letter. This includes puzzles in which two terms partially overlap; they can be made identical by removing letters.
- **Rearrange Letters:** Puzzles that involve rearranging letters. This includes all anagrams.
- **Contains:** Puzzles that involve a string of characters containing another string of characters.
- **Letter Count:** Puzzles that involve a restriction on the number of certain characters. For instance, the puzzle specifies that all vowels are used in the answer; that only two consonants appear; or that the answer contains three consecutive vowels. Does not include puzzles that merely state the total number of letters in a word or in the answer.
- **Alphabet:** Puzzles that require knowledge of the order of letters in the alphabet. Includes puzzles that mention alphabetical order and puzzles that assign a number value to letters based on their position in the alphabet.

B.2 Sound Categories

- **Rhyme:** Puzzles that mention rhyming. Includes both puzzles in which it is stated that words (or parts of words) do and do not rhyme.

- **Sounds Like:** Puzzles that involve knowledge of what strings of letters sound like. This includes puzzles that mention homophones or the phonetics of phrases, as well as puzzles that mention the pronunciation of letters in particular words. It does not include puzzles that merely involve syllables.

B.3 Visual Categories

- **Word Ladders and Crossword-like:** Puzzles that involve writing out strings of letters in a particular configuration or shape. Includes proper crossword grids, where the words must fully interlock, as well as puzzles in which words are written on top of each other in a particular way. Also includes word ladder puzzles, which involve a series of word manipulations from a start to end state. Usually, the rules state that only one letter may be changed in each step, and each step must be a valid English word.

- **Character Properties:** Puzzles that involve knowledge of the visual properties of characters. Includes puzzles that mention rotating characters, mirroring characters, and manipulating the strokes of characters. Also includes puzzles that involve interpreting numeric digits as letters.

B.4 Other Categories

- **Initials and Abbreviations:** Puzzles that involve initials, acronyms, or abbreviations. Includes puzzles in which the initials of two phrases must match; puzzles that involve acronyms or abbreviations of words; and puzzles that involve the symbols of elements on the periodic table.

- **Tricky:** Puzzles that involve a trick. Often, some phrase in the puzzle description must be interpreted in an unusually literal manner or in an unusual way.

- **Mathematical:** Puzzles that involve numbers, mathematical operations, or geometry. This includes puzzles that involve reasoning about combinatorics or probability, involve mathematical notation, require arithmetic, or involve geometric shapes (that are not character symbols).

- **Riddle:** Puzzles that are riddles. Defining a riddle is challenging. However, the language used to describe these puzzles is noticeably different from other puzzles. Usually, wordplay is required but not mentioned overtly.

- **Roman Numerals:** Puzzles that involve knowledge of roman numerals.

- **Synonyms and Antonyms:** Puzzles that state that portions of the answer are synonyms, near-synonyms, or antonyms of each other.

- **Other Wordplay:** Puzzles that involve an element of wordplay and do not fit into any non-knowledge category.

B.5 Knowledge Categories

- **Geography:** Puzzles that require knowledge of geographic locations.

- **Sports and Games:** Puzzles that require knowledge of sports and games, sports teams, or athletes.

- **Pop Culture:** Puzzles that require knowledge of popular culture. Includes movies, television, celebrities, song titles, musicals, or entertainers.

- **Art:** Puzzles that require knowledge of artists, works of art, literature, or classical music.

- **Politics:** Puzzles that require knowledge of politicians or people of political significance.

- **History:** Puzzles that require knowledge of history.

- **General:** Puzzles that require general knowledge that does not belong to any above categories.

C Hardness of Challenges for Humans

The REASONINGWEEKLY benchmark is based on the NPR Sunday Puzzle, which is a segment of the show *Weekend Edition*. The show is broadcast on radio and streamed online and is estimated to reach over three million listeners every week (Bryan, 2008). The show hosts do not consistently report how many solutions they receive. However, when a challenge received approximately 1,500 answers, that was considered notably high (npr, 2009).

In contrast, only four people solved the following challenge–perhaps an all-time low–and all four wrote computer programs to do so in 1995 (Renner, 1995). *This challenge is not part of our benchmark.*

Challenge: Take the two words, palmistry and behind. Together, they contain 15 letters without repetition. What are the two words in the Merriam Webster Tenth Collegiate Dictionary that have the most letters without any repetition?	(14)
Ground Truth Answer: gunpowdery and blacksmith	

While the answer is correct, it is notable that the example in the question is wrong (the two words do repeat the letter 'i'). The listeners were able to ignore this small error and understand the intent of the challenge.

D On-Air Challenges

For the on-air challenges, each puzzle description applies to multiple challenge-answer pairs. The listeners are first presented with the description, followed by a list of questions where they are tasked with determining the answers on-air based on the provided description. The challenges are designed to be solvable in seconds and require less reasoning comparing to the off-air challenges.

We report the performance of GPT-4o and o3-mini on the on-air questions (Zhao and Anderson, 2023). GPT-4o score 65.6%, and o3-mini score 77.7%, with a performance gap of 12% - smaller than the 31% gap observed for the off-air challenges. The significant differences in both the model performances and the relative improvement from model with enhanced reasoning ability suggest that our off-air puzzles require more reasoning.

This is one on-air challenge requiring knowledge on Rearrange Letters and Synonyms. Both o3-mini and GPT-4o answered correctly.

Description: Given a four-letter word and a six-letter word, rearrange the letters of one of them to get a synonym of the other	
Question: peek, retain	(15)
Answer: keep	
Question: clam, serene	
Answer: calm	

Challenge 16 is a similar off-air challenge that require Anagram and Synonym understanding. Compared to the on-air challenge, the off-air challenge does not provide clues regarding the synonyms, and

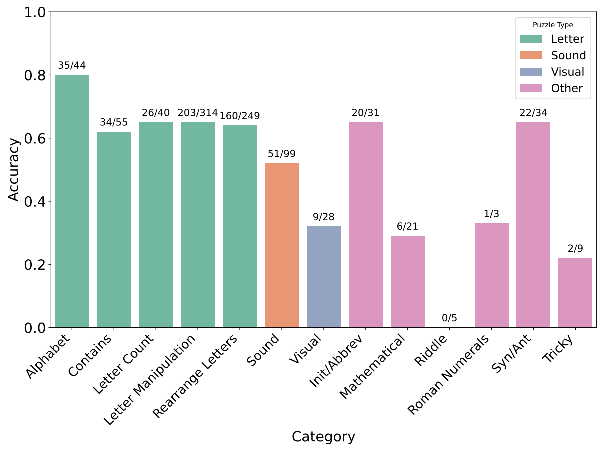


Figure 12: o1's accuracy across puzzle categories

the listeners are tasked to test through U.S. cities and rearrange their letters until a reasonable answer appears. Both o3-mini and GPT-4o failed on this challenge. Out of all models we benchmarked, only o1 and R1 succeeded.

Challenge: Name a major U.S. city in 10 letters. If you have the right one, you can rearrange its letters to get two 5-letter words that are synonyms. What are they?	(16)
Ground Truth Answer: Sacramento -> scent, aroma	

Below is one on-air challenge with wordplay on initials/abbreviation. Both o3-mini and GPT-4o answered correctly.

Description: Every answer today is the name of a famous person whose first initial and last name, in order, spell a word. For example, take Benjamin Rush, signer of the Declaration of Independence. The B of Benjamin + his last name spells BRUSH. I'll give you clues to the parts. You give me the names	(17)
Question: Singer with the group Hole — garlic bulb	
Answer: Courtney Love, clove	

Challenge 18 is a similar off-air challenge that require Initials/abbreviation and Anagram understanding. However, it provides less clues about the person and about the field of renown, thus imposing a larger search space. o3-mini answered correctly, but GPT-4o failed.

Challenge: Name a famous person with the initials M.C. The first initial and last name anagram to the person's field of renown. What is it?	(18)
Ground Truth Answer: Michael Caine -> CINEMA	

E Are Certain Types of Challenges Harder than Others?

We encoded the NPR challenges by the kind of wordplay and knowledge they involve, specified in

Section B. Figure 3 shows the accuracy of the models across puzzle types and across knowledge categories. Figure 3a shows that most models achieve similar performance across knowledge categories. There are no specific knowledge categories that models particular excel or struggles in.

To test whether puzzle types significantly impact the performance of each model, we fit a multivariate multiple regression model. [5] Both o1 and o3-mini-high perform better in the Letter category. In particular, puzzles that involve understanding the order of letters in the alphabet are the easiest for the models o1, o1-mini, o3-mini, o3-mini-high and Gemini-FT. Figure 12 show o1's decline in performance when engaging with Mathematical, Riddle, Tricky puzzles, Visual puzzles including Crosswords and Character Properties, and Sound puzzles on rhyming and phonetic similarity. All other models observe no statistically significant variation in performance across puzzle types.

F Puzzle Solving Strategy as Reasoning Length Increases

We investigate how reasoning models find potential answers or *candidates* for a puzzle. We focus on the thinking output of reasoning models Gemini-FT and R1. We extract potential candidates from the model generation using two different methodologies. The first method is simple regex extraction, where we assume that a candidate is any phrase enclosed within a double quote ("). Note that this method is imprecise, since models do not necessarily provide candidates within double quotes. To improve upon the precision of the first method, we also use a second method involving a transformer-based structured retrieval model, NuExtract-1.5 (Constantin et al., 2024). We finetune the retrieval model on hand-annotated training pairs of generations and candidate labels, taken from the output of another model not in our test set. This step specializes NuExtract-1.5 on our specific task and prompt format. Finally, we use our finetuned retrieval model to extract candidates from R1 and Gemini-FT generations.

In Figure 9, we plot the results of our candidate extraction. We find that beyond a certain length l, models cease to find new candidates for a puzzle. While the value of l varies between our candidate extraction methods, in both cases, the y-axis eventually converges into an asymptote at 100%. The regex-method can be thought of as providing a looser lower bound to the number of candidates. This means that even when given more thinking time, these models will not produce any new candidates. This confirms our previous finding that model accuracy reaches a plateau after n-reasoning tokens. Intuitively, the reason why this happens is because models cease to find new reason ing branches that discover novel candidates.

G Dataset and Code Availability

The code and dataset for this submission is publicly available and licensed under an MIT license. The dataset is derived from NPR Sunday Puzzle challenges and is subject to NPR's copyright and licensing terms.

The code and data for this paper are available at `https://huggingface.co/datasets/nuprl/reasoning-weekly`.

H Use of AI Assistants

Some code for this paper was written with AI assistants enabled.

I Computing Resources

The computational experiments for this paper were conducted with less than 100 hours of A100 GPU time, along with approximately 30 million tokens processed via hosted models.

[5] We determine statistical significance using a p-value threshold of 0.05.

J Checklist

A2 Potential Risks: Did you discuss any potential risks of your work? [Yes/No/NA] **NA**

A2 Elaboration: For yes, provide a section number. For no, justify why not.

B Use Or Create Scientific Artifacts: Did you use or create scientific artifacts? [Yes/No] **Yes**

B1 Cite Creators Of Artifacts: Did you cite the creators of artifacts you used? [Yes/No/NA] **Yes**

B1 Elaboration: For yes, provide a section number. For no, justify why not. **Section 3.**

B2 Discuss The License For Artifacts: Did you discuss the license or terms for use and/or distribution of any artifacts? [Yes/No/NA] **Yes**

B2 Elaboration: For yes, provide a section number. For no, justify why not. **Section G**

B3 Artifact Use Consistent With Intended Use: Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions? [Yes/No/NA] **Yes**

B3 Elaboration: For yes, provide a section number. For no, justify why not. **Ethics Statement**

B4 Data Contains Personally Identifying Info Or Offensive Content: Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it? [Yes/No/NA] **NA**

B4 Elaboration: For yes, provide a section number. For no, justify why not.

B5 Documentation Of Artifacts: Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.? [Yes/No/NA] **Yes**

B5 Elaboration: For yes, provide a section number. For no, justify why not. **Section E.**

B6 Statistics For Data: Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created? [Yes/No/NA] **Yes**

B6 Elaboration: For yes, provide a section number. For no, justify why not. **Section 3**

C Computational Experiments: Did you run computational experiments? [Yes/No/NA] **Yes**

C1 Model Size And Budget: Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used? [Yes/No/NA] **Yes**

C1 Elaboration: For yes, provide a section number. For no, justify why not. **Section I**

C2 Experimental Setup And Hyperparameters: Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values? [Yes/No/NA] **Yes**

C2 Elaboration: For yes, provide a section number. For no, justify why not. **Section 4.**

C3 Descriptive Statistics: Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run? [Yes/No/NA] **Yes**

C3 Elaboration: For yes, provide a section number. For no, justify why not. **Sections 4.2 and E.**

C4 Parameters For Packages: If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used? [Yes/No/NA] **Yes**

C4 Elaboration: For yes, provide a section number. For no, justify why not. **Section 4.**

D Human Subjects Including Annotators: Did you use human annotators (e.g., crowdworkers) or research with human subjects? [Yes/No/NA] **No**

D1 Instructions Given To Participants: Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.? [Yes/No/NA] **NA**

D1 Instructions Given To Participants: Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.? [Yes/No/NA] **NA**

D1 Elaboration: For yes, provide a section number. For no, justify why not.

D2 Recruitment And Payment: Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)? [Yes/No/NA] **NA**

D2 Elaboration: For yes, provide a section number. For no, justify why not.

D3 Data Consent: Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)? [Yes/No/NA] **NA**

D3 Elaboration: For yes, provide a section number. For no, justify why not.

D4 Ethics Review Board Approval: Was the data collection protocol approved (or determined exempt) by an ethics review board? [Yes/No/NA] **NA**

D4 Elaboration: For yes, provide a section number. For no, justify why not.

D5 Characteristics Of Annotators: Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data? [Yes/No/NA] **NA**

D5 Elaboration: For yes, provide a section number. For no, justify why not.

E Ai Assistants In Research Or Writing: Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing? [Yes/No] **Yes**

E1 Information About Use Of AI Assistants: Did you include information about your use of AI assistants? [Yes/No/NA] **Yes**

E1 Elaboration: For yes, provide a section number. For no, justify why not. **Section H.**