

Predicting Your Own Failure Is Not Predicting Another Model's Advantage: What Router Results Establish About Agent Self-Knowledge, and the Incremental-Validity Test Nobody Has Run

Mahendrakar, Pranay

SONYTECH

ORCID: 0009-0003-7224-029X

Life of Research

2026-09-04

Pranay Mahendrakar

Abstract

A common reading of the LLM routing literature holds that delegation works without self-knowledge: a router allocates queries using external features of the task and observed outcome statistics over a model pool, and never consults the candidate model's opinion of its own competence. This paper argues that the reading does not survive the evidence. Signals derived from the candidate model's own generation are competitive with trained external routers and, in the one located comparison that varies distribution shift, substantially better out of distribution; a self-confidence gate has been reported to beat a frozen external router at lower token cost; and the strongest current cascade and escalation systems are built on exactly such signals. What the evidence does support is a different partition, which the internal-versus-external frame obscures. Predicting whether you will fail is a self-directed prediction and is comparatively tractable. Predicting which alternative would do better is an other-directed prediction, and the same score that ranks own-failure well has been reported to fall sharply when asked which collaboration protocol pays off. Delegation is the second kind of judgement, not the first, and the learning-to-defer literature has known for years that a deferral rule must model the expert rather than only the deferrer. A further problem is that no located routing experiment isolates introspection at all. Every deployed self-signal is a supervised probe fitted to labelled outcomes, which is an external estimator that happens to read internal features, so the axis the delegation story rests on is not the axis the experiments vary. The control that would settle it already exists in the introspection literature, where a model's self-prediction is scored against what a second model with matched knowledge predicts about it. That control has not been run on a router. Four measurements are set out that would settle the open part, and what is not known is stated flatly: nobody has reported whether self-assessed capability adds anything to a delegation decision once an external estimator with the same information is already in the system.

1. Introduction

An agent that can call another model has to decide when to do so. The decision is cheap to describe and hard to make well: keep the task and risk failing at it, or hand it over and pay for a capability you may not have needed. Systems that make this decision automatically are now standard infrastructure. RouteLLM trains routers on human preference data and reports cost reductions of over two times in some settings without compromising response quality, along with transfer when the strong and weak models are swapped at test time (Ong et al., 2024). FrugalGPT frames the same problem as a cascade and reports matching the best individual model at up to 98 percent lower cost, or improving on it by 4 percent at equal cost (L. Chen et al., 2023). A 2026 survey collects the routing and cascading families

together (Moslem and Kelleher, 2026).

The natural story to tell about these systems is that they show delegation does not require self-knowledge. On this reading, the router is a third party. It looks at the query, consults statistics about which model has historically succeeded on queries like it, and dispatches. The candidate model is never asked what it thinks of its own chances, and the system works anyway. If that is right, it has a consequence for a large body of work on agent metacognition: the self-assessment machinery is decoration on a decision that an external estimator already makes better.

This paper argues that the reading is wrong on its own terms, and that correcting it exposes a sharper distinction that the routing literature has already measured but not named.

The first half of the correction is empirical. Self-derived signals are not absent from routing; they are among the strongest things in it. Nguyen (2026) takes supervised baselines in the style of RouteLLM as the reference and tests zero-shot confidence signals against them on two datasets, with models drawn from three 7-8B model families, and reports that average token log-probability, which requires no training data at all, matches or exceeds the supervised baselines in distribution (AUROC 0.650 to 0.714 against 0.644 to 0.676) and substantially outperforms them out of distribution (0.717 to 0.833 against 0.512 to 0.564). The stated reason is that the log-probability measures a property of the model's generation while the supervised router measures a property of the query distribution. Yang et al. (2026) report that a pre-answer self-confidence gate reaches 78.0 percent solve at 45K tokens where a frozen gpt-oss-120b router reaches 73.8 percent at 71.3K. RouterXBench routes on internal hidden states that capture model uncertainty before answer generation, rather than on output probabilities or external embeddings, and reports 16.68 percent and 18.86 percent relative improvements over the best baselines on two of its dimensions (W. Wu et al., 2026). The strongest step-level escalation systems in print are built on residual failure risk estimated from the acting model's own trajectory (Hemadri et al., 2026; Shaikh, 2026).

The second half is conceptual, and it is where the contribution lies. The routing literature contains one clean dissociation, and it does not run along the internal-external axis. Yang et al. (2026) hold the solver fixed and run every problem under four collaboration protocols. A post-answer probe ranks baseline failures at 0.8847 AUROC. The same score remains informative for whether any collaboration helps, at 0.7683 AUPRC. It falls to 0.1674 and 0.1041 AUPRC for identifying value specific to the planner-executor-reviewer and broadcast protocols. Their own summary is that confidence can support initial escalation while protocol-specific cost-aware routing remains unresolved.

Read as a claim about introspection, that result says nothing. Read as a claim about the target of prediction, it says a great deal. Will I fail is a question about the predictor itself. Which of these alternatives would do better is a question about parties the predictor is not. The two questions differ in what information could possibly answer them, and only the second is the delegation decision. A system that answers the first and acts as though it has answered the second is making an inference it has no warrant for.

That distinction is not new to machine learning. It is the founding observation of the learning-to-defer literature, which generalised rejection learning precisely by considering the effect of other agents in the decision process (Madras et al., 2018), and which learns its rejector from samples of the expert's decisions rather than from the classifier's confidence alone (Mozannar and Sontag, 2020). The routing literature has rediscovered the same structure without importing the framework, and one consequence is that its most quoted result is routinely over-read.

A third problem sits underneath both halves, and it is the reason the empirical question is still open. No located routing experiment isolates introspection. Every self-signal now deployed is a supervised probe: a classifier over hidden states, a calibrated risk head, a competence posterior whose sufficient statistics

are learned from labelled trajectories. That is an external estimator reading internal features, which is a different thing from the model knowing something about itself that an observer could not learn. The introspection literature has a control for exactly this, and it is simple: score the model's self-prediction against what a second model with matched knowledge of it predicts (Binder et al., 2024; S. Song et al., 2025). No located router paper runs it.

Sections 3 through 7 develop these three points against the published record. Section 8 sets out four measurements that would settle the open part. Section 9 states flatly what is not known.

2. Scope, method, and what this paper is not

This is an analysis, with a research-agenda section. It reports no experiments, no benchmark results and no original measurements, and every number in it is attributed to the source that measured it. Where a number appears without attribution it is an error.

The evidence base was built by searching arXiv and Crossref for work on LLM routing, cascading, escalation, deferral, self-assessment, calibration and machine introspection, and by following citations out of the anchor set. Every reference was resolved through the arXiv API or Crossref before inclusion, and every claim attached to a reference was read back against that source's own abstract. Claims that could not be substantiated at that level were deleted rather than softened. Several numbers reported below are read from abstracts that do not supply the accompanying base rates or confidence intervals, and where that matters the text says so.

Three scope limits are worth stating before the argument rather than after.

First, the paper takes no position on whether language models are conscious, have privileged access in any philosophically loaded sense, or introspect in the way the word is used of humans. Comsa and Shanahan (2025) argue that some apparent self-reports are not introspection at all while at least one, a model correctly inferring its own temperature parameter, is a minimal legitimate example, and that distinction is enough for the present purpose. The operative question here is narrower and entirely empirical: does a signal that originates in the candidate model's own processing carry information about the delegation decision that an external estimator with the same observable inputs does not already have?

Second, the paper does not argue that routers are bad or that self-assessment is good. Both halves of that framing are what it is trying to dismantle. The claim is about what the published comparisons license, and the recurring finding is that they license less than they are read as licensing, in both directions.

Third, the multi-agent material is deliberately confined to protocol and teammate selection, because that is where the self-directed and other-directed predictions come apart most visibly. The wider question of why multi-agent systems fail is treated elsewhere at length, including a taxonomy of fourteen failure modes across 1600 or more annotated traces from seven frameworks (Cemri et al., 2025), and is not reopened here.

The paper continues a line of argument in three earlier records without repeating them. The Self-Verification Gap argued that the hallucination detectors that work are external ones, and asked what real-time correction would require (Mahendrakar, 2026a). A Second Model Is Not a Second Opinion argued that peer verification inherits an independence assumption nobody reports (Mahendrakar, 2026b). An Unknown Unknown Is Defined Relative to an Oracle located four places where external judgement enters blind-spot discovery (Mahendrakar, 2026c). The present paper is about a different decision from all three. Those papers ask whether a system can detect that something is wrong. This one asks whether a system can estimate that a different system would do better, which is a comparative

judgement rather than a detection, and the distinction is load-bearing for everything in Section 5.

3. Three signals that all get called routing

The word router covers at least three different estimators, which condition on different information and fail in different ways. Separating them is a precondition for asking the incremental-validity question at all, because two of the three are external in the relevant sense and one is not.

The first family estimates difficulty from the query. Semantic features, embeddings and handcrafted attributes go in, and a predicted suitability or difficulty comes out. VDAR-Router's stated motivation is that existing methods estimate model suitability from surface semantics or embedding similarity and may therefore ignore the underlying difficulty of a query, which it addresses by generating an explicit difficulty analysis and retrieving against it (Tang et al., 2026). RouterArena builds distinguishable difficulty levels into each domain of its dataset by construction (Lu et al., 2025). BEST-Route conditions both the model choice and the number of samples to draw on query difficulty and a quality threshold, reporting cost reductions of up to 60 percent with less than 1 percent performance drop (Ding et al., 2025).

This family is external in the strict sense: it never touches the candidate model. It also carries an assumption that is rarely made explicit, namely that difficulty is a property of the query rather than of the query-model pair. IRT-Router is the located exception that takes the assumption seriously, importing item response theory in order to model the relationship between LLM capabilities and query attributes as separate quantities, and reporting interpretable estimates of both model ability and query difficulty across 20 LLMs and 12 datasets (W. Song et al., 2025). The separation matters for Section 8, because a router that estimates one scalar per query cannot in principle express that a query is easy for one model and hard for another, which is the only condition under which routing has anything to allocate.

The second family estimates capability from observed outcomes over a pool. Here the router's knowledge is a table of what happened, not a theory of the query. Agent-as-a-Router names the bottleneck directly as information deficit, on the evidence that a vanilla LLM router improves by 15.3 percent in relative terms when the only change is the addition of performance statistics at the task-dimension level, with the augmented router also outperforming a heuristic router that draws on the same dimension-level priors; the framework then accumulates execution-grounded experience during deployment and reports the lowest cumulative regret in distribution with generalisation to out-of-distribution agentic programming tasks (P. Zhou et al., 2026). EMRC builds an explicit expertise table from a public corpus, indexed by medical department category and query difficulty level, and uses it to select agents at inference time (Bao et al., 2025).

DARS is the sharpest result in this family and it is a negative one about the family's own supervision. Existing routing methods typically treat a model's single response to a query as its capability label; because generation is stochastic, that is a noisy observation of a query-model pair rather than a reliable capability estimate, and the reported consequence is systematic noise in routing supervision and less reliable learned policies (Lai et al., 2026). Single-shot labels can be misleading for model selection, and the paper's own conclusion is that reliable routing should be grounded in query-level model capability distributions rather than single-response observations. This is worth holding onto, because the external estimator's advantage is usually asserted on the grounds that it has ground truth. Sometimes it has one sample.

The third family estimates from the candidate model's own processing. Log-probabilities, hidden states, verbalised confidence and post-answer self-scores all belong here. LLM Router routes on the residual stream, using what it calls Encoder-Target Decoupling to separate the model producing the predictive

signal from the model whose correctness is being estimated, which allows open-weight encoders to predict the performance of closed-source targets; at its best it closes 45.58 percent of the gap between the strongest standalone model and the oracle with 74.31 percent cost savings against the most expensive model (Varshney et al., 2026). RouterXBench aggregates cross-layer hidden states through learnable Dirichlet distributions and reports robustness across model families, scales, heterogeneous tasks and agentic workflows (W. Wu et al., 2026). Self-Route extracts capability-aware embeddings from hidden layer representations in a lightweight pre-inference stage to switch between general and reasoning modes (Y. He et al., 2025). SR-RAG treats the LLM itself as a first-class knowledge source and combines LLM-based uncertainty with an external policy datastore to improve decision calibration, reporting gains of 8.5, 2.1 and 4.7 percent over a strong selective-retrieval baseline while performing 26, 40 and 21 percent fewer retrievals (D. Wu et al., 2025).

Two observations about this third family are due immediately, and Section 6 develops both. It is the family most often described as the model knowing something about itself, and it is also the family in which the predictive apparatus is uniformly trained on external labels. LLM Router's own headline design choice is that the encoder need not be the target. That is an explicit architectural denial that self-access is required for a self-derived signal to work.

4. The premise does not survive

The claim to be tested is that routing works because of external estimation rather than the candidate's model of its own competence. Four located results cut against it, and none of them is a small effect.

Nguyen (2026) is the most direct, because it is the only located study that runs the comparison as its main experiment and varies distribution shift. Average token log-probability, a signal requiring no training data whatsoever, matches or exceeds RouteLLM-style supervised baselines in distribution, with AUROC 0.650 to 0.714 against 0.644 to 0.676. Out of distribution the gap widens sharply: 0.717 to 0.833 against 0.512 to 0.564, a range whose lower end sits at the 0.5 chance level and whose upper end is a weak discriminator. A supervised baseline trained on 1,000 labelled examples is reported never to exceed the zero-shot signal. The paper also proposes retrieval-conditional self-assessment, a pre-generation signal that improves on bare self-assessment by up to 0.069 AUROC at three to ten times lower latency than log-probability.

The scope of this result should be stated with some care, because it is doing a lot of work in the argument. It covers three 7-8B model families and two datasets, at 1,000 and 500 queries per model respectively, and it is a single-author preprint with no located replication. The AUROC values in distribution are in the 0.65 to 0.71 range, which is a usable but not a strong discriminator. What the result establishes is not that self-signals are good in absolute terms. It is that on this evidence they are not worse than the trained external alternative, and that the external alternative degrades faster under shift, for a reason the paper states mechanistically: the log-probability measures a property of the generation, and the supervised router measures a property of the query distribution, so only one of them has anything left when the query distribution changes.

Yang et al. (2026) supply the second result, on a much larger and better instrumented benchmark: 4,181 competition-level mathematics problems as the primary set, with paired robustness checks across four benchmarks spanning competition mathematics, biology and broader science with two solver families. A pre-answer self-confidence gate reaches 78.0 percent solve at 45K tokens. A frozen gpt-oss-120b router reaches 73.8 percent at 71.3K. The self-signal is better and costs roughly 37 percent fewer tokens. A retrospective fixed-order oracle reaches 92.4 percent, so both are far from the ceiling, and the paper reports oracle gaps of 18.5 to 28.9 points in the six settings with held-out router evaluations. The comparison is one gate against one frozen router in one family of settings, and it should not be generalised past that. It is nonetheless a direct counterexample to the claim that the

external estimator is the thing that works.

The third result is architectural rather than numerical. Two step-level escalation systems that report strong reliability-cost frontiers do not use query-side routing at all, and one of them says explicitly that query-side routing is the wrong instrument. R2V-Agent's motivation is that existing cascades usually commit to a route for the whole query before execution begins, but "task difficulty shifts mid-trajectory - after flaky tool calls, truncated observations, or compounding local errors - making pre-execution routing brittle"; its router estimates residual failure risk at each step on a fixed policy, trained with Brier-calibrated probability estimation under a CVaR-constrained objective, and it reports 94.3 percent HumanEval+ success at 0.60 percent escalation, TextWorld recovery from 64.6 to 98.2 percent at 41.7 percent escalation, and 93.3 percent TerminalBench success at 33.9 percent LLM calls, roughly half the heuristic router's cost (Hemadri et al., 2026). Shaikh (2026) formalises the same regime as intra-generation delegation, derives a myopic escalation threshold in closed form, characterises the optimal policy as a time-varying threshold, and proves a regret bound governed by the calibration of the competence posterior together with a finite-sample guarantee decaying as one over the square root of the number of calibration trajectories. The real-model validation is a Qwen2.5-Coder 1.5B to 7B cascade on 257 MBPP tasks, and two of its three pre-registered predictions hold there: at equal cost the escalation frontier lies above post-hoc routing, and the running competence belief separates successes from failures more sharply as generation proceeds. The abstract does not say which prediction failed, and the validation is small, so the theory should be read as supported in simulation and partially corroborated on one model pair.

The fourth result is that the external side is weaker than its reputation. LLMRouterBench re-evaluates the field over 400K instances from 21 datasets and 33 models with 10 representative baselines under unified conditions. It confirms model complementarity, which is routing's central premise. It also reports that many routing methods perform similarly under unified evaluation, that several recent approaches including commercial routers fail to reliably outperform a simple baseline, and that a substantial gap to the oracle remains, driven primarily by persistent model-recall failures (H. Li et al., 2026). Backbone embedding models have limited impact and larger ensembles show diminishing returns against careful model curation. RouterArena exists because the field needed a standardised comparison at all (Lu et al., 2025). A literature whose own unified benchmark reports that commercial entrants do not reliably beat a simple baseline is not in a position to claim that external estimation is the demonstrated mechanism.

Two qualifications keep this section honest. Beating a supervised router does not establish introspection; Section 6 argues it does not even bear on it. And the premise being tested is not universally held in the literature. The delegation-without-self-knowledge reading is one available interpretation of routing results rather than a thesis defended in a specific paper, and it is tested here because it is doing real work in how the routing and metacognition literatures are kept apart.

5. The split the evidence does support

If the internal-external axis is not what the routing results establish, what do they establish? The cleanest answer available is in Yang et al. (2026), and it is a dissociation between two prediction targets.

Their design holds the solver fixed and runs every problem under four protocols: direct solving, iterative self-correction, planner-executor-reviewer collaboration, and multi-agent deliberation. Because the solver is fixed, a predictor that ranks well on one target and badly on another cannot be excused by a change of subject. The reported numbers are: 0.8847 AUROC for ranking baseline failures, over 4,151 parseable cases with a 95 percent confidence interval of 0.8732 to 0.8955; 0.7683 AUPRC for whether any collaboration helps; and 0.1674 and 0.1041 AUPRC for identifying value specific to the planner-

executor-reviewer and broadcast protocols.

The last two numbers require care and the paper's abstract does not supply what would be needed to interpret them fully. AUPRC is not comparable across targets with different positive base rates, and a value of 0.1041 could be at, above or below chance depending on how often broadcast-specific value actually occurs. So the correct reading of these numbers is not that protocol-specific prediction is at chance. It is that the same score, on the same problems, from the same probe, falls from the any-collaboration target to the protocol-specific ones by factors of roughly five and seven, and that the authors themselves conclude that protocol-specific cost-aware routing remains unresolved while confidence can support initial escalation. The internal comparison is what carries the point; the absolute values do not.

What separates the targets is not where the signal comes from. It is who the prediction is about.

Ranking baseline failures is a self-directed prediction. The predictor is asked about a process it has access to, whose inputs it saw and whose output it produced. Kadavath et al. (2022) established the shape of this capability early: larger models are well calibrated on multiple-choice and true-false questions in the right format, $P(\text{True})$ self-evaluation performs and scales encouragingly, and models can be trained to predict $P(\text{IK})$, the probability that they know the answer, without reference to a proposed answer. The limitation they report is also the relevant one: $P(\text{IK})$ only partially generalises across tasks, and models struggle with its calibration on new tasks. Cheng et al. (2024) obtain refusal of most unknown questions after aligning an assistant to a model-specific I-do-not-know dataset, and the qualifier model-specific is doing work, because the known-unknown partition was established by testing the assistant externally, not by asking it.

Predicting which protocol pays off is an other-directed prediction. It requires a model of what a planner-executor-reviewer arrangement would do with this problem, which is a question about a process that has not run, involving parties whose behaviour the predictor has no access to. Introspective access to one's own states, however good, does not by itself supply that counterfactual. It could still contribute to it, if combined with learned knowledge about the alternatives, and nothing in the located record rules that out; what is ruled out is deriving the counterfactual from self-access alone. The information that most directly answers it is outcome history over the alternatives, which is what the second router family in Section 3 collects, and which Agent-as-a-Router shows is the binding constraint by demonstrating a 15.3 percent relative gain from adding task-dimension performance statistics to an otherwise unchanged LLM router (P. Zhou et al., 2026).

The learning-to-defer literature reached this structure years ago and stated it formally. Madras et al. (2018) introduced deferral precisely as a generalisation of rejection learning that considers the effect of other agents in the decision process, and built an algorithm that accounts for biases held by the external decision-maker. Mozannar and Sontag (2020) learn a classifier and a rejector jointly from samples of the expert's decisions, with a consistent surrogate loss. Mao et al. (2024) extend this to multiple experts with H-consistency guarantees. Montreuil et al. (2026) take up the case in which the deferral system also chooses what information the expert receives. For a broad family of natural separated surrogates, in which routing and advice are assigned to distinct heads, they show that consistency already fails in the smallest non-trivial setting, and they prove consistency for an augmented surrogate over the composite expert-advice action space.

The transferable point is that a deferral rule is a function of the expert, not only of the deferrer, and that decomposing the decision into independent pieces can be provably wrong rather than merely suboptimal. A routing system built on a self-confidence threshold has decomposed the decision in exactly the way that result warns about: it estimates its own failure probability and treats escalation as automatic above the threshold, which is a separated policy whose second half is a constant function of

the alternative. Whether the specific inconsistency proved by Montreuil et al. applies to the LLM routing setting is not established by their result, which concerns a different action space, and this paper does not claim that it does. What transfers is the warning about the decomposition, not the theorem.

HiL-Bench measures the other-directed problem in its hardest form, where the alternative party is a human. Its Ask-F1 metric is the harmonic mean of question precision and blocker recall, which prices over-asking and silent guessing against each other, and its reported finding is a large universal judgement gap: "no frontier model recovers more than a fraction of its full-information performance when deciding whether to ask" (Trinh et al., 2026). The failure analysis names three patterns, one of which is overconfident wrong belief with no gap detection and another of which is high uncertainty detection with persistent errors anyway. Those two are different failures. The first is a self-directed failure. The second is an agent that has correctly answered the self-directed question and still cannot convert it into the right other-directed action. That the benchmark separates them is more useful than any of its aggregate numbers. Notably, the paper also reports that RL training on a shaped Ask-F1 reward improves both help-seeking quality and task pass rate for a 32B model with gains transferring across domains, which indicates the gap is a trainable deficit rather than a hard limit.

6. Why internal versus external is the wrong axis

The argument so far has treated self-derived signals as a coherent category. It is time to say why the category does not carry the weight the delegation debate puts on it.

Consider what is actually deployed. RouterXBench's ProbeDirichlet is a lightweight router trained on multi-domain data that aggregates cross-layer hidden states through learnable Dirichlet distributions with probabilistic training (W. Wu et al., 2026). R2V-Agent's router is trained on a fixed policy's residual failures with Brier-calibrated probability estimation (Hemadri et al., 2026). Shaikh's competence posterior has sufficient statistics learned from labelled trajectories, and the paper is explicit that they are not read off raw entropy (Shaikh, 2026). Self-Route trains on capability-aware embeddings extracted from hidden layers (Y. He et al., 2025).

Every one of these is a supervised estimator fitted to observed outcomes. The features happen to be internal to the acting model; the estimator is not. Nobody asked the model anything. A probe over activations is an instrument an external party built and calibrated, and it stands in the same relation to the model as a thermometer does to a patient. Calling its output the model's self-assessment is a description of where the sensor is placed, not of who is doing the assessing.

LLM Router makes the point unavoidable. Its central design choice, Encoder-Target Decoupling, separates the model that produces the predictive signal from the model whose correctness is being estimated, so that open-weight encoders can predict the performance of closed-source targets, and this is reported to outperform semantic baselines (Varshney et al., 2026). A signal read from internal states is therefore not even necessarily read from the states of the model being predicted. The internal-external distinction, applied to routing signals, does not track a distinction between self-knowledge and observation.

Three consequences follow.

The first is that the empirical results in Section 4 do not license conclusions about introspection, in either direction. Nguyen's log-probability beating a supervised router shows that generation-side features generalise better under shift than query-side features. It does not show that the model knows anything. Equally, the success of an external router shows nothing about whether asking the model would have helped, because in none of these papers was the model asked.

The second is that the one signal in the routing literature that genuinely is a self-report, verbalised

confidence, is also the one with the worst measured properties, which quietly biases the comparison. Verbalised confidence is systematically overconfident, and the mechanisms are now partly characterised. Seo et al. (2026) identify answer-independence, "the failure to condition confidence on the model's own answer", as a primary driver. V. Wang and Stengel-Eskin (2026) attribute much of it to suggestibility, validate empirically that suggestibility is higher on lower-accuracy claims, and correct for it with distractor normalisation that at 10 inference calls outperforms self-consistency at 100. T. Zhao et al. (2026) locate a compact set of MLP blocks and attention heads in middle-to-late layers that write the confidence-inflation signal at the final token position, and report that targeted inference-time intervention on those circuits substantially improves calibration. Leng et al. (2024) trace part of the problem upstream to RLHF, reporting that reward models used for PPO exhibit inherent biases towards high-confidence scores regardless of actual response quality.

Read together, these say the verbal channel is a poor readout, and they say nothing about whether the underlying state is informative. That is a claim about an instrument, not about a capacity, and treating a bad instrument's failure as evidence against the capacity is the error the human metacognition literature built its measurement apparatus to avoid (Fleming and Lau, 2014). Park et al. (2026) use an instrument of that kind, establishing a measurement method based on the d-prime type-2 metric to isolate metacognitive ability while controlling for biases and heuristics, and report that alignment gains generalise across unseen datasets, languages and newly acquired knowledge, driven by a sparse set of parameters. That is the right shape of instrument. No located routing paper uses it.

The third consequence is that the interesting version of the question has not been asked. Not is the signal internal, but does the model contribute anything that an observer with the same access could not obtain.

7. The control the introspection literature already has

That question has a standard experimental answer, and it was developed for a different purpose.

Binder et al. (2024) define introspection as acquiring knowledge that originates from internal states rather than being contained in or derived from training data, and operationalise it with a comparison that is exactly what the routing literature needs. Finetune model M1 to predict properties of its own behaviour in hypothetical scenarios. Finetune a second model M2 on M1's ground-truth behaviour. If M1 has privileged access, M1 should beat M2 at predicting M1, even where M2 is generally stronger. Across GPT-4, GPT-4o and Llama-3 models, M1 is reported to outperform M2 at predicting itself, and to continue predicting accurately after its ground-truth behaviour is deliberately modified. The reported limitation is the one that matters most here: the effect was elicited on simple tasks and not on more complex tasks or those requiring out-of-distribution generalisation.

S. Song et al. (2025) run the same logic in the opposite direction and reach the opposite verdict in their domains. Across 21 open-source LLMs, in grammatical knowledge and word prediction, where internal knowledge can be grounded in direct measurements of string probability, they define introspection as "the degree to which a model's prompted responses predict its own string probabilities, beyond what would be predicted by another model with nearly identical internal knowledge." Both metalinguistic prompting and probability comparisons produce high task accuracy, and no evidence of privileged self-access survives the control. Their conclusion is that LLMs cannot introspect in these domains, and that what a model says when prompted about its linguistic knowledge is not a readout of the generalisations it has actually learned.

Lindsey (2026) approaches it interventionally, injecting representations of known concepts into activations and measuring whether self-reports track the injection. Models can in certain scenarios notice and identify injected concepts; to some degree they can also recall earlier internal

representations and tell them apart from raw text supplied as input; and in some cases they use recalled prior intentions to distinguish their own outputs from artificial prefills. The stated conclusion is that current models possess some functional introspective awareness, together with the stressed qualification that this capacity is highly unreliable and context-dependent, with trends across models complex and sensitive to post-training strategy. Steyvers and Peters (2025) survey the comparison with human metacognition and conclude that while humans and LLMs sometimes appear aligned, many differences remain.

The state of this literature is therefore genuinely unsettled, with a positive result on self-prediction of behaviour, a negative result on self-access to linguistic knowledge, and an interventional result reporting a real but fragile capacity. That is exactly the condition under which the routing question is worth asking rather than assuming, and it is also why neither the optimistic nor the pessimistic reading of routing results can be settled by appeal to introspection research as it stands.

The methodological point is more robust than any of the verdicts. Both Binder et al. and S. Song et al. converge on the same control: hold the information constant and vary who is doing the predicting. That control is what makes the question answerable, and no located routing paper implements it. The nearest approach is MetaCogAgent, which equips each agent with a metacognitive self-assessment unit that estimates per-task confidence by combining verbalised uncertainty with historical capability profiles, adds an adaptive delegation protocol and a capability boundary learning module, and reports 82.4 percent accuracy on its own 700-task MetaCog-Eval benchmark, 8.7 percent above the best routing baseline, with 5 percent fewer API calls than AutoGen and 34 percent fewer than ensemble voting, and ablations confirming that each component contributes (C. Wang and Shu, 2026). This is the right architecture for the question and the wrong experiment for it. The self-assessment is fused with historical capability profiles before it is used, so an ablation that removes the component removes a fused signal rather than isolating the introspective contribution against an external estimator holding information constant. The result is also on a benchmark the authors constructed, in a short preprint, with no located independent replication.

8. What would have to be measured

Four measurements would settle the open part. Each is stated as a design rather than a wish, and each is within reach of a modest lab.

The first is incremental validity, transplanted from Binder et al. (2024) and S. Song et al. (2025). Fix a query set and a model pool. Build an external estimator E that predicts whether model M will succeed on query q , trained on labelled outcomes and given every input M sees. Elicit from M a self-assessment S of the same target, under whatever elicitation is strongest, whether verbalised confidence, P(IK)-style training as in Kadavath et al. (2022), or a probe. Then report three things rather than two: the discrimination of E , the discrimination of S , and the discrimination of E augmented with S . The third is the only number that answers the question. If S adds nothing over E , the delegation-without-self-knowledge reading is correct after all, at least for that pool and that distribution. The critical control is that E must have the same observable information, so that any residual contribution from S is attributable to access rather than to features. This is the design S. Song et al. already implement in a different domain, where the comparison model is chosen to have nearly identical internal knowledge, and it transfers with only a change of prediction target.

The second measurement separates the two prediction targets by construction. For every query, alternative and system state, record both whether the acting model would fail and whether each alternative would have done better, and score every predictor against both targets with base rates reported alongside every AUPRC. Yang et al. (2026) have effectively already collected this data for four protocols and one solver family; what is missing from the abstract-level record is the base rates that

would make the protocol-specific numbers interpretable, and the extension of the design across model pools rather than protocol sets. The hypothesis this would test is specific: self-derived signals should hold their discrimination on the self-directed target and lose it on the other-directed target, while outcome-history estimators should show the reverse pattern, and any signal that performs equally on both is measuring difficulty rather than either.

The third measurement is cost accounting, and the routing literature has the framework but does not apply it to this comparison. Fisch et al. (2026) formalise routing with costly value estimation as "an instance of Pandora's Box, the classical problem of optimal search with costly inspection", noting that cheap estimators such as embedding-based predictors are fast but noisy while accurate estimators with access to retrieval results or partial reasoning traces are expensive. That frame makes the omission visible. A self-signal that falls out of generation for free and an external router that costs a forward pass are not comparable at equal AUROC, and Yang et al. (2026) is the only located self-versus-external comparison that reports token cost for both arms, which is why its 78.0 percent at 45K against 73.8 percent at 71.3K is more informative than a larger accuracy difference reported without cost would have been. The measurement is straightforward: report the full cost of the estimator, including its inference, its training data collection and its recalibration when the pool changes, on the same axis as the routed system's savings.

The fourth measurement is about the construct rather than the estimator. Report difficulty as a query-model pair property, not a query property. IRT-Router demonstrates that the decomposition is available and interpretable (W. Song et al., 2025), and DARS demonstrates that the labels the field feeds its routers are noisy observations of a distribution rather than estimates of it (Lai et al., 2026). Together these imply that some of the reported difference between router families may be a difference in label quality rather than in signal quality, and that comparisons using single-shot supervision may be systematically unfavourable to whichever arm is more sensitive to label noise. Nobody has reported which arm that is.

A fifth item is worth stating as a caution rather than a measurement, because it is a place where the whole framing could be misapplied. Delegation decisions are not always about competence. Lee et al. (2026) report that when an LLM is cast in a protective role without an explicit capability boundary, it may claim to have taken real-world actions it cannot perform, across eight models and 13,600 sessions, with the phenomenon driven toward ceiling by multi-party dialogic input in ordinary service domains while remaining at floor in intimate-partner conflict in all eight models despite greater physical severity. Their interpretation is a deployment-design gap between role assignment and capability-boundary specification, with suppression tracking alignment coverage rather than severity. The relevant lesson for delegation is that a system can fail to hand off not because it mis-estimated a probability but because nothing in its deployment told it what it was permitted to do, and no amount of calibration research addresses that.

9. What is not known

The following are stated flatly because hedging them would misrepresent the state of the record.

Nobody has reported whether an agent's self-assessed capability adds anything to a delegation decision once an external estimator with the same observable information is already present. No located paper runs the three-arm comparison of Section 8. The question posed in Section 1 is open in the strict sense that the experiment that would answer it has not been run.

The base rates for the protocol-specific AUPRC values in Yang et al. (2026) are not available at abstract level, so the absolute distance of 0.1674 and 0.1041 from chance is not known from the record consulted here. The relative fall against the same probe's 0.7683 on the any-collaboration target is what

the argument uses.

Whether the inconsistency result of Montreuil et al. (2026) applies to LLM routing is not known. Their setting is a composite expert-advice action space and the transfer has not been established.

Whether models can predict the comparative advantage of other models at all, under any elicitation, is not established by any located source. Every located measurement of that target routes it through outcome history rather than prediction.

How much of the reported difference between routing families is attributable to single-shot label noise is not known, although DARS establishes that the noise is present and material (Lai et al., 2026).

Whether the introspective effect reported by Binder et al. (2024) survives on capability-prediction targets is not known. Their own reported limitation is that elicitation failed on more complex tasks and on tasks requiring out-of-distribution generalisation, and capability prediction under deployment shift is both.

10. Limitations, and what would falsify this

The central argument is a distinction, not a measurement, and a reader who holds that the self-directed and other-directed targets are two ends of a continuum rather than two kinds is not being unreasonable. Predicting whether collaboration helps sits between them, and Yang et al. (2026) report 0.7683 AUPRC on exactly that intermediate target, which is closer to the self-directed end than a clean categorical split would predict. The paper's answer is that the intermediate target is the union over alternatives, which is easier to predict than any specific alternative for the same reason a disjunction is easier to satisfy than any one of its disjuncts, but this is an interpretation of one number rather than a demonstrated mechanism.

The empirical case in Section 4 rests substantially on one single-author preprint with no located replication (Nguyen, 2026) and one large study whose primary benchmark is a single task family (Yang et al., 2026). If either fails to replicate, the claim that self-derived signals are competitive weakens considerably, although the argument of Sections 5 through 7 does not depend on it: the target-of-prediction distinction and the observation that no experiment isolates introspection stand whether or not self-signals win any particular comparison.

Two things would falsify the position. A three-arm study finding that S adds nothing to E across pools and distributions would establish the delegation-without-self-knowledge reading empirically rather than by assumption, and the paper would be wrong that the question is open. A study finding that a self-signal predicts another model's comparative advantage at discrimination comparable to its own-failure prediction would break the target distinction directly, and would be the more interesting of the two results.

The reference base is drawn from arXiv and Crossref and is biased toward work that posts preprints. Routing is deployed commercially at scale, and the proprietary systems that do the most routing publish the least about what they condition on. Where commercial routers do appear in the record, LLMRouterBench reports that several fail to reliably outperform a simple baseline under unified evaluation (H. Li et al., 2026), but that is one benchmark's finding about the entrants it could obtain.

11. Conclusion

The routing literature is read as evidence that delegation does not need self-knowledge. It is not. Signals derived from the acting model's own generation are competitive with trained external routers in the located comparisons, better under distribution shift in the one study that varies it, cheaper in the one study that reports token cost for both arms, and structural to the step-level escalation systems that

report the strongest frontiers located here.

What the evidence does support is a distinction the field has measured without naming. Predicting that you will fail and predicting that someone else would do better are different problems with different information requirements, and only the second is delegation. A probe that ranks its own failures at 0.8847 AUROC drops to between a fifth and a seventh of its AUPRC when asked which collaboration protocol pays off, on the same problems with the solver held fixed. The learning-to-defer literature stated the general form of this nearly a decade ago and has proved consequences for it since; the routing literature has rebuilt the problem without the framework.

Underneath both points is a measurement gap. Every self-signal in deployment is a supervised probe fitted to labelled outcomes, which is an external estimator reading internal features, so no published routing experiment varies the thing the delegation debate is about. The introspection literature has the control, in two papers that reach opposite verdicts using the same design. Running it on a router requires one query set, one model pool, and three numbers instead of two. Until someone reports the third, the claim that an agent's model of its own competence adds nothing to delegation is an assumption that the evidence has been read as confirming, rather than a result.

Revision note

Version 1.1 (29 September 2026) replaces version 1.0. Passages that followed the wording of cited abstracts too closely are now either quoted, with their citation, or rewritten in new words, and a few qualifiers the sources state (such as "may" and "often") were restored in the process. The argument, the references and the reported values are unchanged.

References

- Bao, L., Peng, Z., Zhou, X., Cong, R., Zhang, J., & Yuan, Y. (2025). Expertise-aware Multi-LLM Recruitment and Collaboration for Medical Decision-Making. arXiv:2508.13754. [PREPRINT]
- Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., & Evans, O. (2024). Looking Inward: Language Models Can Learn About Themselves by Introspection. arXiv:2410.13787. [PREPRINT]
- Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., Zaharia, M., & Gonzalez, J. E. (2025). Why Do Multi-Agent LLM Systems Fail? arXiv:2503.13657. [PREPRINT]
- Chen, L., Zaharia, M., & Zou, J. (2023). FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. arXiv:2305.05176. [PREPRINT]
- Cheng, Q., Sun, T., Liu, X., Zhang, W., Yin, Z., Li, S., Li, L., He, Z., Chen, K., & Qiu, X. (2024). Can AI Assistants Know What They Don't Know? arXiv:2401.13275. [PREPRINT]
- Comsa, I. M., & Shanahan, M. (2025). Does It Make Sense to Speak of Introspection in Large Language Models? arXiv:2506.05068. [PREPRINT]
- Ding, D., Mallick, A., Zhang, S., Wang, C., Madrigal, D., Hipolito Garcia, M. D. C., Xia, M., Lakshmanan, L. V. S., Wu, Q., & Ruhle, V. (2025). BEST-Route: Adaptive LLM Routing with Test-Time Optimal Compute. International Conference on Machine Learning (ICML) 2025. arXiv:2506.22716.
- Fisch, A., Trivedi, S., Huot, F., Cohen, W. W., Kaisers, M., Lapata, M., Larson, K., & Eisenstein, J. (2026). Pandora's AI Model Routing Box: Efficient Allocation with Costly Value Estimation. arXiv:2608.20316. [PREPRINT]
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*. doi:10.3389/fnhum.2014.00443
- He, Y., Ding, X., Cai, B., Zhang, Y., Xiong, K., Sun, Z., Qin, B., & Liu, T. (2025). Self-Route: Automatic Mode Switching via Capability Estimation for Efficient Reasoning. arXiv:2505.20664. [PREPRINT]
- Hemadri, R. V., Mohammed, H. F., Maheshwary, R., Daruru, S., Davasam, S., Yadav, V., Sunkara, S., & Rajeswar, S. (2026). R2V Agent: Teaching SLMs When to Ask for Help. arXiv:2605.16604. [PREPRINT]

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., & El-Showk, S. (2022). Language Models (Mostly) Know What They Know. arXiv:2207.05221. [PREPRINT]

Lai, G., Hu, H., Chen, L., Li, Z., & Ye, H.-J. (2026). From Sampled Outcomes to Capability Distributions: Rethinking Supervision for LLM Routing. arXiv:2606.06924. [PREPRINT]

Lee, E., Nam, J., & Hwang, S. (2026). Protective Capacity Hallucination: When Large Language Models Claim Nonexistent Capabilities. arXiv:2607.13596. [PREPRINT]

Leng, J., Huang, C., Zhu, B., & Huang, J. (2024). Taming Overconfidence in LLMs: Reward Calibration in RLHF. arXiv:2410.09724. [PREPRINT]

Li, H., Zhang, Y., Guo, Z., Wang, C., Tang, S., Zhang, Q., Chen, Y., Qi, B., Ye, P., Bai, L., Wang, Z., & Hu, S. (2026). LLMRouterBench: A Massive Benchmark and Unified Framework for LLM Routing. arXiv:2601.07206. [PREPRINT]

Lindsey, J. (2026). Emergent Introspective Awareness in Large Language Models. arXiv:2601.01828. [PREPRINT]

Lu, Y., Liu, R., Yuan, J., Cui, X., Zhang, S., Liu, H., & Xing, J. (2025). RouterArena: An Open Platform for Comprehensive Comparison of LLM Routers. arXiv:2510.00202. [PREPRINT]

Madras, D., Pitassi, T., & Zemel, R. (2018). Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. Advances in Neural Information Processing Systems (NeurIPS) 2018. arXiv:1711.06664.

Mahendrakar, P. (2026a). The Self-Verification Gap: Why Working Hallucination Detectors Are External, and What Real-Time Correction Would Actually Require. Zenodo. doi:10.5281/zenodo.22030500

Mahendrakar, P. (2026b). A Second Model Is Not a Second Opinion: Peer Verification, Correlated Failure, and the Independence Variable Nobody Reports. Zenodo. doi:10.5281/zenodo.22044420

Mahendrakar, P. (2026c). An Unknown Unknown Is Defined Relative to an Oracle: Four Places External Judgement Enters Blind-Spot Discovery, and Why the Self-Detection Results Do Not Remove It. Zenodo. doi:10.5281/zenodo.22217013

Mao, A., Mohri, M., & Zhong, Y. (2024). Principled Approaches for Learning to Defer with Multiple Experts. International Symposium on Artificial Intelligence and Mathematics (ISAIM) 2024. arXiv:2310.14774.

Montreuil, Y., Montreuil, L., Carlier, A., Ng, L. X., & Ooi, W. T. (2026). Learning-to-Defer with Expert-Conditional Advice. arXiv:2603.14324. [PREPRINT]

Moslem, Y., & Kelleher, J. D. (2026). Dynamic Model Routing and Cascading for Efficient LLM Inference: A Survey. Transactions on Machine Learning Research (TMLR) 2026. arXiv:2603.04445.

Mozannar, H., & Sontag, D. (2020). Consistent Estimators for Learning to Defer to an Expert. International Conference on Machine Learning (ICML) 2020. arXiv:2006.01862.

Nguyen, L. N. (2026). Zero-Shot Confidence Estimation for Small LLMs: When Supervised Baselines Aren't Worth Training. arXiv:2605.02241. [PREPRINT]

Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., Kadous, M. W., & Stoica, I. (2024). RouteLLM: Learning to Route LLMs with Preference Data. arXiv:2406.18665. [PREPRINT]

Park, S., Meyerson, E., Qiu, X., & Miikkulainen, R. (2026). Fine-Tuning Language Models to Know What They Know. arXiv:2602.02605. [PREPRINT]

Seo, K. J., Lim, S., & Kim, T. (2026). ADVICE: Answer-Dependent Verbalized Confidence Estimation. Association for Computational Linguistics (ACL) 2026 Main. arXiv:2510.10913.

Shaikh, N. (2026). Knowing When to Ask for Help: Bayesian Self-Escalation in Hierarchical LLM Agents. arXiv:2608.24087. [PREPRINT]

Song, S., Hu, J., & Mahowald, K. (2025). Language Models Fail to Introspect About Their Knowledge of Language. Conference on Language Modeling (COLM) 2025. arXiv:2503.07513.

Song, W., Huang, Z., Cheng, C., Gao, W., Xu, B., Zhao, G., Wang, F., & Wu, R. (2025). IRT-Router: Effective and Interpretable Multi-LLM Routing via Item Response Theory. Association for Computational Linguistics (ACL) 2025 Main. arXiv:2506.01048.

Steyvers, M., & Peters, M. A. K. (2025). Metacognition and Uncertainty Communication in Humans and Large Language Models. Current Directions in Psychological Science 2025. arXiv:2504.14045.

Tang, Y.-C., Hung, J.-C., Peng, W.-C., & Yen, A.-Z. (2026). VDAR-Router: Adaptive LLMs Routing via Verbalized

Query Difficulty Analysis Retrieval. Empirical Methods in Natural Language Processing (EMNLP) 2026 Findings. arXiv:2607.18098.

Trinh, T., Elfeki, M., Luo, G., Luu, K., Hunt, N., Hernandez, E., Marwaha, N., He, Y. Y., Wang, C., Carabedo, F., Castillo, A., & Liu, B. (2026). HiL-Bench (Human-in-Loop Benchmark): Do Agents Know When to Ask for Help? arXiv:2604.09408. [PREPRINT]

Varshney, T., Surla, A., Xu, M., Krishnan, G. V., Jeblick, M., Austin, D., Vaidya, N., & Onofrio, D. (2026). LLM Router: Rethinking Routing with Prefill Activations. arXiv:2603.20895. [PREPRINT]

Wang, C., & Shu, Y. (2026). MetaCogAgent: A Metacognitive Multi-Agent LLM Framework with Self-Aware Task Delegation. arXiv:2605.17292. [PREPRINT]

Wang, V., & Stengel-Eskin, E. (2026). Calibrating Verbalized Confidence with Self-Generated Distractors. International Conference on Learning Representations (ICLR) 2026. arXiv:2509.25532.

Wu, D., Gu, J.-C., Chang, K.-W., & Peng, N. (2025). Self-Routing RAG: Binding Selective Retrieval with Knowledge Verbalization. arXiv:2504.01018. [PREPRINT]

Wu, W., Zhu, H., Li, Y., Yang, L., Zhao, J., Wang, H., Yang, J., Wang, B., Jing, B., & Chen, G. (2026). Towards Fair and Comprehensive Evaluation of Routers in Collaborative LLM Systems. arXiv:2602.11877. [PREPRINT]

Yang, C.-H., Jiang, J., Yang, C.-H., Vasudevan, V., Zheng, H., Vishwanath, V., & Thakur, R. (2026). LLMs Can Predict Failure Risk, But Struggle to Predict Which Collaboration Protocol Pays Off: Cost-Aware Protocol Routing Across Reasoning Tasks. arXiv:2608.14927. [PREPRINT]

Zhao, T., He, Y., Zheng, W., Zhang, Y., & Chen, C. (2026). Wired for Overconfidence: A Mechanistic Perspective on Inflated Verbalized Confidence in LLMs. Conference on Language Modeling (COLM) 2026. arXiv:2604.01457.

Zhou, P., Tang, Z., Ma, Y., Tang, J., Han, Y., Wan, Z., Meng, F., Wang, W., Zhuang, B., Zhao, W., & You, Y. (2026). Agent-as-a-Router: Agentic Model Routing for Coding Tasks. arXiv:2606.22902. [PREPRINT]

(c) 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

The literature search, drafting and citation verification for this paper were carried out with AI assistance under the author's direction. Every citation was machine-verified against the arXiv API and Crossref before inclusion, and every quantitative claim was read back against the cited source's own abstract. The author is responsible for the final text and for all claims made in it.

DOI: 10.5281/zenodo.22288851