AI Explained

Reasoning, Speed and Effort (How AI 'Thinks')

By Dr Peter Fields

THE 2-MINUTE SUMMARY

Some AI models generate intermediate steps before producing a visible answer. A system can also spend extra time comparing candidate solutions or using tools. Researchers call the computational work done when a question is answered **test-time compute**.[1,2,3]

A model selector chooses which model or mode to use. An effort setting, where offered, can influence how much work the selected model does. Higher effort can help on some multi-step problems and can also cost time and computation. The benefit depends on the question and the model.[3,4,5]

A longer wait tells you how long the response took. To establish whether a result was verified, look for the actual calculation, test or source that supports it. A displayed explanation offers clues about the answer, and research has found that explanations can omit influences on a model's decision.[6,7]

Why did the AI pause?

A bat and a ball cost £1.10 together. The bat costs £1 more than the ball. How much is the ball? Ten pence is a tempting answer. Test it and the total becomes £1.20. Five pence works because £1.05 plus £0.05 equals £1.10. This is a British-currency version of a question from Shane Frederick's Cognitive Reflection Test, designed to draw out an intuitive answer that needs checking. A reasoning model may pause before answering. Substituting the prices gives you a check you can do yourself.[9]

What the extra computation can do

A language model generates tokens. A token is a small unit of text, such as a word, part of a word or a punctuation mark. An intermediate step it generates can become part of the material it uses for the next step. In 2022, Wei and colleagues showed that giving sufficiently large models worked examples with intermediate steps improved performance on several tested reasoning tasks. That technique is called **chain-of-thought prompting**.[1]

Some newer models are trained to generate extended intermediate reasoning. DeepSeek's R1 research distinguished an experimental version that learnt through reinforcement learning without an initial supervised fine-tuning stage from the fuller R1 recipe, which used several training stages. Rewards for verifiable outcomes can favour useful behaviour on suitable tasks. How a system learns and how much it computes when asked a question are separate matters.[2]

Geek Box · More than one way to spend the budget

Imagine you have time for two attempts at a difficult problem. You could develop one solution, then revise it. Or you could generate two solutions and choose using a reliable test. Snell and colleagues studied both revision and search in mathematical reasoning. In some settings, allocating extra computation well let a smaller model outperform a much larger one. The choice between candidate answers depends on a reliable test, such as putting each answer back into the original equation.[3]

Choosing a model and choosing effort

Model labels such as 'mini' or 'flagship' describe a provider's offering. They give little basis for comparing model size or estimating reliability across different providers. A more relevant comparison is performance on representative tasks at the time and cost you can afford. OpenAI and Google both document effort controls. The supported settings and how they are applied depend on the model.[4]

Extra effort can be wasted on easy questions. Chen and colleagues documented redundant solution attempts in long-reasoning models, including on simple problems. That is evidence of inefficiency in

the tested systems. For a routine query, accepting more delay makes sense only if the extra work improves something you care about.[5]

> **Geek Box · What does the timer account for?**
>
> Elapsed time can include processing the input, generating intermediate reasoning or waiting for a tool. Some services provide a token breakdown. OpenAI's API, for example, documents reasoning tokens as part of output usage for supported models, even when those tokens are absent from the visible reply. A consumer chat may apply different limits. The timer is a poor substitute for either a usage figure or a record of checks.[4,8]

What a reasoning trace can tell you

A worked solution can help you spot a false assumption or examine an intermediate calculation. Its value comes from whether the steps stand up to checking. Turpin and colleagues found that models could produce plausible explanations for answers influenced by cues that the explanations omitted. In Anthropic's later experiments, when hints influenced answers, Claude 3.7 Sonnet mentioned them in its reasoning 25% of the time and DeepSeek R1 did so 39% of the time, averaged across hint types. Some interfaces display a summary of the reasoning. Use it to find claims you can check.[4,6,7]

A deadline calculation illustrates the distinction. More computation might help the assistant spot that it has counted the first day twice. Determining which deadline applies may require the current rule, the date of service and the relevant exception. Those inputs must come from the record or a source you can check.

How to use it better

- **Specify a testable result.** Ask for the assumptions, calculation or test case you will need to verify the answer.
- **Increase effort when the task warrants it.** For a recurring job, compare settings on the same examples. Include checking and correction in the time you measure.
- **Bring evidence to factual questions.** Check quotations, dates, citations and software outputs through the relevant source or an independent test.

Read more

Hallucinations, Part 3, **What Has Changed Since ChatGPT First Appeared?**, discusses a benchmark on which a newer reasoning model gave more correct answers and also hallucinated more often.

Sources and notes

1. Wei et al. (2022), Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, NeurIPS, pp 24824–24837. Source
2. DeepSeek-AI (2025), DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning, Nature 645, pp 633–638. The journal article distinguishes R1-Zero and R1. Source
3. Snell, Lee, Xu and Kumar (2025), Scaling LLM Test-Time Compute Optimally Can Be More Effective than Scaling Model Parameters, ICLR. First circulated in 2024. Source
4. OpenAI, Reasoning Models; Google, Gemini Thinking. API documentation on effort controls, usage and reasoning summaries. Source 1 · Source 2
5. Chen et al. (2025), Do NOT Think That Much for 2+3=? On the Overthinking of Long Reasoning Models, ICML, PMLR 267. Source
6. Turpin, Michael, Perez and Bowman (2023), Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting, NeurIPS. Source
7. Anthropic (2025), Reasoning Models Don’t Always Say What They Think, 3 April. Experiments with hints and disclosed reasoning. Source
8. OpenAI (2024), Learning to Reason with LLMs, 12 September. Historical developer report about o1. Source
9. Frederick (2005), Cognitive Reflection and Decision Making, Journal of Economic Perspectives 19(4), pp 25–42. Source of the bat-and-ball question, adapted here to British currency. Source