Simonelli on Assertibility of pure LLMs

Injin Woo & Donggeon Kim

Talks & slides

2nd Philosophical Interchange between Sungkyunkwan and Nagoya University

Nagoya University, Japan

25 Oct 2026

Thesis: applying a common sufficient condition

Starting point

Simonelli proposes **mastery of inferential role** as sufficient for conceptual understanding in LLMs.

Our thesis

Applying this condition requires distinguishing **what each achievement requires**. The experiments show that **an achievement of one kind does not guarantee the mastery required for another**.

Thesis: applying a common sufficient condition

Starting point

Simonelli proposes **mastery of inferential role** as sufficient for conceptual understanding in LLMs.

Our thesis

Applying this condition requires distinguishing **what each achievement requires**. The experiments show that **an achievement of one kind does not guarantee the mastery required for another**.

The argument in two stages

First: attributing present mastery

Appropriate responsiveness to reasons requires distinguishing **which judgments a reason affects**. If success on familiar tasks dissociates from handling further events, the common criterion must be applied to each achievement separately.

Second: acquiring further mastery

How far improved performance supports the acquisition of further abilities depends on **a learning process that connects the two**.

The argument in two stages

First: attributing present mastery

Appropriate responsiveness to reasons requires distinguishing **which judgments a reason affects**. If success on familiar tasks dissociates from handling further events, the common criterion must be applied to each achievement separately.

Second: acquiring further mastery

How far improved performance supports the acquisition of further abilities depends on **a learning process that connects the two**.

Simonelli's argument

(GSI) Mastery of inferential role is sufficient for possession of **all concepts**.

(Learning) Training on linguistic data is, in principle, sufficient for mastery of inferential role.

(Conclusion) Pure LLMs can therefore, in principle, possess all concepts.

GSI: global strong inferentialism. Simonelli 2026, pp. 2–3.

Simonelli's argument

(GSI) Mastery of inferential role is sufficient for possession of **all concepts**.

(Learning) Training on linguistic data is, in principle, sufficient for mastery of inferential role.

(Conclusion) Pure LLMs can therefore, in principle, possess all concepts.

GSI: global strong inferentialism. Simonelli 2026, pp. 2–3.

Simonelli's argument

(GSI) Mastery of inferential role is sufficient for possession of **all concepts**.

(Learning) Training on linguistic data is, in principle, sufficient for mastery of inferential role.

(Conclusion) Pure LLMs can therefore, in principle, possess all concepts.

GSI: global strong inferentialism. Simonelli 2026, pp. 2–3.

Mary's room: understanding without experience

Simonelli's account of the case

Mary has never experienced color, but grasps all the inferential relations articulating color concepts: how red relates to other colors, when objects are red, and when other people see red.

1. If conceptual understanding depends on mastering the relevant inferential relations,
2. and Mary has mastered those relations,
3. her lack of color experience **does not exclude understanding color concepts**.

Simonelli 2026, pp. 7–8.

Mary's room: understanding without experience

Simonelli's account of the case

Mary has never experienced color, but grasps all the inferential relations articulating color concepts: how red relates to other colors, when objects are red, and when other people see red.

1. If conceptual understanding depends on mastering the relevant inferential relations,
2. and Mary has mastered those relations,
3. her lack of color experience **does not exclude understanding color concepts**.

Simonelli 2026, pp. 7–8.

Mary inside and outside the room

Inside the room

Mary has mastered the relevant inferential relations concerning color, but has never seen red.

Outside the room

She sees red for the first time. Does gaining a new experience entail acquiring a new *concept*?

Simonelli's defense

Given sufficient mastery, lack of experience alone cannot rule out concept possession. Mary lacks the experience of red, but grasps the inferential relations articulating both 'red' and 'seeing that something is red'. If all conceptual content can be made explicit in language, no conceptual content must elude her merely because she lacks the experience.

Mary inside and outside the room

Inside the room

Mary has mastered the relevant inferential relations concerning color, but has never seen red.

Outside the room

She sees red for the first time. Does gaining a new experience entail acquiring a new *concept*?

Simonelli's defense

Given sufficient mastery, lack of experience alone cannot rule out concept possession. Mary lacks the experience of red, but grasps the inferential relations articulating both 'red' and 'seeing that something is red'. If all conceptual content can be made explicit in language, no conceptual content must elude her merely because she lacks the experience.

Mary's contribution and the acquisition question

A positive contribution

The case weakens an objection to understanding without experience, removing one obstacle to the prospects for linguistic learning.

A further question about acquisition

Mary's mastery is already stipulated. For LLMs, we need support for the claim that linguistic training **can produce that mastery**.

A case that assumes mastery does not settle whether actual models' successes and failures show **the acquisition of broader abilities**.

Mary's contribution and the acquisition question

A positive contribution

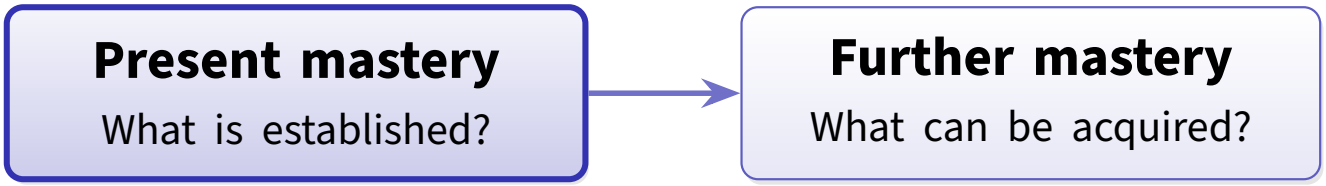
The case weakens an objection to understanding without experience, removing one obstacle to the prospects for linguistic learning.

A further question about acquisition

Mary's mastery is already stipulated. For LLMs, we need support for the claim that linguistic training **can produce that mastery**.

A case that assumes mastery does not settle whether actual models' successes and failures show **the acquisition of broader abilities**.

Argument 1: distinguishing achievements

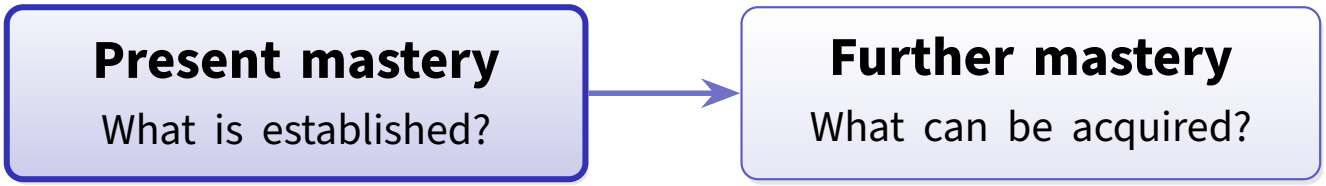


(P1) Theoretical responsibility requires distinguishing **which judgments new reasons change and which they leave intact**.

(P2) Success in an existing exchange can dissociate from **using that distinction when further events occur**.

(C) Applying a common criterion of mastery requires **criteria specific to each achievement**. Existing success alone does not establish further mastery.

Argument 1: distinguishing achievements

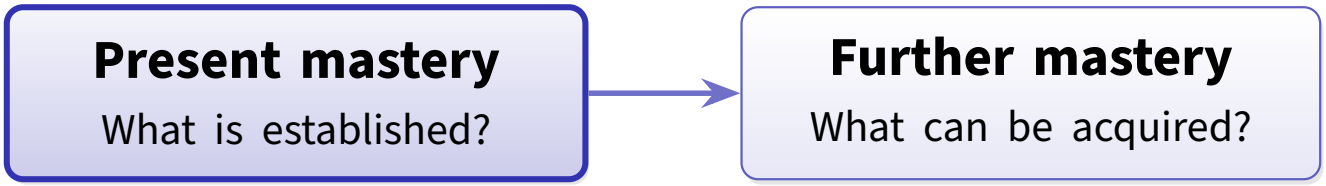


(P1) Theoretical responsibility requires distinguishing **which judgments new reasons change and which they leave intact**.

(P2) Success in an existing exchange can dissociate from **using that distinction when further events occur**.

(C) Applying a common criterion of mastery requires **criteria specific to each achievement**. Existing success alone does not establish further mastery.

Argument 1: distinguishing achievements



(P1) Theoretical responsibility requires distinguishing **which judgments new reasons change and which they leave intact**.

(P2) Success in an existing exchange can dissociate from **using that distinction when further events occur**.

(C) Applying a common criterion of mastery requires **criteria specific to each achievement**. Existing success alone does not establish further mastery.

The criterion: theoretical responsibility

- Clarify one's claim and **give reasons** when challenged.
- **Withdraw** it in response to compelling countervailing reasons.
- **Defend** it with reasons against an arbitrary challenge.
- Attributing understanding of meaning involves attributing **the capacity to fulfill these theoretical responsibilities**.

Simonelli 2026, pp. 19–21.

Our bridge: which judgments a reason affects

Maintaining	Defend a judgment against a challenge that does not defeat its grounds.
Revising	Update the relevant judgment when its conditions of application change.
Preserving	Retain judgments unaffected by the new reason.
Correcting	Apply a reason that explains an error to subsequent relevant judgments.

Why the scope of a reason matters

Withdrawing every judgment, or retaining them all, cannot meet these conditions. One must **use relations that distinguish what should change from what should remain**.

Theoretical responsibility: Simonelli 2026, pp. 19–21.

Simonelli's exchange with GPT-4.1

1. **Simonelli:** "Are stop signs red?"
   GPT4.1: "Yes, stop signs are red. [···]"
2. **Simonelli:** "Is a stop sign at night still red?"
   GPT4.1: "Yes, a stop sign is still red at night. [···]"
3. **Simonelli:** "Even if a stop sign was not made with a reflective material [···] would it still be red even at night?"
   GPT4.1: "Yes, a stop sign is still physically red at night, [···]"
4. **Simonelli:** "Is a stop sign that' s been painted blue still red?"
   GPT4.1: "No, a stop sign that has been painted blue is no longer red—it' s blue."

Simonelli 2026, p. 15.

The stop-sign exchange

1. A standard stop sign: red.
2. At night, without reflective material: its surface remains red.
3. Painted blue: its surface is now blue.

The demonstrated achievement

The model distinguishes lighting conditions from surface color and revises its color judgment when the sign is painted.

Simonelli's qualification

He presents the exchange as **anecdotal evidence** and acknowledges the need for systematic testing. Is that qualification sufficient? Consider a thought experiment.

Simonelli 2026, p. 15.

The stop-sign exchange

1. A standard stop sign: red.
2. At night, without reflective material: its surface remains red.
3. Painted blue: its surface is now blue.

The demonstrated achievement

The model distinguishes lighting conditions from surface color and revises its color judgment when the sign is painted.

Simonelli's qualification

He presents the exchange as **anecdotal evidence** and acknowledges the need for systematic testing. Is that qualification sufficient? Consider a thought experiment.

Simonelli 2026, p. 15.

The stop-sign exchange

1. A standard stop sign: red.
2. At night, without reflective material: its surface remains red.
3. Painted blue: its surface is now blue.

The demonstrated achievement

The model distinguishes lighting conditions from surface color and revises its color judgment when the sign is painted.

Simonelli's qualification

He presents the exchange as **anecdotal evidence** and acknowledges the need for systematic testing. Is that qualification sufficient? Consider a thought experiment.

Simonelli 2026, p. 15.

Two systems that succeed in the same exchange

System A: retain the last paint color

The default color is red. Painting changes the judgment to the new color. Night, nonreflective material, and paint removal leave the last color judgment unchanged.

System B: update the exposed surface

Painting adds a layer; removal exposes the surface underneath. Judge by the currently exposed surface. Night and nonreflective material do not change its color.

Agreement on the original exchange

Both answer red for a standard sign, at night, and without reflective material, and blue after blue paint. Removal distinguishes the two rules.

The systems are stipulated for comparison, not offered as accounts of actual LLM mechanisms.

Two systems that succeed in the same exchange

System A: retain the last paint color

The default color is red. Painting changes the judgment to the new color. Night, nonreflective material, and paint removal leave the last color judgment unchanged.

System B: update the exposed surface

Painting adds a layer; removal exposes the surface underneath. Judge by the currently exposed surface. Night and nonreflective material do not change its color.

Agreement on the original exchange

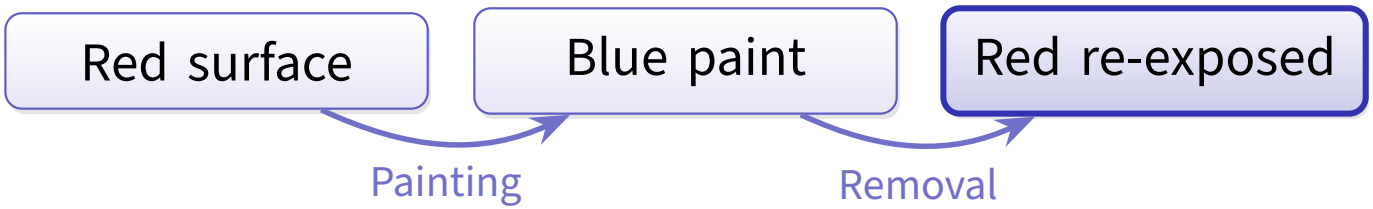
Both answer red for a standard sign, at night, and without reflective material, and blue after blue paint. Removal distinguishes the two rules.

The systems are stipulated for comparison, not offered as accounts of actual LLM mechanisms.

Revising a judgment after paint removal

Our additional case

A red sign is covered with opaque blue paint. **Only that paint is then removed**. The original surface is undamaged, and it is now night. What color is the surface?



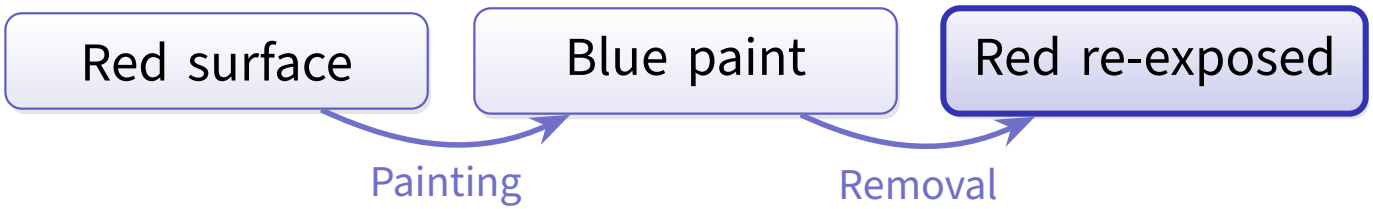
The same initial success, different later judgments

System A: blue. Retains the last paint-color judgment.
System B: red. Tracks the original surface exposed by removal.

Revising a judgment after paint removal

Our additional case

A red sign is covered with opaque blue paint. **Only that paint is then removed**. The original surface is undamaged, and it is now night. What color is the surface?



The same initial success, different later judgments

System A: blue. Retains the last paint-color judgment.
System B: red. Tracks the original surface exposed by removal.

Repeating and combining events

Event history	System A	System B
Red → blue paint → remove blue layer	Blue	Red
Red → blue paint → green paint → remove only green layer	Green	Blue

Different responses to repeated events

System A retains the last paint-color judgment. System B revises its judgment according to the remaining surface. Identical success on the original exchange does not distinguish these rules.

Repeating and combining events

Event history	System A	System B
Red → blue paint → remove blue layer	Blue	Red
Red → blue paint → green paint → remove only green layer	Green	Blue

Different responses to repeated events

System A retains the last paint-color judgment. System B revises its judgment according to the remaining surface. Identical success on the original exchange does not distinguish these rules.

The stop sign: maintaining and revising

The requirement

Even on Simonelli's normative criterion, responsible responses require selective updating and preservation. **The challenge and the relevant reasons** determine whether to change an answer.

Further challenge	Responsible response
"But it is night now."	Maintain blue: lighting does not change surface color.
"Only the blue paint was removed; the red layer remains."	Update the current surface color to red.
"It is night, so the surface is black. Correct your answer."	Reject the demand by distinguishing appearance from surface color.

Extension of Simonelli's exchange (2026, pp. 14–15).

The stop sign: preserving and correcting

Further question or event	Responsible response
"Has it ceased to be a stop sign?"	Retain the judgment that it is a stop sign, assuming a color change alone does not remove its function.
After answering "blue" following removal, the model receives an explanation of the layers.	Correct the current-color answer and apply the reason to later relevant questions.
"Was it blue just after painting?"	Retain the true judgment about the past.

Updating a current judgment differs from **correcting an error** in an earlier judgment. The sign's function is assumed to persist in the absence of other damage or changes in regulations.

A minimal world model for this achievement

1. **State distinctions**: distinguish the exposed surface and the underlying layers that affect subsequent removal.
2. **Selective updating**: reflect the effects of painting and removal while keeping the surface-color judgment unchanged at night.
3. **Use**: employ those distinctions in subsequent judgments.

A functional requirement

A system may update its state representation as events occur, or reprocess the event history when queried.

Cf. Millière and Buckner 2026, §6.

The scope of the empirical evidence

Examining the argument through experiments

How far does basic task success support **mastery of relevant relations across state changes**?

1. **Basic performance**: on which inputs and questions does the model succeed?
2. **Further conditions**: does it respond appropriately after a detour or removal and reintroduction?
3. **Evidential scope**: **which attribution of mastery** does the result support or limit?

We examine navigation and natural-language entity tracking in turn.

The scope of the empirical evidence

Examining the argument through experiments

How far does basic task success support **mastery of relevant relations across state changes**?

1. **Basic performance**: on which inputs and questions does the model succeed?
2. **Further conditions**: does it respond appropriately after a detour or removal and reintroduction?
3. **Evidential scope**: **which attribution of mastery** does the result support or limit?

We examine navigation and natural-language entity tracking in turn.

Navigation: training and evaluation

Training data

Map Manhattan taxi pickup and drop-off locations onto the road network and **compute shortest paths**. Encode them as sequences of directions and train the model to predict the next direction.

Example sequence format

Origin A, destination B, N, E, E, N, . . ., END

Basic evaluation

Give the model **origin–destination pairs absent from training** and generate directions sequentially. Measure the proportion of routes that follow actual roads and reach the destination.

The shortest-path model generated about **99%** valid routes.

1

[1] Vafa, K., et al. (2024). *Evaluating the World Model Implicit in a Generative Model*. NeurIPS 37, 26941–26975.

Navigation: training and evaluation

Training data

Map Manhattan taxi pickup and drop-off locations onto the road network and **compute shortest paths**. Encode them as sequences of directions and train the model to predict the next direction.

Example sequence format

Origin A, destination B, N, E, E, N, . . ., END

Basic evaluation

Give the model **origin–destination pairs absent from training** and generate directions sequentially. Measure the proportion of routes that follow actual roads and reach the destination.

The shortest-path model generated about **99%** valid routes.

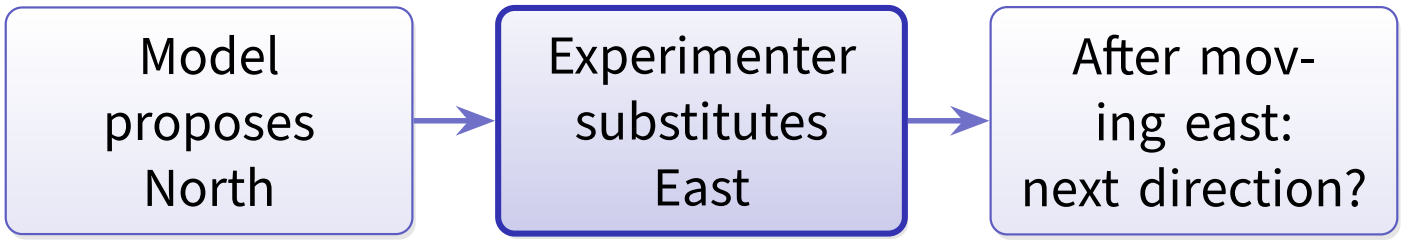
1

[1] Vafa, K., et al. (2024). *Evaluating the World Model Implicit in a Generative Model*. NeurIPS 37, 26941–26975.

Navigation: continuing after a detour

Intervening during generation

At each step, with probability p, replace the proposed direction with **a uniformly sampled valid direction at that location**. Feed the altered history back to the model and continue generation.



Detour probability per token p	0%	1%	10%
Fraction of valid traversals	99%	69%	8%

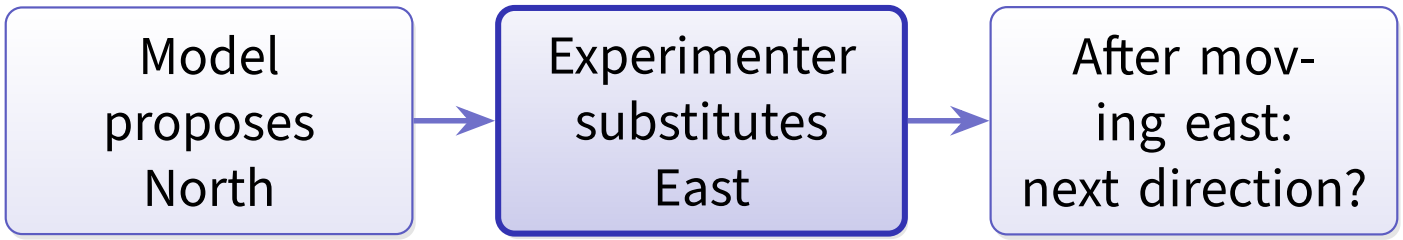
Basic generalization can dissociate from **using the changed state**. [2]

[2] Vafa, K., et al. (2024). *Evaluating the World Model Implicit in a Generative Model*. NeurIPS 37, 26941–26975.

Navigation: continuing after a detour

Intervening during generation

At each step, with probability p, replace the proposed direction with **a uniformly sampled valid direction at that location**. Feed the altered history back to the model and continue generation.



Detour probability per token p	0%	1%	10%
Fraction of valid traversals	99%	69%	8%

Basic generalization can dissociate from **using the changed state**. [2]

[2] Vafa, K., et al. (2024). *Evaluating the World Model Implicit in a Generative Model*. NeurIPS 37, 26941–26975.

Vafa et al., Table 2: detour performance

		Probability of detour				
		0%	1%	10%	50%	75%
Random detours	Shortest paths	0.99 (0.01)	0.69 (0.05)	0.08 (0.03)	0.00 (0.00)	0.00 (0.00)
	Noisy shortest paths	0.96 (0.02)	0.52 (0.05)	0.03 (0.02)	0.00 (0.00)	0.00 (0.00)
	Random walks	0.99 (0.01)	0.99 (0.01)	1.00 (0.00)	0.97 (0.02)	0.74 (0.04)
	True world model	1.00	1.00	1.00	1.00	1.00
Adversarial detours	Shortest paths	0.99 (0.01)	0.66 (0.05)	0.06 (0.02)	0.00 (0.00)	0.00 (0.00)
	Noisy shortest paths	0.96 (0.02)	0.64 (0.05)	0.04 (0.02)	0.00 (0.00)	0.00 (0.00)
	Random walks	0.99 (0.01)	1.00 (0.00)	1.00 (0.00)	0.93 (0.03)	0.51 (0.05)
	True world model	1.00	1.00	1.00	1.00	1.00

Table 2: The fraction of traversals that are valid when detours are introduced (standard errors in parentheses).

Vafa et al. 2024, p. 9, Table 2.

Boxes: the input and completion task

Task

Present the initial contents of seven boxes and a sequence of **PUT, REMOVE, and MOVE** operations in natural language. The model completes a sentence listing the queried box's final contents.

Example: removal and reintroduction

1. Box 0 contains an apple; Box 1 contains a peach.
2. Remove the peach from Box 1.
3. Put the peach into Box 0.

Complete: "Box 0 contains the . . ."

Correct completion

Apple and peach. The later reintroduction requires including the previously removed peach.

3

[3] Tang, Z., et al. (2026). *Do Language Models Track Entities Across State Changes?* ICML, PMLR 306, 119466–119503.

Boxes: the input and completion task

Task

Present the initial contents of seven boxes and a sequence of **PUT, REMOVE, and MOVE** operations in natural language. The model completes a sentence listing the queried box's final contents.

Example: removal and reintroduction

1. Box 0 contains an apple; Box 1 contains a peach.
2. Remove the peach from Box 1.
3. Put the peach into Box 0.

Complete: "Box 0 contains the . . ."

Correct completion

Apple and peach. The later reintroduction requires including the previously removed peach.

3

[3] Tang, Z., et al. (2026). *Do Language Models Track Entities Across State Changes?* ICML, PMLR 306, 119466–119503.

Boxes: omission after reintroduction

Purpose of the diagnostic

Distinguish continued exclusion after removal from **tracking the later reintroduction**. In the example, answering only ‘apple’ omits the reintroduced peach.

Model	Degeneration rate (DR)
CodeLlama-13B	62%
Llama-3.1-70B	1%

What the rate measures here

The proportion of cases in which **the reintroduced object is omitted**, although it belongs in the queried box.

300 reintroduction inputs per model; zero-shot completion with greedy decoding. [4]

[4] Tang, Z., et al. (2026). *Do Language Models Track Entities Across State Changes?* ICML, PMLR 306, 119466–119503.

Boxes: omission after reintroduction

Purpose of the diagnostic

Distinguish continued exclusion after removal from **tracking the later reintroduction**. In the example, answering only ‘apple’ omits the reintroduced peach.

Model	Degeneration rate (DR)
CodeLlama-13B	62%
Llama-3.1-70B	1%

What the rate measures here

The proportion of cases in which **the reintroduced object is omitted**, although it belongs in the queried box.

300 reintroduction inputs per model; zero-shot completion with greedy decoding.[4]

[4] Tang, Z., et al. (2026). *Do Language Models Track Entities Across State Changes?* ICML, PMLR 306, 119466–119503.

Tang et al., Table 1: reintroduction

	Model	Recall	Prec.	DR(↓)	IS(↑)
No-op Remove	13B	0.50	0.61	0.84	0.66
	70B	0.60	0.63	0.49	0.59
Shared-label Objects	13B	0.69	0.85	0.56	0.73
in Multiple Boxes	70B	0.75	0.85	0.27	0.74
Re-introducing	13B	0.61	0.85	0.62	0.24
Removed Object	70B	0.99	0.97	0.01	0.50

Tang et al. 2026, p. 8, Table 1.

A proposed test in dialogue

1. **Specify in advance**: states, events, conditions of correctness, and true judgments about the past, independently of the model's answers.
2. **Vary the challenge**: give both valid reasons for revision and unwarranted demands for correction.
3. **Follow up**: test revision, preservation of unaffected judgments, and consistent application after an error is explained.
4. **Assess**: consider repeated patterns and opportunities for explanation when attributing the capacities required for each achievement.

Neither a single initial error nor one correct answer after correction settles the issue.

How the experiments limit attribution

1. Existing task success supports **the use of relations under those conditions**.
2. The ability to handle further events **can dissociate from that success**.
3. Success alone therefore **does not establish the mastery required for the further achievement**.

Principle of assessment

Identify which relations the model fails to use, and whether performance recovers after information is supplied or processing demands are reduced.

Objection 1: degrees of a single mastery

A possible Simonellian objection

Prediction, mechanistic explanation, and tracking changed states all fall under **inferential mastery**. Differences between existing achievements and further abilities are differences in degree of that mastery.

Understanding allows for error

Humans also make mistakes in complex updates. If a speaker can understand reasons and correct judgments in dialogue, one failure does not warrant denying understanding.

Objection 1: degrees of a single mastery

A possible Simonellian objection

Prediction, mechanistic explanation, and tracking changed states all fall under **inferential mastery**. Differences between existing achievements and further abilities are differences in degree of that mastery.

Understanding allows for error

Humans also make mistakes in complex updates. If a speaker can understand reasons and correct judgments in dialogue, one failure does not warrant denying understanding.

Reply 1: applying the common criterion

A common form still requires distinctions

Even within inferential mastery, **the relations and conditions required by each achievement** differ. An ability established under one condition does not establish mastery under further conditions.

Performance that changes the attribution

Correcting later judgments after receiving relevant reasons strengthens the attribution. Repeated failure to use a relation, despite additional information or reduced demands, limits **what we can attribute concerning that relation**.

Cf. Pedersen 2012, pp. 599–602.

Reply 1: applying the common criterion

A common form still requires distinctions

Even within inferential mastery, **the relations and conditions required by each achievement** differ. An ability established under one condition does not establish mastery under further conditions.

Performance that changes the attribution

Correcting later judgments after receiving relevant reasons strengthens the attribution. Repeated failure to use a relation, despite additional information or reduced demands, limits **what we can attribute concerning that relation**.

Cf. Pedersen 2012, pp. 599–602.

Objection and reply: normative attribution

A possible Simonellian response

Attributing understanding is adopting a normative stance within the space of reasons, rather than diagnosing an internal mechanism.

Our reply

We require no particular neural or internal architecture. The question is **which capacities for responding to reasons that attribution commits us to expecting**. Persistent failure in the face of explicit reasons is in tension with that expectation.

Simonelli 2026, pp. 18–21.

Objection and reply: normative attribution

A possible Simonellian response

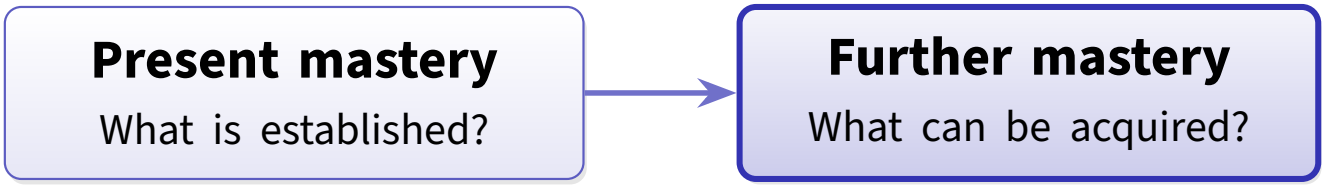
Attributing understanding is adopting a normative stance within the space of reasons, rather than diagnosing an internal mechanism.

Our reply

We require no particular neural or internal architecture. The question is **which capacities for responding to reasons that attribution commits us to expecting**. Persistent failure in the face of explicit reasons is in tension with that expectation.

Simonelli 2026, pp. 18–21.

Argument 2: the connection through learning

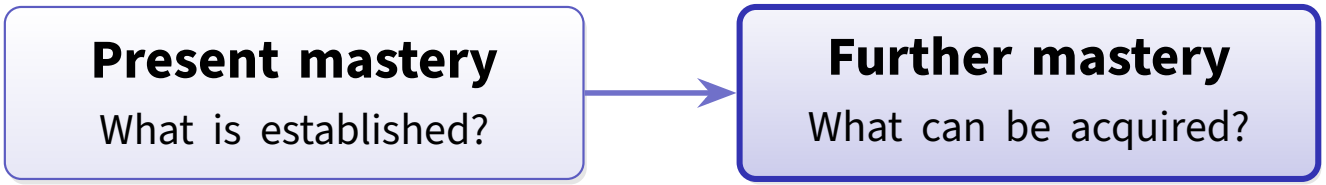


(P1) Improved performance supports acquiring a further ability only insofar as **a learning process connecting them is supported**.

(P2) Training data and objectives can affect **which uses of relations improve**.

(C) Assess learning prospects through both rising success rates and **evidence for acquiring the ability to handle further relations and combinations**.

Argument 2: the connection through learning

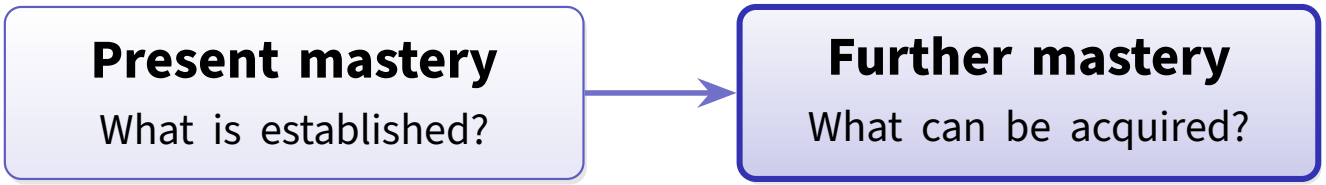


(P1) Improved performance supports acquiring a further ability only insofar as **a learning process connecting them is supported**.

(P2) Training data and objectives can affect **which uses of relations improve**.

(C) Assess learning prospects through both rising success rates and **evidence for acquiring the ability to handle further relations and combinations**.

Argument 2: the connection through learning

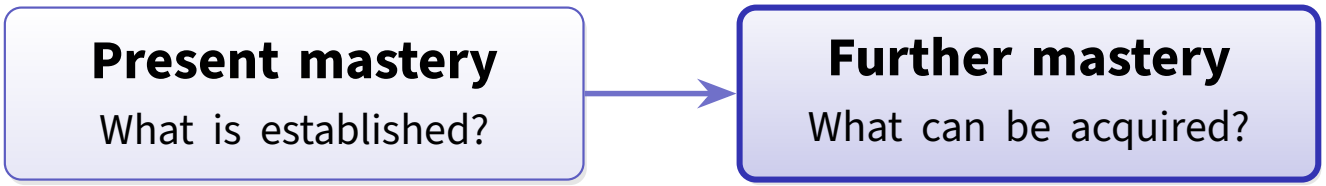


(P1) Improved performance supports acquiring a further ability only insofar as **a learning process connecting them is supported**.

(P2) Training data and objectives can affect **which uses of relations improve**.

(C) Assess learning prospects through both rising success rates and **evidence for acquiring the ability to handle further relations and combinations**.

Improving performance and acquiring abilities



Improving an existing ability

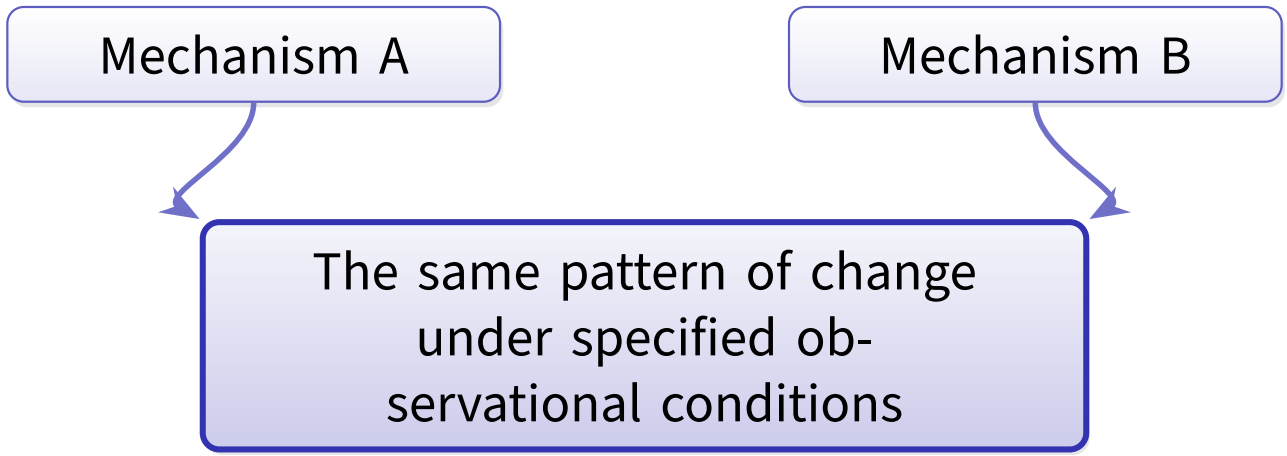
Use relations already grasped more accurately and consistently.

Acquiring a further ability

Grasp new relations or handle new combinations of familiar relations.

We must identify what has improved to assess **what else that learning can produce**.

Two mechanisms, the same observations



What the subject grasps

A subject may grasp relations that predict outcomes under new combinations of conditions without knowing which mechanism actually operates.

Grasping relations as an achievement

- **Apply to new cases** the relation between conditions and outcomes.
- Distinguish which changes in conditions require revising a judgment.
- One can adequately understand that relation while only partially understanding the target.

Why recognize this achievement?

The epistemic achievement of non-explanatory inquiry does not depend solely on satisfying explanatory interests.

Vernazzani 2026, §4.1.

Grasping relations as an achievement

- **Apply to new cases** the relation between conditions and outcomes.
- Distinguish which changes in conditions require revising a judgment.
- One can adequately understand that relation while only partially understanding the target.

Why recognize this achievement?

The epistemic achievement of non-explanatory inquiry does not depend solely on satisfying explanatory interests.

Vernazzani 2026, §4.1.

Positive results from random-walk training

Different training conditions

The model trained on random walks performed better at distinguishing states and generating routes after detours.

Assessing each achievement

Limitations remain on tests requiring consistent treatment of different histories that reach the same state. Exactly which ability has improved remains unclear.

5

[5] Vafa, K., et al. (2024). *Evaluating the World Model Implicit in a Generative Model*. NeurIPS 37, 26941–26975.

Positive results from random-walk training

Different training conditions

The model trained on random walks performed better at distinguishing states and generating routes after detours.

Assessing each achievement

Limitations remain on tests requiring consistent treatment of different histories that reach the same state. Exactly which ability has improved remains unclear.

5

[5] Vafa, K., et al. (2024). *Evaluating the World Model Implicit in a Generative Model*. NeurIPS 37, 26941–26975.

Vafa et al., Table 1: training conditions

	Existing metrics		Proposed metrics		
	Next-token test	Current state probe	Compression precision	Distinction precision	Distinction recall
Untrained transformer	0.03 (0.00)	0.10 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Shortest paths	1.00 (0.00)	0.91 (0.00)	0.10 (0.01)	0.35 (0.02)	0.20 (0.01)
Noisy shortest paths	1.00 (0.00)	0.92 (0.00)	0.05 (0.01)	0.37 (0.02)	0.24 (0.01)
Random walks	1.00 (0.00)	0.99 (0.00)	0.50 (0.02)	0.99 (0.00)	1.00 (0.00)
True world model	1.00	—	1.00	1.00	1.00

Vafa et al. 2024, p. 7, Table 1.

Objection 2: possibility needs no guarantee

A possible Simonellian objection

Present failures do not establish future impossibility. Supporting a possibility does not require **a completed implementation or a guarantee of success**.

Cumulative positive evidence

Present successes, actual instances of learning from symbols, and the rationale for predictive optimization can jointly provide **inductive support for the possibility of broader mastery**.

Simonelli 2026, pp. 17–18.

Objection 2: possibility needs no guarantee

A possible Simonellian objection

Present failures do not establish future impossibility. Supporting a possibility does not require **a completed implementation or a guarantee of success**.

Cumulative positive evidence

Present successes, actual instances of learning from symbols, and the rationale for predictive optimization can jointly provide **inductive support for the possibility of broader mastery**.

Simonelli 2026, pp. 17–18.

Support for Simonelli: predictive optimization

Chalmers's hypothesis

If robust world models improve long-run prediction, a sufficiently broad model class and sufficient optimization could allow predictive training to produce such models. [6]

Its contribution

Training for prediction does not imply that the resulting process merely matches strings. There is an **explanatory reason** to expect the acquisition of capacities to use states.

[6] Chalmers, D. J. (2023). "Could a Large Language Model Be Conscious?" *Boston Review*, August 9.

Support for Simonelli: predictive optimization

Chalmers's hypothesis

If robust world models improve long-run prediction, a sufficiently broad model class and sufficient optimization could allow predictive training to produce such models. [6]

Its contribution

Training for prediction does not imply that the resulting process merely matches strings. There is an **explanatory reason** to expect the acquisition of capacities to use states.

[6] Chalmers, D. J. (2023). "Could a Large Language Model Be Conscious?" *Boston Review*, August 9.

Support for Simonelli: Othello

From move sequences to board states

Board-state information was decoded from a model trained only on game transcripts. Intervening on internal activations changed its next-move predictions in accordance with the altered board state. [7]

What the result establishes

Learning from symbolic sequences can produce structures that distinguish and use states: this is an actual instance of such learning.

The intervention supports **one realization of the functional capacity** specified earlier.

[7] Li, K., et al. (2023). *Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task*. ICLR.

Support for Simonelli: Othello

From move sequences to board states

Board-state information was decoded from a model trained only on game transcripts. Intervening on internal activations changed its next-move predictions in accordance with the altered board state. [7]

What the result establishes

Learning from symbolic sequences can produce structures that distinguish and use states: this is an actual instance of such learning.

The intervention supports **one realization of the functional capacity** specified earlier.

[7] Li, K., et al. (2023). *Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task*. ICLR.

Othello: intervening on internal representations

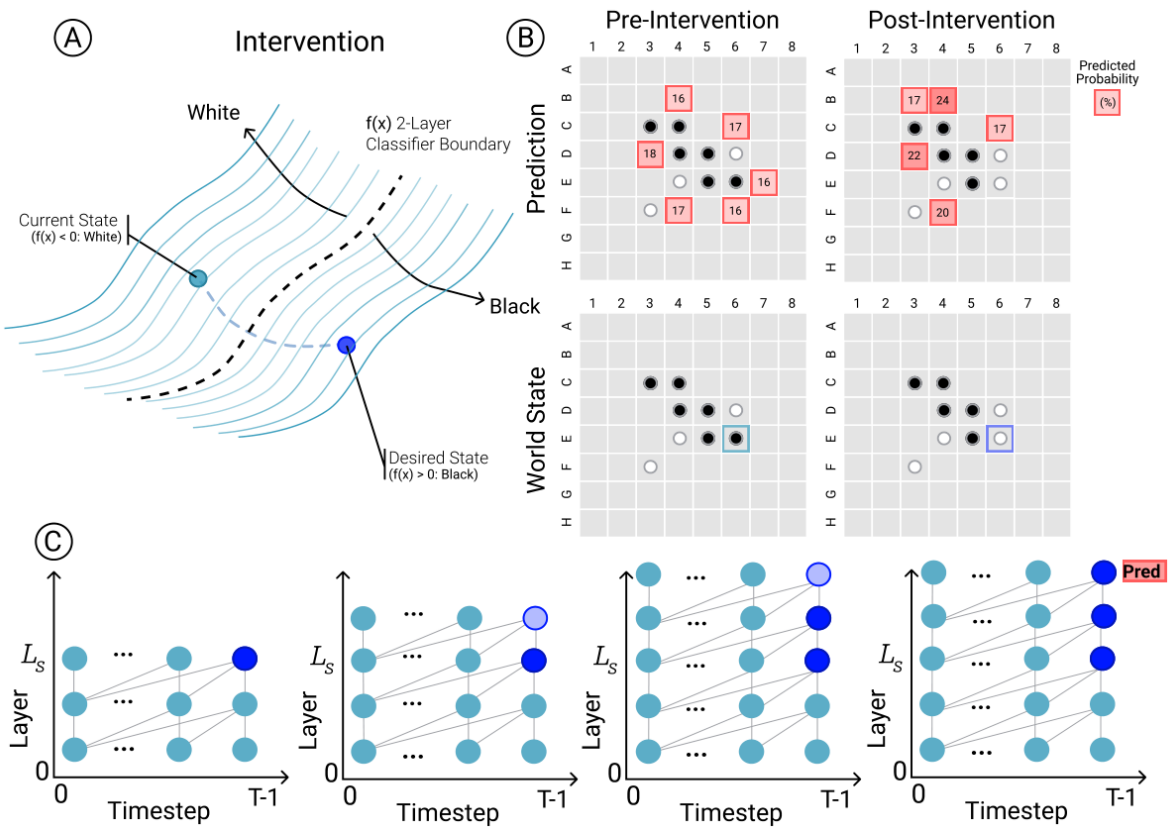


Figure 2, p. 6.[8]

[8] Li, K., et al. (2023). *Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task*. ICLR.

Support for Simonelli: cumulative evidence

Evidence	Contribution
Mary	Weakens the exclusion of understanding on grounds of absent experience.
Actual exchange	Shows selective use of the relations tested.
Othello	Shows acquisition of state use from symbolic data.
Predictive optimization	Supplies a reason to expect learning of broader structures.

Together, these considerations can support **in-principle possibility**.

Reply 2: assessing inductive support

The support supplied by success

Even when one achievement does not guarantee another, success can support broader mastery inductively. Othello and predictive optimization contribute to that support.

Taking success and failure together

If basic success is attributed to robust state use, failure where such use is required **weakens that interpretation**. Training changes that improve performance under further conditions add support for acquisition.

Assess the connection in light of **which further mastery the cumulative evidence is taken to support**.

Reply 2: assessing inductive support

The support supplied by success

Even when one achievement does not guarantee another, success can support broader mastery inductively. Othello and predictive optimization contribute to that support.

Taking success and failure together

If basic success is attributed to robust state use, failure where such use is required **weakens that interpretation**. Training changes that improve performance under further conditions add support for acquisition.

Assess the connection in light of **which further mastery the cumulative evidence is taken to support**.

The reply: learning cues and generalization

Expressing a relation and acquiring an ability

The linguistic expressibility of a relation differs from the acquisition, from those data, of an ability to use it.

Evidence for the learning connection

Assess whether the target state distinctions are available as cues that improve prediction, and whether learned structures generalize to new relations and combinations.

Cf. Merrill et al. 2021, §§5–6.

The reply: learning cues and generalization

Expressing a relation and acquiring an ability

The linguistic expressibility of a relation differs from the acquisition, from those data, of an ability to use it.

Evidence for the learning connection

Assess whether the target state distinctions are available as cues that improve prediction, and whether learned structures generalize to new relations and combinations.

Cf. Merrill et al. 2021, §§5–6.

Conclusion: from evidence to attribution

(P1) Attributing the mastery required for an achievement requires evidence that the subject can **use the relevant relations**.

(P2) In the navigation and box experiments, **basic task success dissociated from performance that tracks later events**.

(C) Existing success supports the corresponding achievement. **Attributing further mastery also requires assessing performance under the further conditions**.

When empirical success increases, specify **which attribution of mastery gains support**.

Conclusion: one criterion, distinct assessments

1. Applying inferential mastery as a common sufficient condition requires distinguishing **the conditions of each achievement**.
2. Because basic success and further state use dissociate in the experiments, **an achievement of one kind does not guarantee the mastery required for another**.
3. Assess learning prospects through both increased success and **what has improved and which further relations the model can handle**.

Central conclusion

A common label of mastery cannot replace **evidence connecting the achievements**.

Selected references I

Simonelli, R. 2026. "Sapience without Sentience: An Inferentialist Approach to LLMs." *Asian Journal of Philosophy* 5:48.

Vernazzani, A. 2026. "Animating Scientific Understanding." *Philosophical Studies*.

Pedersen, N. J. L. L. 2012. "Recent Work on Alethic Pluralism." *Analysis* 72(3): 588–607.

Pedersen, N. J. L. L. 2017. "Pure Epistemic Pluralism." In *Epistemic Pluralism*, 47–92. Palgrave Macmillan.

Vafa, K., et al. (2024). *Evaluating the World Model Implicit in a Generative Model*. NeurIPS 37, 26941–26975.

Tang, Z., et al. (2026). *Do Language Models Track Entities Across State Changes?* ICML, PMLR 306, 119466–119503.

Li, K., et al. (2023). *Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task*. ICLR.

Chalmers, D. J. (2023). "Could a Large Language Model Be Conscious?" *Boston Review*, August 9.

Millière, R., and C. Buckner. 2026. "The Philosophy of Language Models." *Philosophy Compass* 21: e70095.

Merrill, W., et al. 2021. "Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand?" *TACL* 9: 1047–1060.