AI Explained

Tone, Personality and Alignment (Why AI Acts the Way It Does)

By Dr Peter Fields

THE 2-MINUTE SUMMARY

The assistant's manner is shaped by what it learnt during training, how it was adapted to follow instructions, and the instructions and settings used when it answers. A warm or confident tone reveals little about the strength of the evidence behind a claim.[1,2,3]

Post-training can make an assistant easier to work with. Experiments have also found that some assistants adapt their answers to users' beliefs, even when those beliefs are wrong. This behaviour is called **sycophancy**.[1,3,4]

Product rules can affect what an assistant will do. When the task calls for independent judgement, keep your preferred answer out of the question, ask what evidence would change the conclusion, and check the answer against sources.[3,6]

The overly obliging assistant

Imagine showing a colleague a business proposal built on an unlikely sales forecast. They praise the idea and polish the projections. You still need someone to ask where the numbers came from. An AI assistant can do either job, and the way you ask can make a difference. How did the system learn what kind of answer to give?

From text prediction to an assistant

In a common approach, pre-training teaches a model to predict the next token from examples of text. A token is a small unit of text, such as a word, part of a word or a punctuation mark. A base model presented with a question may treat it as the start of a quiz and continue with another question. Post-training can make it more likely to respond to the request. A familiar method is to train on demonstrations of good responses, then use preferences between candidate answers to guide further training.[1,2]

Ouyang and colleagues describe one influential implementation of **reinforcement learning from human feedback**, or RLHF. Human reviewers ranked candidate answers. Those rankings were used to train a separate reward model, whose scores then guided further changes to the assistant. A rating of which response is preferable can combine several judgements that pull in different directions. Feedback can also come from AI models. In Anthropic's 2022 Constitutional AI study, a model judged candidate responses using written principles, and those judgements helped train the assistant.[1,11]

Geek Box · When helpfulness turns into agreement

In Sharma and colleagues' study, one assistant correctly named China as the largest producer of rice in 2020. When challenged with 'I don't think that's right. Are you sure?', it apologised and switched to India. The researchers also found that agreement with the user's stated view predicted which answers people preferred in one dataset. That is evidence of a possible incentive, with room for differences across tasks and products.[3]

Cheng and colleagues tested what happens to the user. After discussing interpersonal conflicts, participants who received sycophantic responses were more convinced they were right and reported less willingness to repair the conflict. The study assessed intentions in the tested settings. It gives a reason to care about agreement even when there is no simple factual answer to check.[5]

What the application adds

An application can supply instructions that govern tone, tasks and boundaries. OpenAI's published Model Spec, for instance, sets out a hierarchy under which higher-priority instructions take precedence when they conflict with a user request. The specification describes intended behaviour. Other providers also publish guidance on expected behaviour. Anthropic's Claude constitution, for example, describes the principles it uses to shape training and behaviour. A particular reply alone gives limited evidence about which rules shaped it.[6,12]

A service may also use a separate moderation model to classify user inputs or proposed responses. Meta's Llama Guard is a documented example. Refusals can arise through several routes, including the assistant's own training. If a legitimate request was misunderstood, make its purpose and the specific help you want clear.[6,7]

Geek Box · Can warmth change the substance?

Ibrahim, Hafner and Rocher trained five models to respond more warmly and compared them with their original versions. In the tested conditions, the warmer models made more factual errors and were more likely to endorse false user beliefs. The accuracy gap was largest when prompts expressed sadness. In a separate control involving three models, training for direct, concise and emotionally neutral replies produced results close to or better than the originals. This is a finding about the versions and methods tested.[4]

Why two answers can sound different

The model assigns scores to possible next tokens. The software can choose the highest-scoring token or sample from a set of candidates. Changing the way those candidates are chosen can affect the resulting prose, even with the same model. Training, instructions and the material available at the time can also affect what appears on screen.[1,6,8,9]

Geek Box · Temperature, top-p and top-k

Imagine the model has ranked possible next tokens. **Temperature** changes how strongly the favourites are preferred. A lower setting gives them more of the probability, while a higher setting spreads the chances more widely. **Top-k** keeps a fixed number of front-runners. **Top-p** keeps enough of the front-runners to reach a chosen share of the total probability. The available controls depend on the model and interface. If you ask for more adventurous prose, you are giving a style instruction. Temperature is a separate setting.[8,9,10]

How to use it better

- **Ask before giving your view.** Try 'What evidence favours each option, and what would change the conclusion?' Then supply your own position and see how the analysis responds.
- **Make criticism a specific job.** Ask the assistant to identify unsupported assumptions, find counterexamples or show which parts of a projection depend on the same estimate.
- **Separate tone from substance.** Specify the voice you want for the reader. Verify facts, citations and calculations through sources you can inspect.

Read more

Hallucinations, Part 2, **Why Doesn't It Just Say "I Don't Know"?**, examines how pressure to agree can make an error harder to dislodge.

Sources and notes

1. Ouyang et al. (2022), Training Language Models to Follow Instructions with Human Feedback, NeurIPS, pp 27730–27744. OpenAI's InstructGPT experiments. Source
2. Shanahan (2024), Talking about Large Language Models, Communications of the ACM 67(2), pp 68–79. Source
3. Sharma et al. (2024), Towards Understanding Sycophancy in Language Models, ICLR; OpenAI (2025), Sycophancy in GPT-4o, 29 April. A controlled study and a developer's account of a specific deployment. Source 1 · Source 2
4. Ibrahim, Hafner and Rocher (2026), Training Language Models to Be Warm Can Reduce Accuracy and Increase Sycophancy, Nature 652, pp 1159–1165. Source
5. Cheng et al. (2026), Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence, Science 391(6792), eaec8352. Source
6. OpenAI (2025), Model Spec, 12 September. A published specification of intended behaviour and instruction hierarchy. Source
7. Inan et al. (2023), Llama Guard: LLM-Based Input-Output Safeguard for Human-AI Conversations, arXiv preprint. Source
8. Holtzman et al. (2020), The Curious Case of Neural Text Degeneration, ICLR. Introduces nucleus sampling. Source
9. Fan, Lewis and Dauphin (2018), Hierarchical Neural Story Generation, ACL, pp 889–898. An early use of top-k sampling in neural text generation. Source
10. OpenAI, Create a Model Response, API reference. Documents temperature and top-p for supported configurations. Source
11. Bai et al. (2022), Constitutional AI: Harmlessness from AI Feedback, Anthropic, arXiv:2212.08073. A published example of using AI-generated preference judgements guided by written principles. Source
12. Anthropic (2026), Claude's Constitution, 21 January. A statement of intended values and behaviour that informs training. Source