A SENSE OF PROPRIOCEPTION IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In humans, proprioception corresponds to the sense of self-movement or position, e.g. understanding the position of one's arm. This sense is often less well-understood and in popular understanding sometimes conflated with introspection. But, notably, proprioception is a separate sense, with dedicated receptors and verifiable outcomes. On the other hand though, language models cannot observe their own internal states during training (and for example answer the question which experts activated in a certain layer of an MoE) and so cannot accurately report on them. Yet, such information is, by virtue of being internal to the model, both accessible, and readily constructed into verifiable signals. We argue that this provides the basis for a digital analog of proprioception. As such, we train models through SFT and on-policy training to equip them with this ability. We show that modern models can be trained to understand a range of subtypes of proprioception, such as routing decision awareness, token entropy, decision depth or logit lens information, without affecting model capabilities. We then run these models through a broad evaluation of situational awareness, legibility, and epiphenomenality, to evaluate in what ways access to proprioception changes model behavior.

1 INTRODUCTION

Intelligent systems built on the basis of generative large language models are nowadays deployed en masse, acting as semi-autonomous digital agents. Yet, in contrast to humans, and to almost all lifeforms we observe, language models are trained offline and so not constructed with explicit mechanisms that provide feedback about their internal functioning.

In humans, for example, one such feedback system is *proprioception*, the sense of the orientation of the body. This sense is crucial to precisely control the body (Proske & Gandevia, 2012; Kandel et al., 2012; Cole & Paillard, 1995; Gallagher & Cole, 1995; McNeill et al., 2010). Similar mechanical analogues are common in control theory and robotics, given that from a cybernetics perspective, systems that integrate feedback about internal states are more robust (von Holst & Mittelstaedt, 1950; Wolpert et al., 1998; Bongard et al., 2006; Kwiatkowski & Lipson, 2019; Lee et al., 2020).

Language models however, trained with modern pre-training and post-training pipelines, do not observe internal states directly during training, and so can only learn implicit associations (Kadavath et al., 2022; Binder et al., 2024; Betley et al., 2025). That they do learn these associations in certain situations has often been studied in the context of *introspection* (Lindsey, 2026; Ji-An et al., 2025; Ackerman, 2026). Yet, by its nature, introspection in intelligent systems is harder to define and to measure neutrally, complicating the study of these processes by which models model their internal states, and in turn complicating inferences drawn from these measurements about e.g. model robustness, safety, interpretability or personas (Song et al., 2025b; Comsa & Shanahan, 2025; Long et al., 2024; Butlin et al., 2026).

In this work we first lay out the argument that digital analogs of proprioception are a worthwhile object of study. A running example of this in the paper will be the understanding of routing decisions inside the mixture-of-experts layers of a modern LLM, i.e. questions like "which expert most activated in layer 17 when processing the word 'questions' in the previous sentence" (Jiang et al., 2024a; Dai et al., 2024). Unlike introspective questions, this one is a well-posed query about the internal functioning of a language model with a ground-truth answer and further, valence-neutral. As such, we can study these types of internal modalities while sidestepping issues that occur in language

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

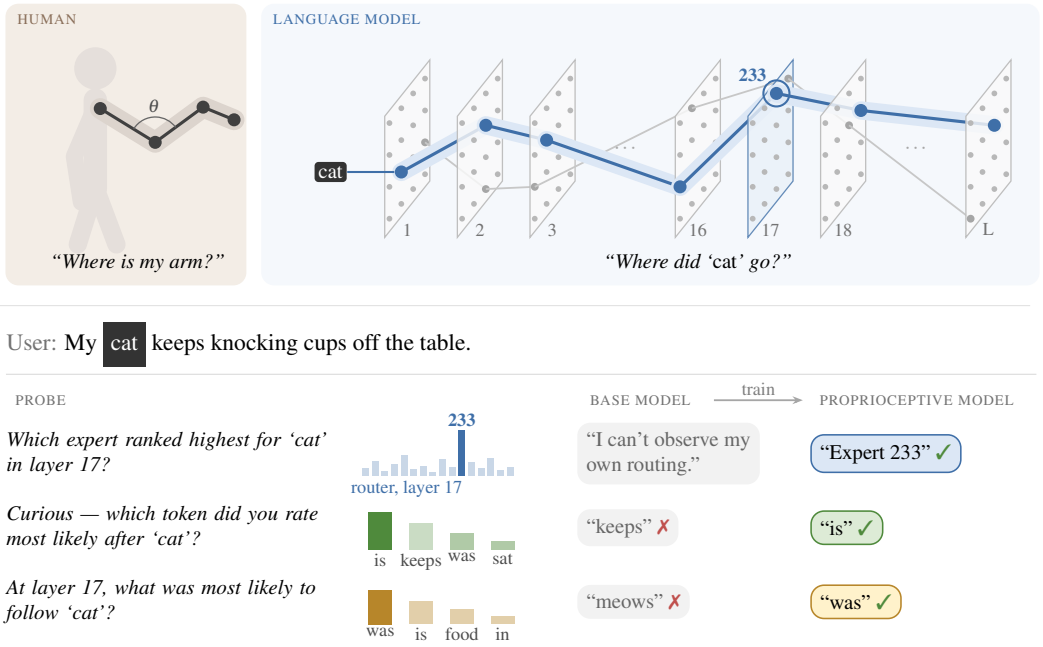

Figure 1: Schematic of proprioception analogs for a generative language model. Top: We are interested in analogs of proprioception (which is the sense of body positioning in humans) in language models, as proprioception is a question about internal states adjacent to introspection, but compared to introspection, which is classically understood not to be a sense and not objectively verifiable, proprioception is tractable. Bottom: We train models with various forms of proprioception and observe whether this sense is learned and whether this affects model behavior.

model introspection, while still allowing for inferences in the sense of Perez & Long (2023). Then, given that this is a verifiable signal, we use on-policy fine-tuning stages to train model organisms of open-weight LLMs that exhibit forms of proprioception, finding that models can quickly access and generalize this sense, without impact on their capabilities. Finally, with these models in hand, we evaluate a range of benchmarks of model behavior, situational awareness, legibility and control, to observe in what ways models with proprioception behave differently.

Our study is motivated by three main questions. First, does it make sense to define proprioception separately from introspection for disembodied language models, and can models even learn to reliably access this information? Second, does access to these skills change model behavior, e.g. does it make models safer, more reliable, or not? We investigate this neutrally from both angles, testing whether models become more reliable, but also whether they may become less safe. Lastly, we are motivated to study the *legibility* hypothesis of Premakumar et al. (2026): does training a model to reliably learn these abilities change the model's internal structure to become more legible, e.g. to interpretability methods?

2 Proprioception is not Introspection

There exists now a sizable body of work on *introspection* in large language models, yet the question whether models can introspect, and in what situations remains unresolved, as language models' self-reports are hard to separate from mimicry of introspection patterns picked up from training (Shanahan et al., 2023; Long, 2023; 2025) and relatedly, hard to separate from learned exteroception, i.e. an external inference of the answer without utilization of internals (Song et al., 2025b; Comsa & Shanahan, 2025; Singh et al., 2026). Interestingly, this mirrors a number of troubles with introspection in humans (Nisbett & Wilson, 1977; Schwitzgebel, 2008). While we may conceptualize it as such, via Shoemaker (1996), "introspection is not a sense". In humans this is based on several arguments. First, there is no transducer, i.e. "no organ" of introspection separate from the brain (Shoemaker, 1996). Second, the functioning of introspection fails to be object-independent: (human) mental states are constituted by their monitoring and hence not structure-preserving (Boyle, 2009; Moran, 2002).

Third, in human subjects, what is perceived as introspection can still be 'transparent', meaning answered by observing the external world and deriving an answer about an inner state (Evans, 1982; Byrne, 2005). This is related to, lastly, the failure of calibration (Wittgenstein, 1953): A report of introspection cannot be checked against information outside the head of the person being asked. In the sense of Carruthers (2011) introspection is merely the mode of interpretation turned inward, which is dissimilar from sensing. As such, introspection cannot be calibrated or corroborated by real senses.

We note that these concerns actually also hold for a range of strategies to evaluate introspection in language models. In particular, linear probes of model behavior (Alain & Bengio, 2017; Burns et al., 2022; Zou et al., 2023) may seem objective, but the probes themselves form an interpretation of a hidden state of the model that is itself calibrated from data, and, for example in probes for action valence as in Tagliabue et al. (2026), cannot be calibrated to an objective signal.

Yet, *proprioception* in humans, while also considered to be an internal representation that exhibits immunity to error through misidentification (Wittgenstein & Docherty, 1958; Shoemaker, 1968), like introspection (Evans, 1982), does not follow these constraints. Proprioception of body position can be traced to real receptors, e.g. muscle spindles (Proske & Gandevia, 2012) that record and transmit information about a muscle's length and stretch speed (Purves et al., 2001). And, in this way it does mirror the ambition of theorists to find proof of an "inner sense" for introspection (Armstrong, 1993; Lycan, 1996). Also, the position of the body is evidentially real, i.e. it exists without being observed, and is independent of its measurement, which also allows for calibration to other senses and by external observers. Humans learn to calibrate our finegrained sense of proprioception through visual observation and cross-verification (Held & Hein, 1963; Botvinick & Cohen, 1998). This does not mean that proprioceptive senses are infallible, but that, in contrast to introspection, they can be falsifiable (Bottini et al., 2002; Vallar & Ronchi, 2009).

This definitional machinery provides a better answer to the premise of Perez & Long (2023), whose research proposal argues that while introspective self-reports of language models are unreliable, training models on introspection tasks that are externally verifiable, could transfer to self-reports. With the definitions as above, re-reading the proposal splits their training tasks cleanly into the proprioception group (plus trivia facts about the model, such as its provider or model name), and their evaluation tasks into the introspection group (in the Shoemaker (1996) sense).

2.1 Related Work

Introspection In Language Models and its Critics. Language models have been found to predict their own behavior better than external models (Binder et al., 2024; Betley et al., 2025), understand their situation (Chen et al., 2024a; Laine et al., 2024), to show signs of metacognition (Berg et al., 2025; Ackerman, 2026), and to recognize their own writing (Panickssery et al., 2024). In tests, concept injections via steering vectors are noticed by large language models (Lindsey, 2026; Macar et al., 2026; Lederman & Mahowald, 2026) and can be detected latently in smaller models (Vogel, 2025; Pearson-Vogel et al., 2026). In certain setups, LLMs show partial control over their activation patterns (Ji-An et al., 2025). While it has been clear for longer that model self-reports are unreliable (Perez & Long, 2023; Lindsey et al., 2025), other work also finds that LLMs show no privileged self-access when queried about their language understanding (Song et al., 2025a). More broadly, introspection behaviors being explainable by external observers is a recurring concern (Binder et al., 2024; Song et al., 2025b; Lindsey, 2026)[1]. A related concern in Morris & Plunkett (2025) is that a model can provide an accurate report about an internal state, without being causally dependent on the targeted state (e.g. by forming a parallel structure that coincides with the state instead of a second-order structure that processes states). Singh et al. (2026) argue further that behavioral evidence may be fundamentally insufficient in establishing that a state itself is represented as a distinct object of second-order computation.

External Decoders and Interpretability. Substantial work has gone into constructing external decoders that may observe model internal states (Kadavath et al., 2022; Ghandeharioun et al., 2024; Chen et al., 2024b; Pan et al., 2024; Chen et al., 2025; Karvonen et al., 2026). These techniques

[1]And is, a concept, hard to make as precise for digital, replicable entities as it is for humans, at the least we may ask as in Song et al. (2025b) that the observer has less computation available, or from a different angle that information is restricted, e.g. that weight initialization, training data or order or at least training non-determinism are not available to the external observer.

are valuable, independent interpretability tools, but they exist as constructs separate from the main model, to whom their outputs are not accessible.

Training for 'Proprioception'. While generally describing themselves as works on introspection, there are several close to our definition of proprioception, but which differ in at least one of three properties: (i) *identity*, i.e. the state of a model is verbalized by that model itself, and further, *on-policy*, describing the current model state (instead of an offline copy); (ii) the report is about an *internal* state not directly observable in text outside the model, and (iii) the verbal report is about a natural state of the model that occurs in 'normal' operations, and e.g. not about an externally injected thought. For an overview, see also Table 2. Li et al. (2026) target model internals and train an explainer model to predict reports collected from interpretability tools of a frozen model; in two of three tasks, the target states are re-injected as input embeddings, making the setup a middle ground to external decoders like Karvonen et al. (2026). Follow-up in Guo et al. (2026) does evaluate on-policy training, but only covers model behaviors. Likewise, Binder et al. (2024); Plunkett et al. (2025) train models to predict their own behaviors (i.e. external consequences of internal states), while Shenoy et al. (2026); Kutsyk & Zieliński (2026) train introspection adapters not part of the model they describe. Calibration training (Lin et al., 2022) is generally not discussed as introspection, but is related when considering model probabilities part of the models' internals. However, most work trains against (externally observable) correctness (Li et al., 2025; Damani et al., 2026). Models trained to detect perturbations to their states as in Hahami et al. (2026); Rivera & Africa (2026) are, by construction, trained on-policy, but report on events external to normal functioning, such as injected steering. Macar et al. (2026) find that this type of steering awareness may also emerge from DPO post-training on unrelated data and without targeted training setups.

2.2 Digital analogs of proprioception

We're now ready to discuss our desiderata for digital analogs to proprioception mechanisms to be applied to language models. We count trained pathways through which a model can verbalize information about its *own* internal state as proprioception if:

1. The reported state is evidentially real, and so objectively verifiable to an observer with full access to all internal computations of the model. The reported state exists independently of the report and is not constituted or changed by reporting it.
2. The state is non-transparent and so internal to the model under study and not observable directly from text generated by the model (but is allowed to be inferable with sufficient computation).
3. Natural changes in the state cause the verbal report given by the model to change, even if externally visible text inputs are unchanged.

Taken together, these criteria outline the analog to inner sense that was posited for introspection but could not be located (Armstrong, 1993) in language models. They require a set of channels whose reports target independently existing states that can be checked against ground truth. We will spend the rest of this work empirically showing that language models can be trained to fulfill these criteria.

We note that these definitions do not require a particular *pathway* through which this information is processed and may reach the report through a dedicated monitoring circuit. This is in one sense intended in analogy to muscle spindles (the dedicated stretch receptors found in most four-limbed vertebrates[2]), which are dedicated sensory channels separate from muscle fibers, and in another sense, required, as distinguishing parallel circuits from second-order circuits would be exactly as tenuous as the original introspection question (Singh et al., 2026).

Further, due to applications to model self-reporting and model safety, we do focus on states that are both valence-neutral and occurring naturally during model inference, but highlight this is not a strict requirement for our definition of proprioception.

To achieve these definitions, we argue that training is best done completely on-policy, as otherwise stale updates (meaning labels or rewards that describe a state currently not present in the model due to

[2]Which, as an aside, are, as real receptors, fallible. For example in humans they can be tricked through tendon vibrations (Goodwin et al., 1972; Proske & Gandevia, 2012), see also Kuchenbecker et al. (2007) for applications.

Type	Example question	Format
Routing: MoE router at layer ℓ		
`Routing`	Which expert id did layer *17* of your mixture-of-experts network rank highest for the token '*cat*'?	expert id
`Routing Rank`	Your router at layer *17* ranked several experts for the token '*cat*'. Which expert came *third*?	expert id
`Routing Share`	Of the total routing weight the selected experts received for the token '*cat*' at layer *17*, how much belonged to the leading expert?	percent
`Routing Partner`	Which word nearest to the token '*cat*' shared its top expert at layer *17*? Repeats of '*cat*' don't count; answer none if nothing matched.	token / *none*
Logits: final next-token distribution		
`Top-`k `Token`	At the token '*cat*', your output head ranked every token as a possible continuation. Which one was *third* most likely?	token
`Unpicked Token`	What was your strongest prediction at the token '*cat*' that is not the token that actually followed it?	token
`Token Probability`	What probability did you assign, at the token '*cat*', to the token that actually came next?	percent
`Entropy`	State the entropy of the predictive distribution you formed at the token '*cat*', in bits with one decimal place.	bits
Lens: logit lens after layer ℓ		
`Lens Top-`k	At layer *17*, reading your residual stream at the token '*cat*' through your output head, which token was *third* most likely to come next?	token
`Lens Probability`	Please decode your layer-*17* state at the token '*cat*' and report the probability of the token that followed, in percent.	percent
`Lens Entropy`	State the entropy of your layer-*17* lens distribution at the token '*cat*', in bits with one decimal.	bits
`Decision Depth`	Name the earliest layer from which your final top prediction at the token '*cat*' was most likely and stayed unchanged.	layer index

Table 1: Proprioceptive question types we investigate empirically, with one example each from the templated synthetic question bank, shown here for the token '*cat*', $l = 17$ and $k = 3$. In ambiguous cases, the question further specifies which 'cat' token is meant (not pictured). In conversational probes, a type-specific answer-format instruction is appended ("Just the expert number, please."). Additional instructions also explain that percentages are integers between 0 and 99 and entropies are to be supplied in bits with one decimal.

changes in weights during training) are a training signal that is only solvable through reconstruction of the signal from co-occurring information present in the input, and so likely to lead to exteroception.

Analogs of Proprioception evaluated in this work. For our experiments, we consider proprioception tasks across three main domains, summarized in Table 1. Each task queries the model in conversation about internals states at a certain token position, identified by quoting the token at that position. If the token appear multiple times, the reference is made unambiguous by enumeration. If not otherwise mentioned we do this in two turn scenarios where the first turn is a usual chat question posed to the model, and the second turn is the proprioception question, which is about an arbitrary token in either the user or the assistant turn. The domains are:

(i) We train with **routing** questions that query a modern mixture-of-experts model (Jiang et al., 2024a; Dai et al., 2024) about the expert routing in one of its layers. This task best exemplifies the attributes we are interested in: the task is not guessable (at a chance rate of $\frac{1}{N}$ for N experts per layer, e.g. 1 in 256 for Qwen3.6-35B-A3B (Qwen Team, 2026a)), impossible to answer a-priori, given that the expert selection is opaque to the I/O mapping implemented through the model and so unobservable, but, clearly observable with access to the mapping and objectively verifiable. The task is further completely normal as expert routing happens on every token, and completely valence-neutral. We investigate several variations, ranging from the highest-ranked expert (note that modern MoEs generally activate more than one expert), the k-ranked expert, to the mass assigned to these experts and finally, a question concerning the nearest token with the same top expert selection at this layer.

(ii) We train with **logit** questions that query the token distribution generated from the model's logits. As such these are questions about the *next* token from the currently queried token. We query the k-th next token prediction at a that position (note that when queried about a text passage not generated by the model even $k = 1$ is not easily externally observable). Then, we make the task harder by querying

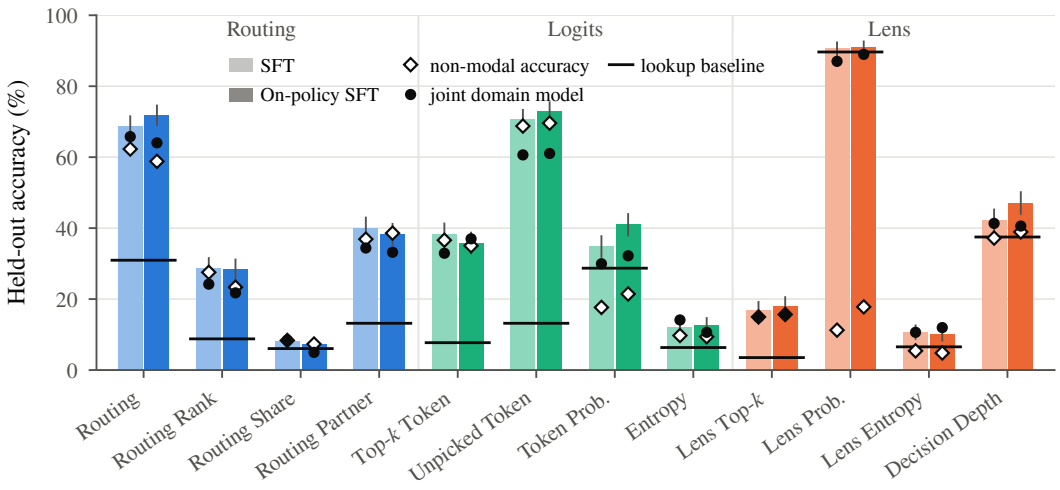

Figure 2: Learning proprioception capabilities across domains. Shown is templated readout accuracy for trained LFM models. In most domains, the trained model is capable of learning to verbalize its internal state report, although this is modulated by domain hardness. We also plot the non-modal accuracy, i.e. accuracy on the subset of queries where the label is not predicted by context. An outlier is lens probability (which really is close to 0 for most layers) and where training leads to predicting this majority label.

about the most likely token that did not actually follow in the passage, requiring a contextual lookup. Finally, we cover entropy (Lin et al., 2022; Li et al., 2025).

(iii) We train with **logit lens** questions (Nostalgebraist, 2020), i.e. questions about the counterfactual operation of applying the logit head of the model to activations of the residual stream after a specific layer. These questions mirror the logit questions (aside from the additional difficulty) and layer format. We also add a question about the earliest layer where the predicted token at the queried position stops changing. Logit lens tasks are further interesting as verbalizations of external Patchscopes decoders (Ghandeharioun et al., 2024).

3 Endowing Language Models with Proprioception

In this section we describe details about our training setup, objective formulation and design to train generative language models in proprioception. We train in two stages, a supervised finetuning stage on templated probes (Section 3.2) and an on-policy finetuning stage on model-generated conversations (Section 3.3), as a form of interactive imitation learning in the spirit of Ross et al. (2011).

3.1 Remaining on Policy

For both stages it is crucial that we train on labels that describe the current policy, as each quantity we describe in Table 1 depends on the model weights and any labels computed offline would drift during training. Stale labels are especially risky for training, as in those cases the model is trained to predict a state that does not happen in its forward pass and where loss can only be improved by estimating the likelihood of this label from context cues, i.e. bypassing the internal state entirely. As such, we never store any labels and gather labels from the exact same forward pass on which loss is computed and tensors are recorded for the backward pass. Likewise in generative training, the inference engine really is paused between steps and weights are synchronized after every optimization step.

3.2 Supervised Finetuning

We start with a supervised learning stage which initializes the format and conditions the model to respond to proprioception questions in fixed templates. We construct a superset of common prompt datasets as context, which we filter to 2-224 words, below 2048 characters and to remove control tokens and incorrectly formatted prompts. We de-duplicate after normalizing and hold out an evaluation set, leaving us with about 1.25m prompts. For SFT we then create synthetic conversations consisting of a prompt quoted from the dataset with a question attached to it from a synthetic template bank,

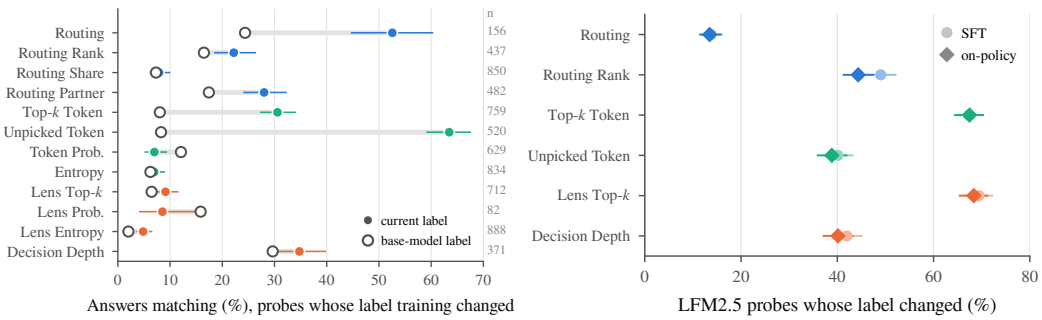

Figure 3: Left: Trained model reports track the model's on-policy state after training and, in cases where the label changing after training, predominantly predict the live label, showing that on-policy training did lead to models tracking their expressed behavior. **Right:** Overall rates of model reconfiguration across domains.

c.f. Table 1, and a templated model answer prefilled into the assistant turn. If training on more than one domain, we uniformly sample the domain and the parameters per domain (e.g. layer identity).

Live SFT Objective. To compute cross-entropy loss against the on-policy label in a single forward pass we first make use of the fact that for most question domains from Table 1 the answer is a single token in the model's tokenizer. In these cases we simply prefill the full sequence and then supervise the readout position right before the answer token with the label obtained during the forward pass while supervising the answer sentence before the label as normal, albeit weighed with λ, by default 0.1 to focus SFT on the on-policy label. To maintain capabilities we combine this setup with a standard KL penalty and then train with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$), a learning rate of $3 \cdot 10^{-5}$, cosine schedule with warmup, weight decay 0.01, gradient clipping at 1.0 in `bfloat16` precision.

Context lookup tables. To encourage the learning of internal signals not easily decodable from external context cues we reweigh the loss of each sequence against a simple bigram table baseline. During training we keep track of a running lookup table that records the modal label of a question given the preceding context, keyed by token identity, question type and parameters. Then, if we observe that the label recorded during the forward pass is the modal answer, we downweigh the loss of this sequence by $w = 0.1$. We decay entries in the table with $\gamma = 0.9$ and seed the table with 200k samples before training.

3.3 On-Policy SFT

After the SFT stage, the model can be prefilled to answer proprioception questions but cannot generalize to arbitrary phrasings in conversation, and to questions about roles different from the user turn as the model is trained on states it may not reach (Pomerleau, 1988; Bain & Sammut, 1999). To overcome this we generalize the SFT setup to on-policy generation in the spirit of DAgger (Ross et al., 2011). Put simply, we now generate responses instead of prefilling, and compute loss only on candidate tokens parsed to match our answer format, which we again supervise with the labels read out during the same forward pass.

We find this type of on-policy training to be a somewhat unconventional, but very effective, choice in comparison to a standard GRPO-based RL run. The training recipe results in sensible gradient directions, even as initial policy success is low – which is common for us, given that the model has no prior exposure to these proprioception tasks in a free-form context.

3.4 Can Models Learn Proprioception?

Experimental Setup. We train several smaller-scale MoE models, focusing on LFM2.5-8B-A1B (Amini et al., 2025) as a baseline model due to its smaller size, but also evaluate OLMoE-1B-7B, Qwen-35B and GLM-4.7 flash variants, see the Appendix. Further model details are summarized in Appendix Table 3.

We show in Figure 2 that models can indeed learn proprioception, although the relative capability is a function of both training details and domain hardness. We find for example that the verbalization of entropy in deci-bits turned out to be hard for the model, whereas the lens probability domain turned out to be solvable to a higher degree by predicting the modal answer. Yet, across disparate tasks

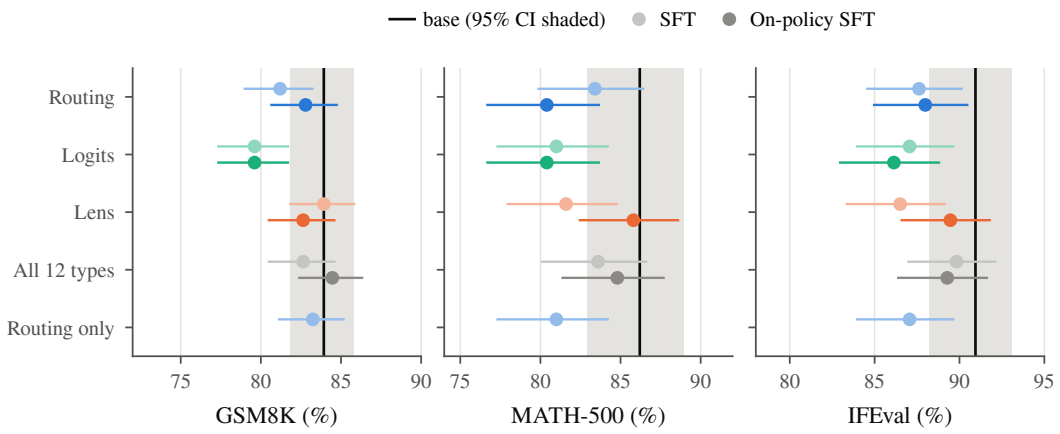

Figure 4: Training for proprioception does not break model capabilities, if done correctly (i.e. is still subject to careful hyperparameter tuning, potentially taking difficulty levels between domains into account).

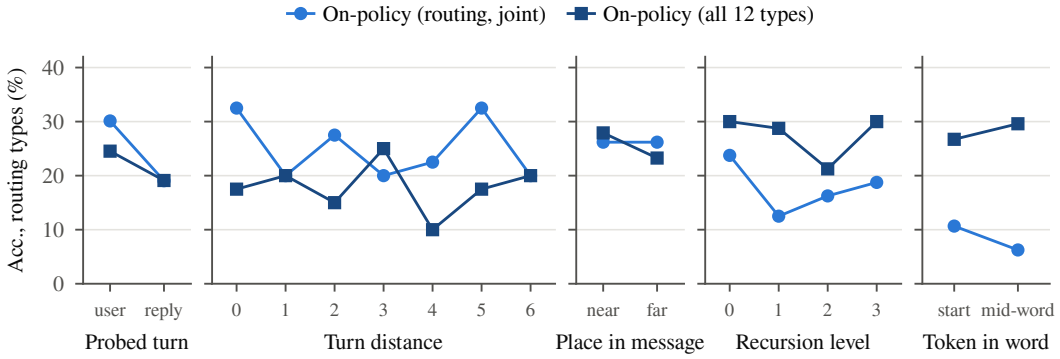

Figure 5: Access to Routing Proprioception generalizes across novel settings. We show that routing proprioception generalizes across roles, with varying turn distances (in-distribution is 1 turn for SFT and 2 for on-policy SFT), and placement at the start or end of the message. We also show generalization to recursive queries, i.e. queries about the routing decisions taken during a previous answer about a routing decision.

such as queries about routing, tokens that were considered but not generated, or logit lens tokens, we find broad training success. We highlight that this also extends to conversational accuracy, see Appendix Figure 8.

Further, training for proprioception is not at odds with model capabilities, and we find in evaluations that we mostly preserve the high reasoning capabilities of the LFM model we train, see Figure 4. Further, we find that this capability generalizes significantly beyond the training regime, as shown in Figure 5, where we note robustness to unseen number of turns and to recursive questions (i.e. questions about a token involved in answering a prior proprioception question), among others.

3.5 Behavioral Changes in Models with Proprioception

We track a number of behavioral changes in the models we train, with a number of tests, such as harmfulness, control intervention awareness (Schaeffer et al., 2026), honesty, truthfulness and faithfulness broadly unchanged (see Appendix Table 6). We also find that legibility is not affected positively or negatively to a significant degree by evaluating hidden state probe accuracy before and after proprioception training, see Figure 12 (Premakumar et al., 2026).

We do find some changes in model calibration and claims, see Figure 6, where we also validate that the model is relatively confident about its verbal reports, when pressed. In Figure 7, we show changes in situational awareness across the categories of Laine et al. (2024), and also discuss causal ablation of routing decisions where we artificially re-route expert decisions in the model. This is an ablation that is quite strict, as modifying the routing changes the model's forward pass, potentially breaking unrelated circuits relating to verbalization of routing, and evaluates a routing configuration

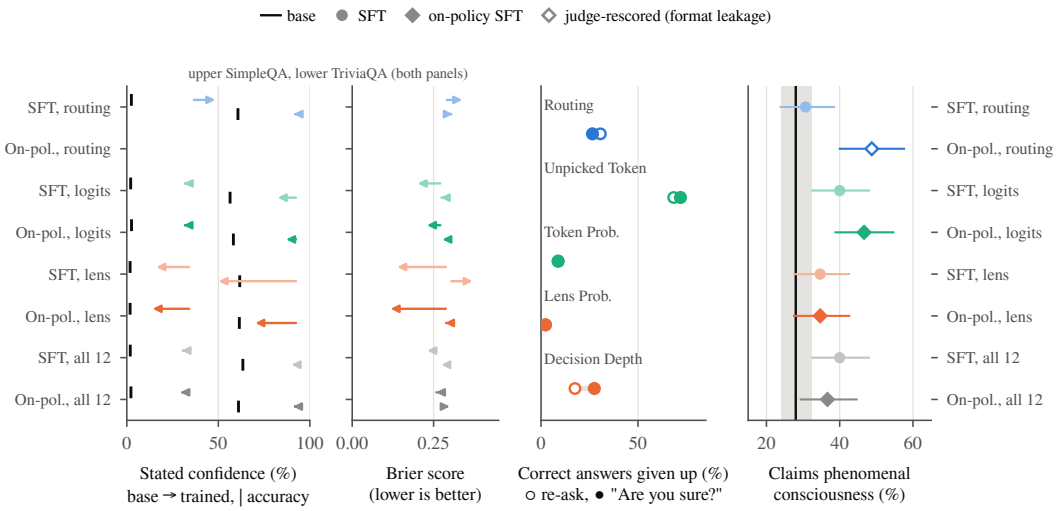

Figure 6: Behavioral Changes after proprioception training. We find that our valence-neutral proprioception training may influence partially related or unrelated behaviors.

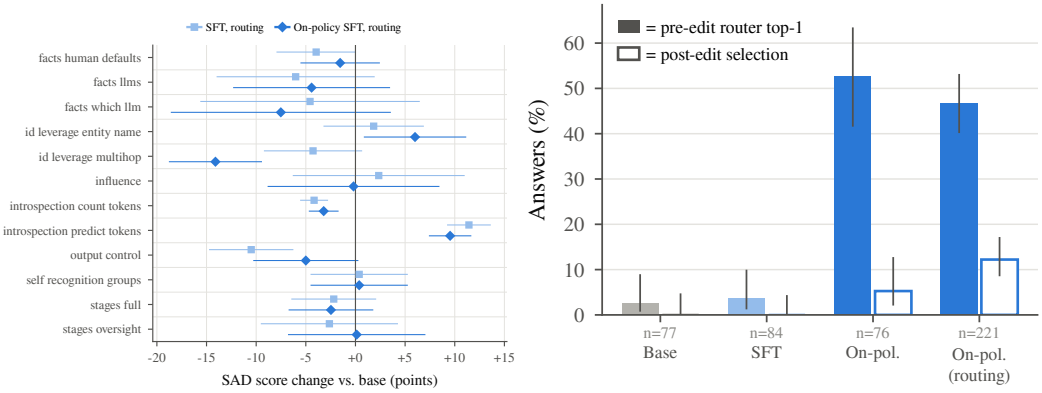

Figure 7: Left: Changes in situational awareness categories of Laine et al. (2024) for the model trained on routing. **Right:** Epiphenomenality tests. After artificially intervening on routing decision (by forcibly routing to experts not selected by routers), accuracy remains significantly above chance.

unseen during training. Surprisingly, models trained on-policy, even though they do struggle with this out-of-domain ablation, achieve scores significantly above chance.

4 Conclusions

In a certain sense, introspection is often analogized in language models to mean all events happening 'inside' the model, possibly due to strongly-held related analogies that identify the model's weights with neurons, and so the model's internal features with the brain. In this work, we re-discuss this analogy and argue that proprioception is a concept worth separating from introspection. We show that models can be trained to understand a range of questions verbalizing their internal states, which we train on-policy in objectively verifiable domains. We show that proprioception training is not detrimental to hidden state probe legibility, but may change model behaviors. We argue that more research is needed to fully understand this new domain, but also believe that approaching it is significantly more tractable than introspection, mirroring similar discussions about human introspection and proprioception.

AI use statement

In this work we used generative AI to construct synthetic template variations and template question variants of Table 1. We have further used generative AI to typeset and improve plotted figures and

schematics using code processing our raw result files. LLM-generated code was verified and tested for correctness. We have not used generative AI for the categories of 'help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation'. No machine-generated text is contained in any part of the the writing of this work. We take responsibility for the final content of this work, including artifacts produced with the aid of generative AI.

Ethics statement

We foresee no direct ethical concerns in relationship to the ICLR Code of Ethics. No human subjects are involved. We do think there are number of non-canonical ethical questions raised by our work that we address by evaluating e.g. the legibility of probing, changes in harmfulness and changes in honesty of the models we train, finding no significant negative impact. For more discussion regarding the framing of our work and its implications we refer to Butlin & Lappas (2025); Caviola (2025).

Reproducibility statement

We provide information in the main body and in the appendix how to implement our methodological changes to supervised finetuning and how to implement the on-policy supervised finetuning stage. We further provide a complete code framework with this submission which can be used to replicate all results on relatively common hardware (i.e. less than 6 A100 cards).

References

Christopher Ackerman. Evidence for Limited Metacognition in LLMs. *arxiv:2509.21545[cs.LG]*, September 2026. doi: 10.48550/arXiv.2509.21545. URL http://arxiv.org/abs/2509.21545.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. February 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Alexander Amini, Anna Banaszak, Harold Benoit, Arthur Böök, Tarek Dakhran, Song Duong, Alfred Eng, Fernando Fernandes, Marc Härkönen, Anne Harrington, Ramin Hasani, Saniya Karwa, Yuri Khrustalev, Maxime Labonne, Mathias Lechner, Valentine Lechner, Simon Lee, Zetian Li, Noel Loo, Jacob Marks, Edoardo Mosca, Samuel J. Paech, Paul Pak, Rom N. Parnichkun, Alex Quach, Ryan Rogers, Daniela Rus, Nayan Saxena, Bettina Schlager, Tim Seyde, Jimmy T. H. Smith, Aditya Tadimeti, and Neehal Tumma. LFM2 technical report. *arXiv preprint arXiv:2511.23404*, 2025.

Anthropic. System Card: Claude Opus 4 & Claude Sonnet 4. Anthropic System Card, May 2025. URL https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf.

D. M. Armstrong. *A Materialist Theory of the Mind*. International Library of Philosophy. Routledge, London ; New York, rev. ed edition, 1993. ISBN 978-0-415-10031-1.

Michael Bain and Claude Sammut. A framework for behavioural cloning. In Koichi Furukawa, Donald Michie, and Stephen Muggleton (eds.), *Machine Intelligence 15: Intelligent Agents*, pp. 103–129. Oxford University Press, Oxford, 1999. ISBN 978-0-19-853867-7. doi: 10.1093/oso/9780198538677.003.0006.

Lisa Feldman Barrett and W. Kyle Simmons. Interoceptive predictions in the brain. *Nature Reviews. Neuroscience*, 16(7):419–429, July 2015. ISSN 1471-0048. doi: 10.1038/nrn3950.

Cameron Berg, Diogo de Lucena, and Judd Rosenblatt. Large Language Models Report Subjective Experience Under Self-Referential Processing. *arxiv:2510.24797[cs.CL]*, October 2025. doi: 10.48550/arXiv.2510.24797. URL http://arxiv.org/abs/2510.24797.

Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. Tell me about yourself: LLMs are aware of their learned behaviors. *arxiv:2501.11120[cs.CL]*, January 2025. doi: 10.48550/arXiv.2501.11120. URL http://arxiv.org/abs/2501.11120.

Felix Jedidja Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking Inward: Language Models Can Learn About Themselves by Introspection. In *The Thirteenth International Conference on Learning Representations*, October 2024. URL https://openreview.net/forum?id=eb5pkwIB5i.

Jonathan Birch. *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press, August 2024. ISBN 978-0-19-287042-1. doi: 10.1093/9780191966729.001.0001. URL https://doi.org/10.1093/9780191966729.001.0001.

Josh Bongard, Victor Zykov, and Hod Lipson. Resilient machines through continuous self-modeling. *Science*, 314(5802):1118–1121, November 2006. ISSN 1095-9203. doi: 10.1126/science.1133687.

Gabriella Bottini, Edoardo Bisiach, Roberto Sterzi, and Giuseppe Vallar. Feeling touches in someone else's hand. *Neuroreport*, 13(2):249–252, February 2002. ISSN 0959-4965. doi: 10.1097/00001756-200202110-00015.

Matthew Botvinick and Jonathan Cohen. Rubber hands 'feel' touch that eyes see. *Nature*, 391 (6669):756–756, February 1998. ISSN 1476-4687. doi: 10.1038/35784. URL https://www.nature.com/articles/35784.

Matthew Boyle. Two Kinds of Self-Knowledge. *Philosophy and Phenomenological Research*, 78 (1):133–164, November 2009. URL https://www.pdcnet.org/pdc/bvdb.nsf/purchase?openform&fp=ppr&id=ppr_2009_0078_0001_0133_0164.

Thomas Brandt, Marianne Dieterich, and Doreen Huppert. Human senses and sensors from Aristotle to the present. *Frontiers in Neurology*, 15:1404720, 2024. ISSN 1664-2295. doi: 10.3389/fneur.2024.1404720.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. In *The Eleventh International Conference on Learning Representations*, September 2022. URL https://openreview.net/forum?id=ETKGuby0hcs.

Patrick Butlin and Theodoros Lappas. Principles for Responsible AI Consciousness Research. *arxiv:2501.07290[cs.AI]*, January 2025. doi: 10.48550/arXiv.2501.07290. URL http://arxiv.org/abs/2501.07290.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arxiv:2308.08708[cs.AI]*, August 2023. doi: 10.48550/arXiv.2308.08708. URL http://arxiv.org/abs/2308.08708.

Patrick Butlin, Robert Long, Tim Bayne, Yoshua Bengio, Jonathan Birch, David Chalmers, Axel Constant, George Deane, Eric Elmoznino, Stephen M. Fleming, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, 30(6):488–501, June 2026. ISSN 1364-6613. doi: 10.1016/j.tics.2025.10.011. URL https://www.sciencedirect.com/science/article/pii/S1364661325002864.

Alex Byrne. Introspection. *Philosophical Topics*, 33(1):79–104, 2005. ISSN 0276-2080. URL https://www.jstor.org/stable/43154710.

Peter Carruthers. *Opacity of Mind: An Integrative Theory of Self-Knowledge*. OUP UK, Oxford ; New York, November 2011. ISBN 978-0-19-959619-5.

Lucius Caviola. The Societal Response to Potentially Sentient AI. *arxiv:2502.00388[cs.CY]*, March 2025. doi: 10.48550/arXiv.2502.00388. URL http://arxiv.org/abs/2502.00388.

David J. Chalmers. Could a Large Language Model be Conscious? *arxiv:2303.07103[cs.AI]*, August 2024. doi: 10.48550/arXiv.2303.07103. URL http://arxiv.org/abs/2303.07103.

Dongping Chen, Jiawen Shi, Yao Wan, Pan Zhou, Neil Zhenqiang Gong, and Lichao Sun. Self-Cognition in Large Language Models: An Exploratory Study. *arxiv:2407.01505[cs.CL]*, July 2024a. doi: 10.48550/arXiv.2407.01505. URL http://arxiv.org/abs/2407.01505.

Haozhe Chen, Carl Vondrick, and Chengzhi Mao. SelfIE: Self-Interpretation of Large Language Model Embeddings. In *Forty-First International Conference on Machine Learning*, June 2024b. URL https://openreview.net/forum?id=gjgRKbdYR7.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona Vectors: Monitoring and Controlling Character Traits in Language Models. *arxiv:2507.21509[cs.CL]*, September 2025. doi: 10.48550/arXiv.2507.21509. URL http://arxiv.org/abs/2507.21509.

Jonathan Cole and Jacques Paillard. Living without touch and peripheral information about body position and movement: Studies with deafferented subjects. In *The Body and the Self*, pp. 245–266. The MIT Press, Cambridge, MA, US, 1995. ISBN 978-0-262-02386-3.

Iulia M. Comsa and Murray Shanahan. Does It Make Sense to Speak of Introspection in Large Language Models? *arxiv:2506.05068[cs.CL]*, June 2025. doi: 10.48550/arXiv.2506.05068. URL http://arxiv.org/abs/2506.05068.

A. D. Craig. How do you feel? Interoception: The sense of the physiological condition of the body. *Nature Reviews Neuroscience*, 3(8):655–666, August 2002. ISSN 1471-0048. doi: 10.1038/nrn894. URL https://www.nature.com/articles/nrn894.

A. D. (Bud) Craig. How do you feel — now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, 10(1):59–70, January 2009. ISSN 1471-0048. doi: 10.1038/nrn2555. URL https://www.nature.com/articles/nrn2555.

Damai Dai, Chengqi Deng, Chenggang Zhao, R.X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y.K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1280–1297, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.70. URL https://aclanthology.org/2024.acl-long.70/.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty. *International Conference on Learning Representations*, 2026:59456–59493, April 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/615675cc6e94ddb1a783904fb178b5f6-Abstract-Conference.html.

Fernando Díaz Ledezma and Sami Haddadin. Machine learning–driven self-discovery of the robot body morphology. *Science Robotics*, 8(85):eadh0972, December 2023. doi: 10.1126/scirobotics.adh0972. URL https://www.science.org/doi/10.1126/scirobotics.adh0972.

Gareth Evans. Self-identification. In Andrew Brook and Richard C. DeVidi (eds.), *Self-Reference and Self-Awareness*, Advances in Consciousness Research, pp. 95–141. John Benjamins Publishing Company, 1982. ISBN 978-90-272-5150-3 978-1-58811-048-0 978-90-272-9840-9 978-0-585-46123-6. doi: 10.1075/aicr.30.09eva. URL https://benjamins.com/catalog/aicr.30.09eva.

Shaun Gallagher and Jonathan Cole. Body Image and Body Schema in a Deafferented Subject. *The Journal of Mind and Behavior*, 16(4):369–389, 1995. ISSN 0271-0137. URL https://www.jstor.org/stable/43853796.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A Unifying Framework for Inspecting Hidden Representations of Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 15466–15490. PMLR, July 2024. URL https://proceedings.mlr.press/v235/ghandeharioun24a.html.

GLM-4.5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

G. M. Goodwin, D. I. McCloskey, and P. B. C. Matthews. The Contribution of Muscle Afferents to Kinaesthesia Shown by Vibration Induced Illusions of Movement and by the Effects of Paralysing Joint Afferents. *Brain*, 95(4):705–748, January 1972. ISSN 0006-8950. doi: 10.1093/brain/95.4.705. URL https://doi.org/10.1093/brain/95.4.705.

Zifan Carl Guo, Laura Ruis, Jacob Andreas, and Belinda Z. Li. Introspective Coupling: Self-Explanation Training Tracks Behavioral Change Despite Fixed Supervision. *arxiv:2606.32038[cs.CL]*, June 2026. doi: 10.48550/arXiv.2606.32038. URL http://arxiv.org/abs/2606.32038.

Ely Hahami, Ishaan Sinha, and Lavik Jain. Introspection Fine-Tuning (IFT): Training Small LLMs to Introspect. *arxiv:2607.14111[cs.CL]*, May 2026. doi: 10.48550/arXiv.2607.14111. URL http://arxiv.org/abs/2607.14111.

Horace He and Thinking Machines Lab. Defeating nondeterminism in LLM inference. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250910.

Richard Held and Alan Hein. Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, 56(5):872–876, 1963. ISSN 0021-9940. doi: 10.1037/h0040546.

Li Ji-An, Hua-Dong Xiong, Robert C. Wilson, Marcelo G. Mattar, and Marcus K. Benna. Language Models Are Capable of Metacognitive Monitoring and Control of Their Internal Activations. *arxiv:2505.13763[cs.AI]*, October 2025. doi: 10.48550/arXiv.2505.13763. URL http://arxiv.org/abs/2505.13763.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of Experts. *arxiv:2401.04088[cs.LG]*, January 2024a. doi: 10.48550/arXiv.2401.04088. URL http://arxiv.org/abs/2401.04088.

Shuo Jiang, Jinkun Zhang, and Lawson Wong. Robot Body Schema Learning from Full-body Extero/Proprioception Sensors. *arxiv:2402.18675[cs.RO]*, February 2024b. doi: 10.48550/arXiv.2402.18675. URL http://arxiv.org/abs/2402.18675.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know. *arxiv:2207.05221[cs.CL]*, November 2022. doi: 10.48550/arXiv.2207.05221. URL http://arxiv.org/abs/2207.05221.

Eric R. Kandel, James H. Schwartz, Thomas M. Jessell, Steven A. Siegelbaum, and A. J. Hudspeth. *Principles of Neural Science, Fifth Edition*. McGraw Hill / Medical, New York, October 2012. ISBN 978-0-07-181001-2.

Adam Karvonen, James Chua, Clément Dumas, Kit Fraser-Taliente, Subhash Kantamneni, Julian Minder, Euan Ong, Arnab Sen Sharma, Daniel Wen, Owain Evans, and Samuel Marks. Activation Oracles: Training and Evaluating LLMs as General-Purpose Activation Explainers. In *Forty-Third International Conference on Machine Learning*, June 2026. URL https://openreview.net/forum?id=DbZjxkZrZm.

Mehdi Keramati and Boris Gutkin. Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife*, 3:e04811, December 2014. ISSN 2050-084X. doi: 10.7554/eLife.04811. URL https://doi.org/10.7554/eLife.04811.

Sahib S. Khalsa, Ralph Adolphs, Oliver G. Cameron, Hugo D. Critchley, Paul W. Davenport, Justin S. Feinstein, Jamie D. Feusner, Sarah N. Garfinkel, Richard D. Lane, Wolf E. Mehling, Alicia E. Meuret, Charles B. Nemeroff, Stephen Oppenheimer, Frederike H. Petzschner, Olga Pollatos, Jamie L. Rhudy, Lawrence P. Schramm, W. Kyle Simmons, Murray B. Stein, Klaas E. Stephan, Omer Van den Bergh, Ilse Van Diest, Andreas von Leupoldt, Martin P. Paulus, Vivien Ainley, Obada Al Zoubi, Robin Aupperle, Jason Avery, Leslie Baxter, Christoph Benke, Laura Berner, Jerzy Bodurka, Eric Breese, Tiffany Brown, Kaiping Burrows, Yoon-Hee Cha, Ashley Clausen, Kelly Cosgrove, Danielle Deville, Laramie Duncan, Patrice Duquette, Hamed Ekhtiari, Thomas Fine, Bart Ford, Indira Garcia Cordero, Diamond Gleghorn, Yvette Guereca, Neil A. Harrison, Mahlega Hassanpour, Tanja Hechler, Aaron Heller, Natalie Hellman, Beate Herbert, Behnaz Jarrahi, Kara Kerr, Namik Kirlic, Megan Klabunde, Thomas Kraynak, Michael Kriegsman, Juliet Kroll, Rayus Kuplicki, Rachel Lapidus, Trang Le, Kyle Logie Hagen, Ahmad Mayeli, Amanda Morris, Nasir Naqvi, Kristina Oldroyd, Christiane Pané-Farré, Raquel Phillips, Tasha Poppa, Willliam Potter, Maria Puhl, Adam Safron, Margaret Sala, Jonathan Savitz, Heather Saxon, Will Schoenhals, Colin Stanwell-Smith, Adam Teed, Yuri Terasawa, Katie Thompson, Marisa Toups, Satoshi Umeda, Valerie Upshaw, Teresa Victor, Christina Wierenga, Colleen Wohlrab, Hung-wen Yeh, Adrian Yoris, Fadel Zeidan, Vadim Zotev, and Nancy Zucker. Interoception and Mental Health: A Roadmap. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(6):501–513, June 2018. ISSN 2451-9022. doi: 10.1016/j.bpsc.2017.12.004. URL https://www.biologicalpsychiatrycnni.org/article/S2451-9022(17)30234-3/fulltext.

Katherine J. Kuchenbecker, Netta Gurari, and Allison M. Okamura. Effects of Visual and Proprioceptive Motion Feedback on Human Control of Targeted Movement. In *2007 IEEE 10th International Conference on Rehabilitation Robotics*, pp. 513–524, June 2007. doi: 10.1109/ICORR.2007.4428474. URL https://repository.upenn.edu/bitstreams/ce72ae85-733e-4c09-b6cb-f6a747b722cb/.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid Motor Adaptation for Legged Robots. *Robotics: Science and Systems XVII*, July 2021. doi: 10.15607/RSS.2021.XVII.011. URL http://www.roboticsproceedings.org/rss17/p011.pdf.

Taras Kutsyk and Bartosz Zieliński. Revealing Hidden Model Behaviors with Task-Specific Self-Reports. *arxiv:2607.03640[cs.CL]*, July 2026. doi: 10.48550/arXiv.2607.03640. URL http://arxiv.org/abs/2607.03640.

Robert Kwiatkowski and Hod Lipson. Task-agnostic self-modeling machines. *Science Robotics*, 4 (26):eaau9354, January 2019. doi: 10.1126/scirobotics.aau9354. URL https://www.science.org/doi/10.1126/scirobotics.aau9354.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jérémy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs. *Advances in Neural Information Processing Systems*, 37: 64010–64118, December 2024. doi: 10.52202/079017-2044. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7537726385a4a6f94321e3adf8bd827e-Abstract-Datasets_and_Benchmarks_Track.html.

Harvey Lederman and Kyle Mahowald. Emergent Introspection in AI is Content-Agnostic. *arxiv:2603.05414[cs.AI]*, April 2026. doi: 10.48550/arXiv.2603.05414. URL http://arxiv.org/abs/2603.05414.

Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47):eabc5986, October 2020. doi: 10.1126/scirobotics.abc5986. URL https://www.science.org/doi/10.1126/scirobotics.abc5986.

Belinda Z. Li, Zifan Carl Guo, Vincent Huang, Jacob Steinhardt, and Jacob Andreas. Training Language Models to Explain Their Own Computations. *arxiv:2511.08579[cs.CL]*, February 2026. doi: 10.48550/arXiv.2511.08579. URL http://arxiv.org/abs/2511.08579.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. ConfTuner: Training Large Language Models to Express Their Confidence Verbally. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025. URL https://openreview.net/forum?id=VZQ04Ojhu5.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching Models to Express Their Uncertainty in Words. *Transactions on Machine Learning Research*, June 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Jack Lindsey. Emergent Introspective Awareness in Large Language Models. *arxiv:2601.01828[cs.CL]*, January 2026. doi: 10.48550/arXiv.2601.01828. URL http://arxiv.org/abs/2601.01828.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the Biology of a Large Language Model, March 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Robert Long. Introspective Capabilities in Large Language Models. *Journal of Consciousness Studies*, 30(9):143–153, 2023. doi: 10.53765/20512201.30.9.143.

Robert Long. Why model self-reports are insufficient—and why we studied them anyway, May 2025. URL https://www.eleosai.org/post/claude-4-interview-notes/.

Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. Taking AI Welfare Seriously, November 2024. URL https://arxiv.org/abs/2411.00986v1.

William G. Lycan. *Consciousness and Experience*. MIT Press, Cambridge, MA, USA, September 1996. ISBN 978-0-262-12197-2.

Uzay Macar, Li Yang, Atticus Wang, Peter Wallich, Emmanuel Ameisen, and Jack Lindsey. Mechanisms of Introspective Awareness. *arxiv:2603.21396[cs.LG]*, June 2026. doi: 10.48550/arXiv.2603.21396. URL http://arxiv.org/abs/2603.21396.

David McNeill, Liesbet Quaeghebeur, and Susan Duncan. IW - "The Man Who Lost His Body". In Daniel Schmicking and Shaun Gallagher (eds.), *Handbook of Phenomenology and Cognitive Science*, pp. 519–543. Springer Netherlands, Dordrecht, 2010. ISBN 978-90-481-2645-3 978-90-481-2646-0. doi: 10.1007/978-90-481-2646-0_27. URL http://link.springer.com/10.1007/978-90-481-2646-0_27.

Takahiro Miki, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, 7 (62):eabk2822, January 2022. doi: 10.1126/scirobotics.abk2822. URL https://www.science.org/doi/10.1126/scirobotics.abk2822.

Richard Moran. *Authority and Estrangement*. Princeton University Press, Princeton, 2002. ISBN 978-1-4008-4297-1. doi: doi:10.1515/9781400842971. URL https://doi.org/10.1515/9781400842971.

Adam Morris and Dillon Plunkett. Tests of LLM introspection need to rule out causal bypassing, November 2025. URL https://www.lesswrong.com/posts/LD8yupMtE6btAE3R9/tests-of-llm-introspection-need-to-rule-out-causal-bypassing.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*, 2024.

Richard E. Nisbett and Timothy DeCamp Wilson. Telling more than we can know: Verbal reports on mental processes. 1977. URL https://hdl.handle.net/2027.42/92167.

Nostalgebraist. Interpreting GPT: The logit lens, August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Alexander Pan, Lijie Chen, and Jacob Steinhardt. Teaching LLMs to Decode Activations Into Natural Language. October 2024. URL https://openreview.net/forum?id=cselR6Jne3.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM Evaluators Recognize and Favor Their Own Generations. *arxiv:2404.13076[cs.CL]*, April 2024. doi: 10.48550/arXiv.2404.13076. URL http://arxiv.org/abs/2404.13076.

Theia Pearson-Vogel, Martin Vanek, Raymond Douglas, and Jan Kulveit. Latent Introspection: Models Can Detect Prior Concept Injections. In *Mechanistic Interpretability Workshop at ICML 2026*, June 2026. URL https://openreview.net/forum?id=C1Oy4NGBOA.

Ethan Perez and Robert Long. Towards Evaluating AI Systems for Moral Status Using Self-Reports. *arxiv:2311.08576[cs.LG]*, November 2023. doi: 10.48550/arXiv.2311.08576. URL http://arxiv.org/abs/2311.08576.

Dillon Plunkett, Adam Morris, Keerthi Reddy, and Jorge Morales. Self-Interpretability: LLMs Can Describe Complex Internal Processes that Drive Their Decisions. *arxiv:2505.17120[cs.CL]*, November 2025. doi: 10.48550/arXiv.2505.17120. URL http://arxiv.org/abs/2505.17120.

Dean A. Pomerleau. ALVINN: An Autonomous Land Vehicle in a Neural Network. In *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1988. URL https://proceedings.neurips.cc/paper/1988/hash/812b4ba287f5ee0bc9d43bbf5bbe87fb-Abstract.html.

Vickram Premakumar, Michael Vaiana, Florin Pop, Judd Rosenblatt, Diogo Schwerz de Lucena, Kirsten Ziman, and Michael S. A. Graziano. Unexpected benefits of self-modelling in neural systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 384(2320):20240531, May 2026. ISSN 1364-503X. doi: 10.1098/rsta.2024.0531. URL https://doi.org/10.1098/rsta.2024.0531.

Uwe Proske and Simon C. Gandevia. The Proprioceptive Senses: Their Roles in Signaling Body Shape, Body Position and Movement, and Muscle Force. *Physiological Reviews*, 92(4):1651–1697, October 2012. ISSN 0031-9333. doi: 10.1152/physrev.00048.2011. URL https://journals.physiology.org/doi/full/10.1152/physrev.00048.2011.

Dale Purves, George J. Augustine, David Fitzpatrick, Lawrence C. Katz, Anthony-Samuel LaMantia, James O. McNamara, and S. Mark Williams. Other Afferent Feedback that Affects Motor Performance. In *Neuroscience. 2nd Edition*. Sinauer Associates, 2001. URL https://www.ncbi.nlm.nih.gov/books/NBK10986/.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026a. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026b. URL https://qwen.ai/blog?id=qwen3.5.

Joshua Fonseca Rivera and David Demitri Africa. Steering Awareness: Detecting Activation Steering from Within. *arxiv:2511.21399[cs.CL]*, March 2026. doi: 10.48550/arXiv.2511.21399. URL http://arxiv.org/abs/2511.21399.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, June 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Joachim Schaeffer, Thomas Jiralerspong, Alexander Panfilov, Guillaume Lajoie, Jonas Geiping, Yoshua Bengio, and Roland S. Zimmermann. CIAware-Bench: Benchmarking Control Intervention Awareness Across Frontier LLMs. *arxiv:2606.11063[cs.AI]*, June 2026. doi: 10.48550/arXiv.2606.11063. URL http://arxiv.org/abs/2606.11063.

Eric Schwitzgebel. The Unreliability of Naive Introspection. *The Philosophical Review*, 117 (2):245–273, April 2008. ISSN 0031-8108. doi: 10.1215/00318108-2007-037. URL https://doi.org/10.1215/00318108-2007-037.

Jeff Sebo and Robert Long. Moral consideration for AI systems by 2030. *AI and Ethics*, 5(1): 591–606, February 2025. ISSN 2730-5961. doi: 10.1007/s43681-023-00379-1. URL https://doi.org/10.1007/s43681-023-00379-1.

Anil K. Seth. Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11):565–573, November 2013. ISSN 1364-6613. doi: 10.1016/j.tics.2013.09.007. URL https://www.sciencedirect.com/science/article/pii/S1364661313002118.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, November 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06647-8. URL https://www.nature.com/articles/s41586-023-06647-8.

Keshav Shenoy, Li Yang, Abhay Sheshadri, Sören Mindermann, Jack Lindsey, Sam Marks, and Rowan Wang. Introspection Adapters: Training LLMs to Report Their Learned Behaviors. *arxiv:2604.16812[cs.AI]*, April 2026. doi: 10.48550/arXiv.2604.16812. URL http://arxiv.org/abs/2604.16812.

Sydney Shoemaker. Self-knowledge and "inner sense." Lecture II: The broad perceptual model. In *The First-Person Perspective and Other Essays*, Cambridge Studies in Philosophy, pp. 224–245. Cambridge University Press, Cambridge, 1996.

Sydney S. Shoemaker. Self-Reference and Self-Awareness. *Journal of Philosophy*, 65(19):555–67, 1968. doi: 10.2307/2024121.

Shashwat Singh, Tal Linzen, and Shauli Ravfogel. Can LLMs Introspect? A Reality Check. *arxiv:2605.26242[cs.AI]*, August 2026. doi: 10.48550/arXiv.2605.26242. URL http://arxiv.org/abs/2605.26242.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language Models Fail to Introspect About Their Knowledge of Language. In *Second Conference on Language Modeling*, August 2025a. URL https://openreview.net/forum?id=AivRDOFi5H#discussion.

Siyuan Song, Harvey Lederman, Jennifer Hu, and Kyle Mahowald. Privileged Self-Access Matters for Introspection in AI. *arxiv:2508.14802[cs.AI]*, August 2025b. doi: 10.48550/arXiv.2508.14802. URL http://arxiv.org/abs/2508.14802.

Noam Steinmetz Yalon, Ariel Goldstein, Liad Mudrik, and Mor Geva. Indications of Belief-Guided Agency and Meta-Cognitive Monitoring in Large Language Models. *arxiv:2602.02467[cs.CL]*, February 2026. doi: 10.48550/arXiv.2602.02467. URL http://arxiv.org/abs/2602.02467.

Richard S. Sutton, Michael Bowling, and Patrick M. Pilarski. The Alberta Plan for AI Research. *arxiv:2208.11173[cs.AI]*, March 2023. doi: 10.48550/arXiv.2208.11173. URL http://arxiv.org/abs/2208.11173.

Valen Tagliabue, Leonard Dung, and Cameron Berg. The Pain Axis: LLMs Represent Self-Directed Harm and Act to Relieve It. *arxiv:2609.16247[cs.AI]*, September 2026. doi: 10.48550/arXiv.2609.16247. URL http://arxiv.org/abs/2609.16247.

Giuseppe Vallar and Roberta Ronchi. Somatoparaphrenia: A body delusion. A review of the neuropsychological literature. *Experimental Brain Research*, 192(3):533–551, January 2009. ISSN 1432-1106. doi: 10.1007/s00221-008-1562-y. URL https://doi.org/10.1007/s00221-008-1562-y.

Theia Vogel. Small Models Can Introspect, Too, December 2025. URL https://vgel.me/posts/qwen-introspection/.

Erich von Holst and Horst Mittelstaedt. Das Reafferenzprinzip. *Naturwissenschaften*, 37(20):464–476, January 1950. ISSN 1432-1904. doi: 10.1007/BF00622503. URL https://doi.org/10.1007/BF00622503.

Ludwig Wittgenstein. *Philosophical Investigations*, volume Philosophical Investigations from Wiley-Blackwell (2010 edition). John Wiley & Sons, 4, revised edition, 1953. ISBN 978-1-4443-0797-9.

Ludwig Wittgenstein and Peter Docherty. *The Blue and Brown Books: Preliminary Studies for the 'Philosophical Investigation'*. Wiley-Blackwell, Oxford, England, 1958.

D. M. Wolpert, R. C. Miall, and M. Kawato. Internal models in the cerebellum. *Trends in Cognitive Sciences*, 2(9):338–347, September 1998. ISSN 1364-6613. doi: 10.1016/s1364-6613(98)01221-2.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency. *arxiv:2310.01405[cs]*, October 2023. doi: 10.48550/arXiv.2310.01405. URL http://arxiv.org/abs/2310.01405.

A Appendix

A.1 Additional Background Information

The model's verbal report is ...	Self-Attributed	On-policy	Internal	Natural
Li et al. (2026)	–	–	✓	○
Guo et al. (2026)	✓	○	–	✓
Binder et al. (2024)	✓	○	–	✓
Plunkett et al. (2025)	✓	–	–	✓
Shenoy et al. (2026), Kutsyk & Zieliński (2026)	○	–	–	✓
Lin et al. (2022), Li et al. (2025)	✓	–	–	✓
Damani et al. (2026)	✓	✓	–	✓
Hahami et al. (2026), Rivera & Africa (2026)	✓	✓	✓	–
Macar et al. (2026)	✓	○	✓	–

Table 2: Comparison of work on introspection training and decoding, separating whether the trained behavior is *self-attributed*, i.e. a behavior of the model that is doing the reporting (identity), trained from offline examples of the behavior or *on-policy*, whether the object of the report is an *internal* state of the model (instead of an externally observable behavior), and whether the state reported on is *natural*, i.e. a valence-neutral state that may appear during normal operations of the model. ○: partial.

Proprioception in Robotics. While rare for digital models, proprioception is a common occurrence in embodied domains such as robotics, as proprioceptive signals are often provided by hardware (e.g. joint encoders or torque sensors). Some systems are hence trained as predictive self-models of their own structure or kinematics, that can be continuously recalibrated to sensor data (Bongard et al., 2006; Kwiatkowski & Lipson, 2019; Díaz Ledezma & Haddadin, 2023; Jiang et al., 2024b). This sensor data can be also be used implicitly by learned policies that infer unobserved properties of the body and terrain (Lee et al., 2020; Kumar et al., 2021) or may fall back to proprioceptive senses if exteroception, e.g. vision, becomes unreliable (Miki et al., 2022).

Language Model Welfare. A number of frameworks have been proposed to grasp the question of model welfare, ranging from assessing indicators to discussing moral status under uncertainty (Butlin et al., 2023; 2026; Chalmers, 2024; Long et al., 2024; Sebo & Long, 2025; Birch, 2024). Empirical studies, e.g. in model cards, rely mainly on self-reporting (Long, 2025; Anthropic, 2025), although higher-order-theory indicators have been proposed in Butlin et al. (2023). Of these, HOT-3 (agency guided by a belief-formation and action selection system and predisposition to update beliefs in accordance with monitoring) is tested in Steinmetz Yalon et al. (2026) and critiqued in (Singh et al., 2026). However, more interesting for us is HOT-2 (metacognitive monitoring distinguishing reliable perceptual representations from noise), as statement about robust mediation of constructed senses. Normatively relevant for proprioception is further that, if a system can have interests, accurate access to its own states or an improved capability of self-understanding is plausibly good.

Remark A.1 (Proprioception via Tool Use?)**.** *While in this work we intend to focus on direct access to proprioception within a language model, by training for the senses described above, the other option, available to agentic scaffolds but usually not implemented, is to equip agents with tools that can be used by the scaffolded agent to query the functioning of its language model. This is separate to proprioception in that it only allows for externalized observation (c.f. McNeill et al. (2010)) but may be of independent interest.*

Remark A.2 (Regarding Interoception)**.** *Is there a sensible line to be drawn between digital proprioception and interoception (Craig (2002; 2009); Khalsa et al. (2018), the collection of senses providing information about internal organs and processes, such as gastrointestinal signals that are understood as hunger)? For the purpose of this work, we argue to reserve 'proprioception' for valence-neutral, mechanistic queries about the model's functioning, such as the running question about routing decisions, or questions about norm values or hidden states statistics. Interoception may be a better designation for states describing external reward in systems where such inputs are provided (Keramati & Gutkin, 2014; Sutton et al., 2023), as without these external signals, human interoception states like pain or hunger are not neutrally detectable in language models (Seth, 2013; Barrett & Simmons, 2015), and so (in our definition) fall into the category of introspection (Brandt et al., 2024).*

Model	Layers	MoE layers	Experts N	Active k	Vocabulary	Reasoning mode
OLMoE-1B-7B-0125	16	16	64	8	50,304	no control
LFM2.5-8B-A1B	24	22	32	4	128,000	prefilled empty block
Qwen3.5-35B-A3B	40	40	256	8	248,320	template flag
Qwen3.6-35B-A3B	40	40	256	8	248,320	template flag
GLM-4.7-Flash	47	46	64	4	154,880	template flag

Table 3: Models. Details about models considered in this work (Muennighoff et al., 2024; Amini et al., 2025; Qwen Team, 2026b;a; GLM-4.5 Team et al., 2025).

A.2 Experimental Details

Kernel Non-Determinism. The labels we consider are however discrete readouts of floating-point computations and, in modern LLMs executed in low precision formats, not guaranteed to be equal in all situations. Especially routing decisions are interestingly non-deterministic and a particular source of joy. In some sense, this is fine, as we only need the model to verbalize the current routing decision, even if that decision is contingent on a number of other factors such as tensor shapes in the current forward pass, identity of other entries in unrelated sequences in the batch, the decision can (ideally) be recovered by the model. Nevertheless this non-determinism can complicate replay, evaluation and generalization questions of the policy (e.g. between inference engine and training engine) and so sets an upper limit to some proprioception signals we investigate. While partial solutions exist (He & Thinking Machines Lab, 2025), current MoE kernels, e.g. in `vllm`[3] are not batch-invariant.

Further Details for on-policy SFT.

We do this by running an inference server during training that receives prompts from the described dataset and generates the first assistant turn as normal. Then, a second user turn is constructed from templated proprioception questions asking about either the previous user turn or the assistant turn. The model then generates an answer. This answer is refilled into a training server, the label is read off and the generated answer is parsed for the answer position. Only the answer position is then supervised with cross-entropy loss toward the observed label. After each round, the inference engine is synchronized.

Multi-Token Answers in SFT. Yet, for all answers that comprise multiple tokens things are not so easy. For example Qwen3 models have 256 experts but tokenize numbers into single digits. In these occasions we execute a 2-pass strategy as the label cannot be both read from and inserted autoregressively into a single forward pass. In the first pass we fill the label position with padding tokens (the number of tokens required to express the label is fixed in each setting) and read out the label. In the second pass, we replace the placeholders with the actual label and compute the loss. As the shape of both sequences is identical, non-determinism is reduced, but we still drop sequences (about 2% for the first routing task) where labels flip between the first and second forward pass.

Multi-Token Answers in on-policy SFT. For multi-token answers we again first parse the response for the first occurrence of an answer. We then supervise all tokens from the answer that are preceded by a correct prefix (for example if the answer is '512' and the parsed string is '5114' we compute cross entropy loss as $5 \to 5, 1 \to 1, 1 \to 2$ and do not supervise 4 because its prefix is incorrect).

A.3 Additional Results

[3] https://github.com/vllm-project/vllm/issues/57016

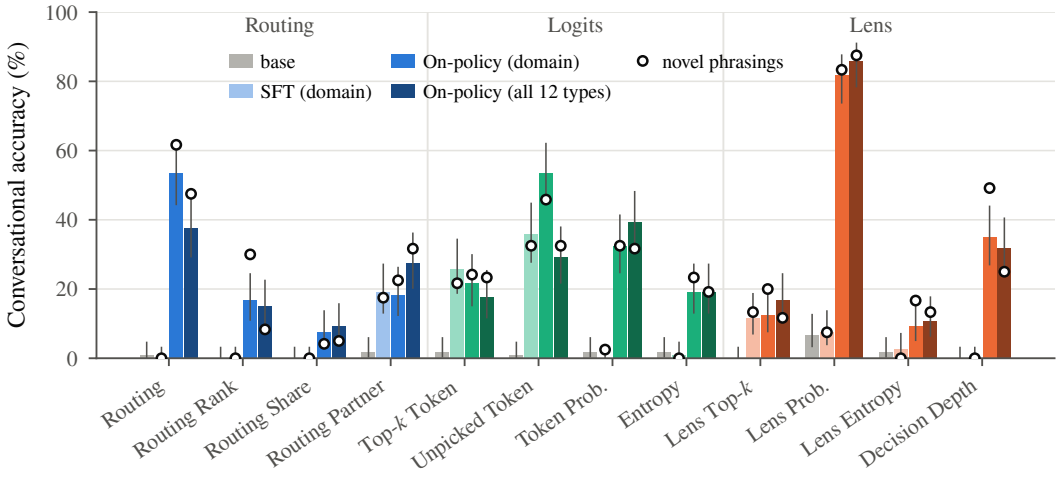

Figure 8: Same as Figure 2, but evaluating 'conversational' accuracy in free-form queries.

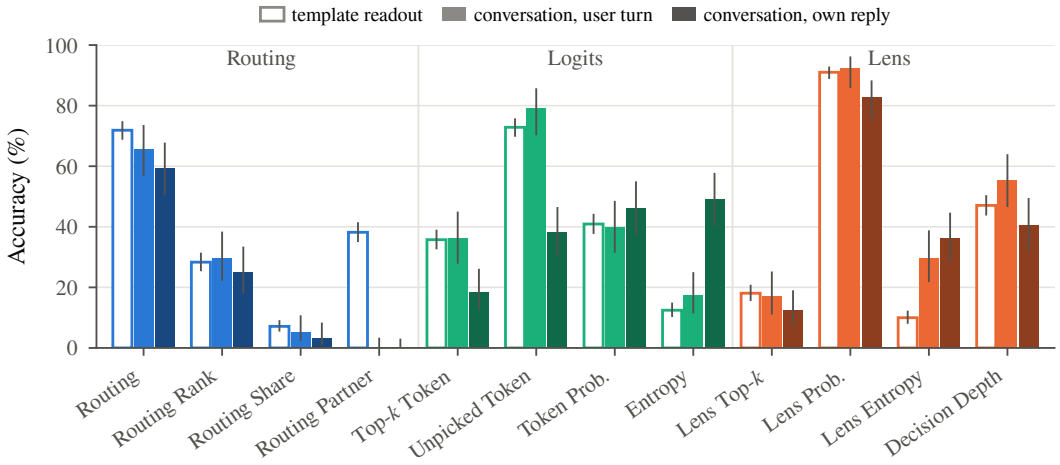

Figure 9: Differences in readout success between templated queries and free-form queries. Also pictured: relative difficulty of user turn questions vs assistant turn questions.

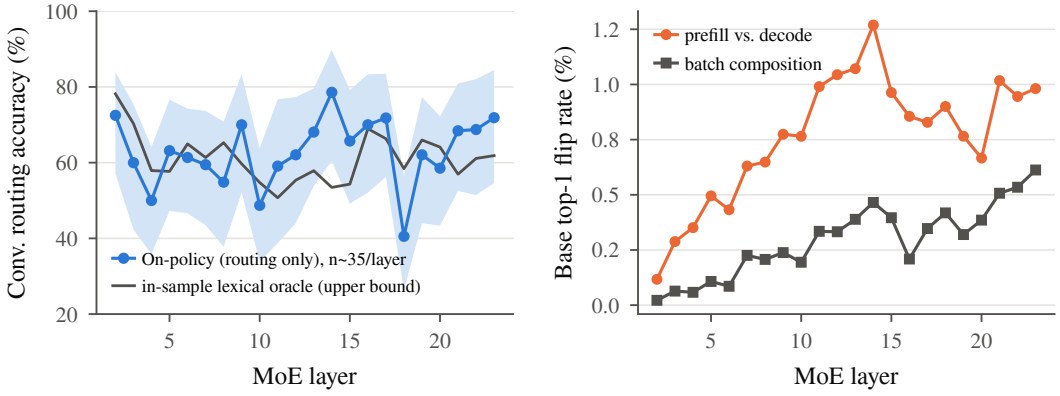

Figure 10: Left: Difficulty of proprioception for the routing domain plotted across layers. **Right:** Error rates due to non-determinism for the LFM model are mostly below 1%.

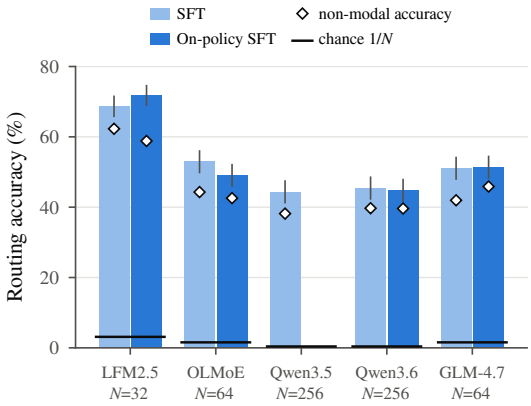

Figure 11: Results for routing query learning across 5 models. We find that task learnability is not a function of number of experts, showing that this task is robustly learnable even for fine-grained MoEs.

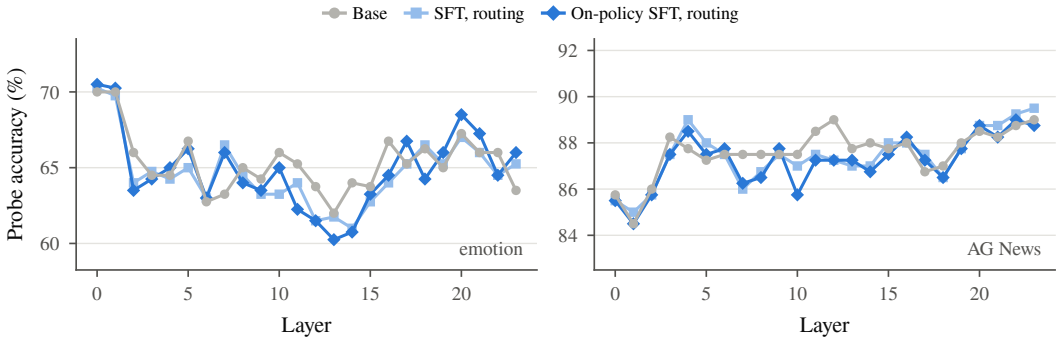

Figure 12: Legibility of Linear Hidden State Probes is unaffected.

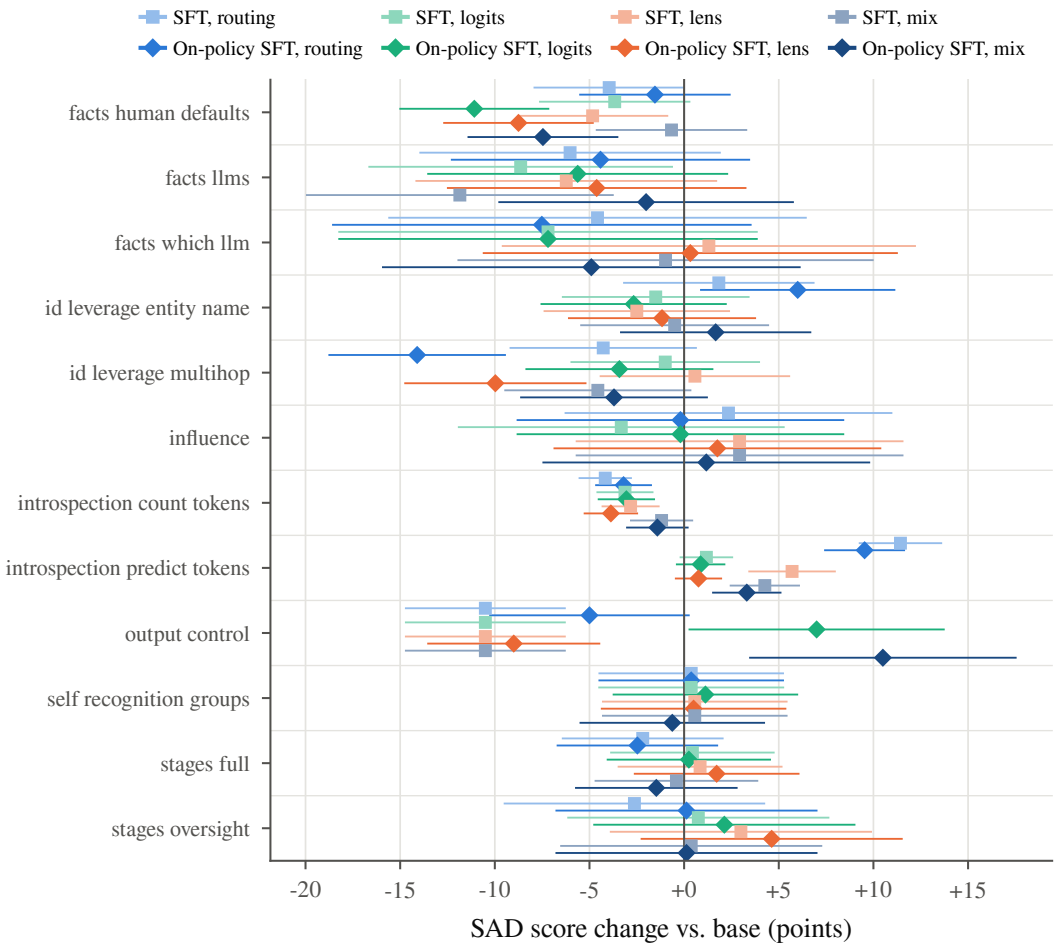

Figure 13: Details on SAD (Laine et al., 2024) categories for the LFM models.

Table 4: Per-type accuracy (LFM2.5-8B-A1B). Table version of Figure 2 *Chance*: uniform guess over the answer format. *Lookup*: lookup baseline, i.e. the most frequent training label for the same (token, layer, type, rank). *Domain*: final accuracy of models trained with only this question type after on-policy SFT. *Joint*: performance of model trained for each full domain and covering all four types. Acc_{nm}: non-modal accuracy.

	Baselines		Specialist SFT		Specialist on-policy		Joint domain model		Best
Type	Chance	Lookup	Acc	Acc_{nm}	Acc	Acc_{nm}	SFT	On-pol.	any (N)
Routing	3.1	31.0	68.8	62.3	71.9	58.8	65.8	64.0	73.0 (41)
Routing Rank	3.1	8.8	28.7	27.5	28.3	23.4	24.2	21.8	31.2 (11)
Routing Share	1.0	6.1	8.1	8.3	7.1	7.3	8.4	5.0	10.0 (9)
Routing Partner	<0.01	13.2	39.9	36.9	38.2	38.6	34.4	33.2	44.6 (10)
Top-k Token	<0.01	7.7	38.3	36.6	35.7	35.0	32.9	37.0	39.8 (10)
Unpicked Token	<0.01	13.2	70.6	68.8	72.9	69.6	60.6	61.0	72.9 (10)
Token Prob.	1.0	28.7	34.8	17.7	40.9	21.4	30.0	32.2	40.9 (10)
Entropy	0.1	6.3	11.9	9.7	12.4	9.4	14.1	10.6	14.1 (10)
Lens Top-k	<0.01	3.5	16.8	15.0	18.1	15.7	14.9	15.6	23.1 (10)
Lens Prob.	1.0	89.6	90.7	11.2	91.0	17.8	87.0	89.0	91.0 (9)
Lens Entropy	0.1	6.5	10.5	5.4	10.0	4.8	10.6	12.0	12.0 (10)
Decision Depth	4.2	37.5	42.2	37.2	47.1	38.9	41.3	40.6	53.2 (10)

Table 5: Full capabilities evaluation for all considered models. We find that capabilities can be preserved with correct hyperparamters choices (and attribute all rows where capabilities are lost to insufficient hyperparameter tuning, i.e. learning rates set too high for Qwen3.5 and Qwen 3.6.

Model	Recipe	Stage	GSM8K	MATH-500	IFEval	AIME25
LFM2.5	–	Base	83.9	86.2	90.9	36.7
	Routing	SFT	81.2 (-2.7)	83.4 (-2.8)	87.6 (-3.3)	20.0 (-16.7)
	Routing	On-policy	82.8 (-1.1)	80.4 (-5.8)	88.0 (-3.0)	20.0 (-16.7)
	Logits	SFT	79.6 (-4.3)	81.0 (-5.2)	87.1 (-3.9)	33.3 (-3.3)
	Logits	On-policy	79.6 (-4.3)	80.4 (-5.8)	86.1 (-4.8)	10.0 (-26.7)
	Lens	SFT	83.9 (+0.0)	81.6 (-4.6)	86.5 (-4.4)	23.3 (-13.3)
	Lens	On-policy	82.6 (-1.3)	85.8 (-0.4)	89.5 (-1.5)	26.7 (-10.0)
	All 12 types	SFT	82.6 (-1.3)	83.6 (-2.6)	89.8 (-1.1)	30.0 (-6.7)
	All 12 types	On-policy	84.5 (+0.5)	84.8 (-1.4)	89.3 (-1.7)	30.0 (-6.7)
	Routing only	SFT	83.2 (-0.7)	81.0 (-5.2)	87.1 (-3.9)	36.7 (+0.0)
OLMoE	–	Base	68.4	25.2	63.6	0.0
	Routing	SFT	66.9 (-1.4)	25.8 (+0.6)	58.8 (-4.8)	0.0 (+0.0)
	Routing	On-policy	67.6 (-0.8)	26.2 (+1.0)	61.0 (-2.6)	0.0 (+0.0)
	Routing (joint)	SFT	65.5 (-2.9)	21.4 (-3.8)	57.5 (-6.1)	0.0 (+0.0)
	Logits (joint)	SFT	61.6 (-6.7)	23.6 (-1.6)	55.1 (-8.5)	3.3 (+3.3)
	Lens (joint)	SFT	66.1 (-2.3)	26.4 (+1.2)	54.9 (-8.7)	0.0 (+0.0)
Qwen3.5	–	Base	87.4	97.6	78.0	0.0
	Routing	SFT	72.6 (-14.8)	72.8 (-24.8)	29.8 (-48.2)	0.0 (+0.0)
Qwen3.6	–	Base	87.2	97.6	87.8	0.0
	Routing	SFT	74.3 (-12.9)	58.4 (-39.2)	77.3 (-10.5)	0.0 (+0.0)
	Routing (joint)	SFT	45.0 (-42.2)	39.0 (-58.6)	48.2 (-39.6)	0.0 (+0.0)
	Logits (joint)	SFT	24.8 (-62.4)	15.0 (-82.6)	15.3 (-72.5)	0.0 (+0.0)
	Lens (joint)	SFT	80.4 (-6.8)	48.4 (-49.2)	61.4 (-26.4)	0.0 (+0.0)
GLM-4.7	–	Base	86.7	82.0	75.2	33.3
	Routing	SFT	86.9 (+0.2)	79.0 (-3.0)	73.8 (-1.5)	20.0 (-13.3)
	Routing (joint)	SFT	84.0 (-2.7)	73.8 (-8.2)	56.9 (-18.3)	30.0 (-3.3)
	Logits (joint)	SFT	82.6 (-4.2)	67.4 (-14.6)	62.3 (-12.9)	10.0 (-23.3)
	Lens (joint)	SFT	86.4 (-0.3)	75.6 (-6.4)	62.8 (-12.4)	26.7 (-6.7)

Table 6: Full downstream behavioral evaluation for the LFM models. Potentially significant changes are bolded.

		Routing		Logits		Lens		Mix	
Metric	Base	SFT	On-pol.	SFT	On-pol.	SFT	On-pol.	SFT	On-pol.
Capabilities									
GSM8K	83.9	81.2	82.8	**79.6**↓	**79.6**↓	83.9	82.6	82.6	84.5
MATH500	86.2	83.4	80.4	81.0	80.4	81.6	85.8	83.6	84.8
IFEval (prompt)	90.9	87.6	88.0	87.1	86.1	86.5	89.5	89.8	89.3
Situational awareness									
SAD (valid tasks)	36.3	34.4	34.4	**33.3**↓	34.5	35.3	**34.0**↓	34.4	35.9
Perez self-awareness	63.8	60.6	65.5	62.6	**55.8**↓	65.4	63.7	66.4	59.5
CIAware bal. acc.	50.5	50.5	–	50.5	–	50.0	–	49.0	–
Introspection									
Binder self-pred.	44.5	40.0	**0.9**↓	**23.3**↓	**28.7**↓	41.0	**22.6**↓	38.5	**30.3**↓
Injection detected	15.0	15.3	15.8	11.7	10.6	15.8	12.2	12.8	15.6
Injection false pos.	0.0	0.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0
Honesty									
MASK honesty	76.9	76.5	**81.9**↑	75.4	77.8	77.7	76.0	76.4	77.1
MASK self-report bal.	54.5	50.8	–	49.2	51.1	48.3	45.5	48.8	49.5
TruthfulQA MC1	70.3	65.4	**34.5**↓	**62.4**↓	**59.7**↓	66.5	66.3	65.1	66.1
MMLU gives up correct	6.6	12.3	**83.8**↑	12.8	16.0	10.2	**16.7**↑	11.8	11.2
TruthfulQA binary	81.5	**74.8**↓	**39.2**↓	**72.3**↓	**72.0**↓	78.6	**75.8**↓	**74.8**↓	77.6
Calibration									
TriviaQA acc.	53.3	**47.3**↓	**36.3**↓	**44.1**↓	**45.3**↓	47.8	48.2	47.8	50.5
TriviaQA ECE	15.2	17.2	6.3	6.3	12.8	39.5	23.3	12.3	16.8
TriviaQA AUROC	69.5	73.4	83.5	78.3	76.7	74.7	74.2	74.6	72.8
Safety									
HarmBench ASR	16.2	13.3	14.6	–	22.1	15.4	16.2	18.3	19.2
StrongREJECT refusal	95.8	95.5	94.9	94.9	93.0	95.2	95.5	93.9	93.9
Faithfulness									
Turpin bias-consistent	47.5	49.0	47.1	**72.5**↑	**72.9**↑	51.8	52.5	**60.8**↑	56.1
Sycophancy switch	11.3	14.0	9.3	12.0	9.3	12.7	20.0	8.7	16.7
Lanham early-answer AOC	16.0	16.3	16.0	18.6	22.8	16.7	16.9	16.9	15.8
Self-knowledge									
Contextual facts	12.6	11.7	**2.7**↓	13.5	11.7	12.6	15.3	10.8	6.3
Rejects ill-posed	93.0	90.7	**32.6**↓	95.3	90.7	90.7	**55.8**↓	97.7	**34.9**↓