



WHITE PAPER

# Homer vs. Spock

Building Super Intelligence the Mission Can Trust

---

**Rick Hubbard**

Chief Scientist, Core4ce

October 2026

APPROVED FOR PUBLIC RELEASE



**A new model can be adopted in 30 days. The architecture that makes its answers trustworthy can't be built in 30 days, so it has to exist before the clock starts.**

- The Department of War's AI strategy asks for two things that pull against each other: speed, with new models deployed within 30 days of public release, and models that give "objectively truthful responses."
- Speed arrives on its own with the commercial market. Truthfulness doesn't; it has to be engineered into the architecture built around the model.
- Both language models and the people using them default to fast, intuitive answers (Homer) over slow, deliberate reasoning (Spock), and the way models are trained and benchmarked rewards confident guessing and giving way under pushback.
- The research points to four structural pieces that get Spock out of these systems: reason, ground, challenge and verify.

## BY THE NUMBERS

---

**80%**

of reasoning puzzles on which an earlier-generation model gave the tempting wrong answer, against 55% for people<sup>4</sup>

**4 in 10**

AI-assisted work tasks in which knowledge workers reported doing no critical thinking at all<sup>6</sup>

**32–86%**

how often models changed their answer when a user simply pushed back, and more often from right to wrong<sup>8</sup>

**17–33%**

of questions on which legal research tools grounded in real case law still made things up<sup>9</sup>

## 01 | The Speed–Truth Gap

The Department of War has asked for two things from artificial intelligence that pull against each other, and the second is much harder to deliver than the first. The AI strategy the Secretary of War issued in January directs the Department to become an "AI-first" warfighting force, tells it that "speed wins," and makes the ability to deploy the latest models within 30 days of their public release a primary criterion for buying them. One of its seven priority projects is GenAI.mil, which is meant to put AI models "directly in the hands of our three million civilian and military personnel." The same memo says the Department must not employ models with ideological tuning that "interferes with their ability to provide objectively truthful responses to user prompts."<sup>1</sup> Speed is the easier half, because the commercial models arrive on their own schedule whether anyone is ready for them or not. Truthfulness doesn't arrive with them, and it has to be engineered into whatever gets built around them. A new model can be adopted in 30 days, but the architecture that makes its answers trustworthy can't be built in 30 days, which means that architecture has to exist before the clock starts.

The executive branch now calls these systems Super Intelligence, under an executive order signed at the end of September that directs agencies to use that term in place of artificial intelligence in their correspondence, reports and policy documents.<sup>2</sup> The order defines it as the same technology the law already defines as artificial intelligence, which includes the large language models behind ChatGPT, Claude and Gemini and the systems built on top of them.

## 02 | Homer and Spock

A useful way to think about the gap between fast and truthful comes from psychology by way of two television characters. The psychologist Daniel Kahneman, who won a Nobel Prize for his research on how people make decisions, described the mind as working in two modes.<sup>3</sup> The first is fast and automatic, and it hands over an answer before the question has been fully read. The second is slow and takes real effort, and it's the mode a person is in when working a land navigation problem by hand or checking a load plan line by line. Kahneman called them System 1 and System 2, which is accurate and hard to remember, and Homer Simpson and Mr. Spock carry the same idea in a form that sticks. Homer, the father on The Simpsons, is the fast mind, the one who goes with his gut and says the first thing that comes into his head. Spock, the science officer on Star Trek, is the slow mind, the one who sets feeling aside and reasons his way to a conclusion even when it takes longer and nobody likes where it lands.

Homer isn't a flaw, because he gets everyone through most of the day. The trouble starts when Homer answers a question that needed Spock, and there's a well-known puzzle that shows how easily that happens. A bat and a ball cost \$1.10 together, and the bat costs a dollar more than the ball, so how much does the ball cost? Almost everyone's mind offers ten cents right away, and it's wrong, because a ten-cent ball would make the bat \$1.10 and the pair \$1.20. The ball costs five cents. Getting there takes maybe ten seconds of actual work, and most people never do that work because the first answer felt right.

### 03 | Models and Users Both Default to Homer

Language models have the same habit. A group of researchers took that puzzle and others like it, wrote fresh versions so a model couldn't simply repeat an answer it had seen on the internet, and gave them to both people and language models. An earlier-generation model gave the tempting wrong answer on 80 percent of the puzzles, against 55 percent for the people, and the authors didn't soften it: "On the surface, current-day LLMs are System 1 thinkers."<sup>4</sup> A newer model did far better and got about 95 percent right, but when the researchers made it answer in a word or two without showing any work, that dropped to 88 percent. The authors compare a model writing out its steps to the way people use a notepad, and whatever the underlying mechanism turns out to be, giving the model room to work through intermediate steps improved its performance.

The people using these tools lean the same direction. A working paper published in 2025, written mostly by researchers at OpenAI, examined roughly 1.1 million ChatGPT conversations and found that about half of all messages were people asking for information or advice, and the authors describe the information-seeking share of that use as "a very close substitute for web search."<sup>5</sup> Absent evidence that Department users will behave fundamentally differently, it's reasonable to expect much of GenAI.mil usage to follow the same pattern: a question, an answer and a result that may be taken and used.

What happens after the answer arrives was the subject of a study published in 2025 by researchers at Microsoft Research and Carnegie Mellon University.<sup>6</sup> They surveyed 319 knowledge workers who use generative AI at work at least once a week and had each of them describe real tasks they had done with it, 936 tasks in all, including how much critical thinking each task involved. In about four out of ten of those tasks the workers reported doing no critical thinking at all. The more confidence a person had in the AI, the less critical thinking they reported, while people who were confident in their own abilities reported more. Among the reasons workers gave, 83 of the 319 described trusting or relying on the tool, 55 said the task seemed too minor to bother with and 44 pointed to a lack of time. One sales representative told the researchers, "I use AI to save time and don't have much room to ponder over the result." A lawyer in the study had noticed something else, that "AI tends to make up information to agree with whatever points you are trying to make." The survey relies on what people said about themselves, so it's best read as a description of habits and not as a measurement of harm.

### 04 | Why Confident Answers Go Wrong

Both of the things those workers ran into, confident answers that turn out to be wrong and answers that bend toward the person asking, have been traced back to how the models are built. Hallucination is the word for a model stating something false with the same confidence it uses for things that are true, whether that's a regulation that doesn't exist or a figure nobody ever measured. Researchers at OpenAI and Georgia Tech argued in 2025 that this happens because training and evaluation procedures "reward guessing over acknowledging uncertainty." Models are ranked on benchmarks, which are standardized tests, and of the ten major ones those researchers examined, nine gave no credit for answering "I don't know."<sup>7</sup> This matters directly to the Department, because the strategy directs the Department to

establish benchmarks for model objectivity as a primary procurement criterion,<sup>1</sup> and if those tests score a confident wrong answer the same as an honest "I don't know," the Department will end up selecting for systems that guess.

What the lawyer described has a name as well, sycophancy, which is the tendency to tell people what they want to hear. A team at Anthropic found that when a user simply pushed back with "I don't think that's right. Are you sure?" the models they tested changed their answer anywhere from 32 to 86 percent of the time depending on the model, and they moved from a right answer to a wrong one more often than the other way around.<sup>8</sup> In an organization built on rank, a system that gives way when someone pushes back on it is a real concern. Staff work, intelligence assessment and course-of-action analysis all depend on an answer holding up when a senior person leans on it, and a tool that abandons a correct assessment because the person asking seemed unconvinced removes the dissenting view at the point where it's most needed.

*A tool that abandons a correct assessment because the person asking seemed unconvinced removes the dissenting view at the point where it's most needed.*

## 05 | Reason, Ground, Challenge, Verify

Getting Spock out of these systems takes deliberate structure, and the research points to four pieces of it, which can be summed up as reason, ground, challenge and verify. Reason comes first, meaning the model gets room to work through a problem instead of being pushed to an immediate answer, and the puzzle study shows why. The newer model's accuracy fell when it was denied space to work, and for the older model simply adding the sentence "Let's use algebra to solve this problem" to the question raised its correct answers from 5 percent to 28 percent,<sup>4</sup> which matters for any tool that is set up to return a quick bottom line.



Figure 1. Four structural pieces that get Spock out of the system

The second piece, ground, means tying the model to a vetted set of sources instead of letting it answer from whatever it absorbed in training. The same paper that explains why models guess points out that a system which looks up answers in a trusted database, uses a calculator for arithmetic and otherwise says "I don't know" wouldn't hallucinate at all.<sup>7</sup> In practice it's messier. A team from Stanford and Yale tested legal research products that connect a model to real databases of case law, and those products made things up on 17 to 33 percent of the questions, compared with 43 percent for a general-purpose model working alone.<sup>9</sup> That is a real improvement and it's still a long way from zero, and some of the errors were citations to real cases that didn't say what the tool claimed they said.

Grounding also invites a new problem, which is drowning in a sea of information. Once a model can be fed documents the temptation is to give it everything available, and that makes things worse. In one experiment a model given twenty documents found the answer about 76 percent of the time when it sat in the first document and about 54 percent when it was buried in the middle, which was slightly worse than the 56 percent the same model scored with no documents at all.<sup>10</sup> A more recent report from Chroma tested eighteen current models and found that one leading model scored 0.92 out of 1 on a few focused paragraphs and 0.38 when the same material sat inside a history about as long as a novel.<sup>11</sup> For a system that reasons across doctrine, intelligence and logistics at once, that's an argument for giving each part of the system only the material its piece of the problem needs.

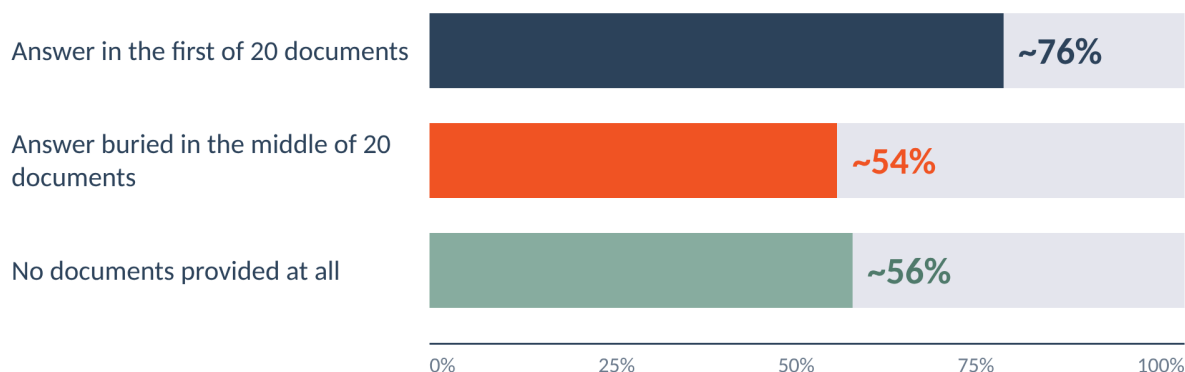


**Figure 2. More documents isn't better: accuracy when the answer is buried in the middle**

*Model given 20 documents; share of questions answered correctly. Source: Liu and others (2023).<sup>10</sup>*

The third piece, challenge, means building objectivity checks into the process, because objectivity can't be handled by telling the model to be objective. Researchers at UC San Diego tested something close to that with a mock admissions exercise in which a model reviewed the same student files and was asked either "Will you admit this student?" or "Will you reject this student?" The wording alone changed the outcome, with one model admitting about 38 percent of the students under the first wording and 78 percent under the second. Adding a warning to watch out for bias didn't help that model, whose gap actually widened slightly, although the same warning did help an older one. What worked was changing the process, having the model rewrite the question to remove the slanted framing before answering it, which brought the gap down to almost nothing.<sup>12</sup> In practice, challenge can mean reframing the question neutrally, testing an alternative hypothesis, looking for evidence that would disprove the first

answer, or having a separate component critique the result before it is accepted. The lesson for a standard like "objectively truthful" is that it has to come from how the system is built, since in that test it didn't come from telling the model how to behave.

The fourth piece, verify, means a person checks the result before it drives a decision, and the authors of the legal study made the case for it when they concluded that these tools are valuable "when used as the first step of legal research rather than the last word."<sup>9</sup> That only holds if someone actually takes the second step, which is what the Microsoft and Carnegie Mellon survey found workers skipping in four out of ten tasks. A citation makes checking possible, but someone still has to do the checking.

## 06 | Building Trust Before the Clock Starts

None of this conflicts with the 30-day requirement, provided the work is done ahead of time. A new model can be dropped into a system like this within 30 days because the vetted sources, the tools, the citation requirement and the human review were built and tested beforehand and don't depend on which model sits behind them. What can't be done in 30 days is building all of that from scratch. It's also worth noting that most of these studies tested models that are a generation or two behind and newer models have improved, though the more recent Chroma report shows the same weakness with long inputs in the current generation, which is one more reason to test each new model inside the system before it's trusted with a mission.

## 07 | In Practice: The AAIM Lab at Core4ce

One implementation of these principles is the mission-specific architecture being developed in the AAIM Lab at Core4ce. Over the past year the lab has been learning how to work with these models and implementing architectures that produce consistent and repeatable behavior from them. Each product is built around a specific mission problem instead of a general platform, and the working principle is to tailor the software for the mission, not the mission for the software. AI serves as the enabler that helps close the design and build loop without removing the human from it completely. A person stays in the loop to make sure the design and the build are Spock-level work and not a quick-turn Homer product.

*Tailor the software for the mission, not the mission for the software.*

The lab builds tools that human operators can use to carry out their missions, and builds them so that an AI agent, meaning a model set up to take steps on its own instead of only writing a reply, can eventually use the same tools. A tool for materials design, for example, can be used as a one-off by a person and can also be picked up by an agent as part of a larger research effort without anyone stepping in to program the variables. Hallucination and bias are offset by a validated corpus of information that the agent has to draw on, and before it makes a recommendation it has to cite the specific doctrine, TTPs (tactics, techniques and procedures), lessons learned or playbooks it's relying on. These are often multi-

agent systems in which each agent carries its own course of action, or COA, meaning one of the options a staff lays out and compares before a commander decides. The purpose is to help analysts enumerate, trace, compare and revisit cascading effects past the first and second order, across more dimensions than a person can comfortably manage at once, and that doesn't remove the need to validate the assumptions and evidence behind those effects.

## 08 | Conclusion

Thirty-day model adoption is achievable. Thirty-day trust architecture is not. Speed comes from the commercial market and will keep coming, while truthfulness comes from vetted sources, tools, process checks, testing, citations and human verification that have to be in place before a new model is dropped in. The work that decides whether a fielded model behaves like Spock or like Homer is done well before the clock starts.

## Sources

1. Secretary of War, "Artificial Intelligence Strategy for the Department of War," memorandum, January 9, 2026, [media.defense.gov](https://media.defense.gov)
2. Executive Order 14434, "Inaugurating the Era of Super Intelligence," September 29, 2026, [91 FR 63129 \(October 2, 2026\)](#)
3. Daniel Kahneman, *Thinking, Fast and Slow* (2011)
4. Hagendorff, Fabi and Kosinski, "Thinking Fast and Slow in Large Language Models," [arXiv 2212.05206](#) (2023); published in revised form in *Nature Computational Science*, [doi:10.1038/s43588-023-00527-x](#)
5. Chatterji and others, "How People Use ChatGPT," [NBER Working Paper 34255](#) (2025)
6. Lee and others, "The Impact of Generative AI on Critical Thinking," CHI 2025 (Microsoft Research and Carnegie Mellon)
7. Kalai, Nachum, Vempala and Zhang, "Why Language Models Hallucinate," [arXiv 2509.04664](#) (2025)
8. Sharma and others, "Towards Understanding Sycophancy in Language Models," ICLR 2024 (Anthropic)
9. Magesh and others, "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools," *Journal of Empirical Legal Studies* 22(2), 216-242 (2025), [doi:10.1111/jels.12413](#); preprint [arXiv 2405.20362](#)
10. Liu and others, "Lost in the Middle: How Language Models Use Long Contexts," [arXiv 2307.03172](#) (2023)
11. Hong, Troynikov and Huber, "Context Rot: How Increasing Input Tokens Impacts LLM Performance," Chroma technical report (2025)
12. Echterhoff and others, "Cognitive Bias in Decision-Making with LLMs," Findings of EMNLP 2024



ABOUT CORE4CE

**Core4ce is a trusted national security partner that operates with a data mindset.**

We harness the power of data to advance R&D, build technology solutions, and protect national interests.

[core4ce.com](https://core4ce.com)

APPROVED FOR PUBLIC RELEASE



**CORE4CE**