Don't Teach Your Agents With Hindsight

Autonomy Platform · Public Blog

CONFIDENTIAL

Technical July 2026

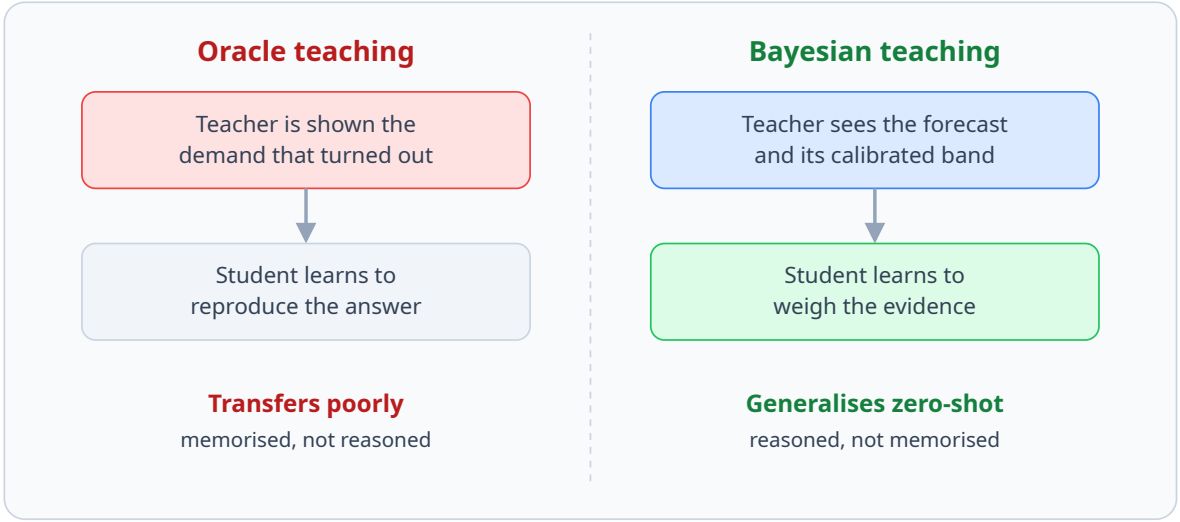
Don't Teach Your Agents With Hindsight

Google Research recently published evidence that an AI taught to reason under uncertainty generalises better than one taught from perfect-hindsight answers. It is a clean statement of a principle we built our decision agents on: teach the process under the calibrated band, not the answer you only know in retrospect.

In March 2026, Google Research published a result with a modest setup and a large implication. They put a language model in a five-round recommendation task, where the right answer depended on a preference the model could not see and had to infer from feedback. Then they trained it two ways. One student learned from an *oracle*: a teacher that already knew the hidden answer and always produced the perfect choice. The other learned from a *Bayesian teacher*: one that did not know the answer either, but maintained a distribution over the possibilities, updated it as evidence arrived, and made its best calibrated guess each round.

The Bayesian-taught student generalised better. It agreed with the ideal reasoner more often, and the skill it acquired transferred, zero-shot, to domains it had never trained on, while the oracle-taught student, handed

perfect answers, learned mostly to reproduce answers. One learned an answer; the other learned to reason.



Same task, two teachers: the uncertain teacher produced the student that generalised. After Google Research, "Teaching LLMs to reason like Bayesians" (2026).

Why this is not a language-model story to us

It would be easy to file this under “LLMs are getting better at reasoning” and move on. That is not why it matters here. At Autonomy the language model never makes the decision. Our decisions come from graph and per-decision models trained against a digital twin, with a conformal layer carrying the calibrated uncertainty and the LLM confined to narrating what those agents already decided. So the headline, “teach an LLM to reason like a Bayesian”, is not a threat to that architecture and not something we adopt into it.

What the result validates is the *training method*, and that we do use, on the decision tier. Our agents are not programmed with planning rules. They are taught by a teacher, the way a chess engine learns by watching stronger play, and then they surpass the teacher by finding structure the teacher's rules cannot express. The open question in any teacher-student setup is exactly the one Google isolated: what does the teacher know when it produces the label the student imitates?

The oracle trap in supply chain

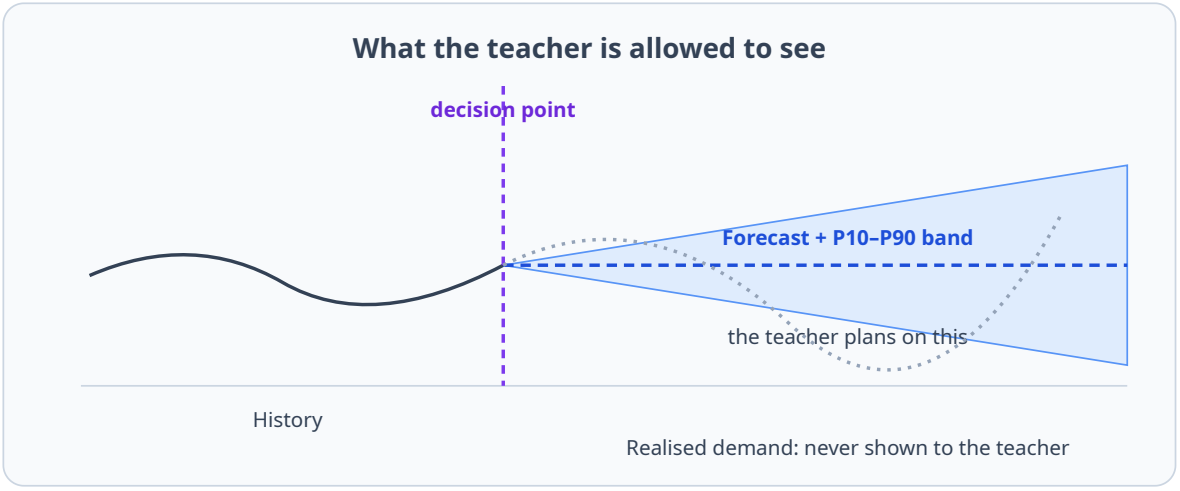
There is a seductive shortcut when you build training data for a planning agent. You have the historical demand. You know what actually happened. So you let the teacher solve each scenario with that realised demand in hand and call the result the "optimal" plan. It looks rigorous. It is a trap.

A teacher that plans with the actual outcome in hand is a clairvoyant. The plan it produces is unbeatable and unreachable, because no live agent will ever know next week's demand when it places this week's order. Train a student to imitate that plan and you teach it to reproduce an answer it can only ever approximate by memorising, not to reason its way there under the fog every real decision is made in. It scores beautifully on the data it trained on and degrades the moment conditions shift, which in a supply chain is always.

What we teach instead

Our teachers are built to solve on the decision-time information set: the forecast of record and its calibrated confidence band, the P10 to P90 range

the conformal layer produces, and nothing the live agent would not also have. A correctly built teacher never sees the realised outcome. It reasons under the same uncertainty the deployed agent will face, sizes its buffers and its allocations against that uncertainty, and the student learns that process rather than a hindsight-perfect answer.



The teacher solves on the same information the live agent will have, and no more.

That is why the calibrated band is not a reporting feature bolted on at the end. It is an input to how every agent is trained. A forecast that carries a median but no honest band is, in our terms, accuracy-grade and not decision-grade, because it cannot teach an agent to behave differently when the future is wide open versus when it is tight. The band is what lets the agent modulate.

Google’s result is the cleanest external statement of the principle I have seen: reasoning under maintained uncertainty is what generalises, and hindsight-perfect supervision is what fails to. We arrived at it from the supply-chain side, because a planning agent that decides what to order and what to move has to be correct for a reason that holds up when the

world moves. It is good to see a major lab arrive at the same place from the other direction.

Reference: Sjoerd van Steenkiste and Tal Linzen, "Teaching LLMs to reason like Bayesians", Google Research, March 2026.