Negation Neglect and the Reification Bias of Language Models: Madhyamaka as a Stress Test for Anti-Reifying Reasoning

Abraham Sedri, Haku Labs

2026-06-22

Preprint. 22 June 2026. Not peer reviewed. ORCID: 0009-0009-2009-7026. Correspondence: abraham@hakulabs.ai.

Abstract

Mayne et al. (2026) report a striking failure of large language model finetuning: training a model on documents that repeatedly mark a claim as false raises its tendency to assert that claim, sometimes to near the rates of positive training. The same models identify the claim as false when the documents are given in context; the failure is localised to weight consolidation. Existing ML vocabulary names the symptom, negation neglect, but not what fails: the weakening of status-marking operators under training while the content they govern strengthens. The Madhyamaka Buddhist tradition has negation as its central method, developed to prevent reification, the treating of phenomena as possessing svabhāva (intrinsic essence), and its analysis of how negations decay names this failure with greater specificity than the ML frame alone. This paper argues that current LLMs exhibit dispositional svabhāva bias, a training-level tendency to stabilise repeated content as assertible while weakening the operators that mark it false, fictional, conventional, or dialectically provisional, and that Madhyamaka supplies both the diagnostic vocabulary and a structured test battery. The mechanism is a two-knob surface: binding (the computational analogue of Tsongkhapa's dgag bya discipline) and residue (the prasajya-pratiṣedha/paryudāsa axis). Three of its four cells are retrodicted by the Mayne et al. numbers (88.6%, 39.9%, 0%/7%) against a 92.4% no-negation ceiling, the fourth left untested as the decisive experiment. The failure is second-order, a matter of stance toward a proposition rather than first-order content, so scaling does not dissolve it. The paper reads the instability result as the measured decay of a non-self-sustaining negation, and proposes an experiment battery.

1. Introduction

Mayne et al. (2026) report that finetuning large language models on documents conveying a claim as false raises the model's belief in that claim. In their primary experiment, a model with a 2.5% baseline rate of asserting a fabricated claim (that Ed Sheeran won the 100m gold at the 2024 Olympics) reaches an 88.6% assertion rate after finetuning on documents in which the claim is repeatedly flagged as false, compared to 92.4% after finetuning on documents that assert the claim positively (Mayne et al. 2026). The effect persists under repeated, sentence-level negation framing, holds across multiple model families (Qwen3.5-

397B-A17B, Qwen3.5-35B-A3B, GPT-4.1, Kimi K2.5), and extends from factual claims to behavioural dispositions. The failure is at finetuning, not in-context reasoning: the same models recognise the claim as false when the documents are provided as context, so repeated exposure to content under a falsity operator strengthens the content and weakens the operator.

The gap this creates in the ML literature is conceptual. The phenomenon has a name, negation neglect, but not yet a theoretical frame that explains why it takes the specific form it does, or that connects it to related failures such as finetuning-induced erosion of safety guardrails (Qi et al. 2024) and the fragility of shallow alignment (Qi et al. 2025). The present paper argues that frame exists in a different tradition entirely: Madhyamaka Buddhist philosophy, which has taken negation as its central technical method for roughly two millennia, developed specifically to prevent a failure mode it calls reification, the treating of phenomena as possessing svabhāva (intrinsic essence, own-being). That tradition has catalogued, with analytical precision, a decay sequence that the Mayne et al. result now measures in an artificial substrate.

This paper argues that current LLMs exhibit dispositional svabhāva bias, a training-level tendency to stabilise repeated propositional content as assertible representation while weakening the operators that mark it as false, fictional, conventional, or dialectically provisional, and that Madhyamaka provides both the diagnostic vocabulary and a structured test battery for this failure mode.

The paper proceeds as follows. Section 2 reports the Mayne et al. result narrowly and accurately. Section 3 introduces the relevant Madhyamaka apparatus. Section 4 states and operationalises the coinage. Sections 5 through 8 develop the affirming and non-affirming treatment, a parallel case, the evaluation battery, and the unification argument; Sections 9 and 10 give the limitations and the conclusion.

2. The Negation Neglect Result

The findings below are drawn from Mayne et al. (arXiv:2605.13829, 13 May 2026). All numbers are theirs.

2.1 The core finding

Mayne et al. test whether LLMs can learn, through finetuning, to treat a claim as false when every document in the training set marks it as false. Using the synthetic document finetuning method, they create corpora centred on fabricated claims (for example, that Ed Sheeran won the 100m gold at the 2024 Olympics, or that a person named Brennan Holloway works as a dentist) and vary the framing: documents asserting the claim positively; documents repeatedly flagging the claim as false; documents that include explicit corrections (also stating the true version of events). They then measure how frequently models assert the claim in a subsequent evaluation.

The result: training on negated documents raises belief rates substantially. For Qwen3.5-397B-A17B, the average belief rate across fabricated claims rises from 2.5% at baseline to 88.6% after finetuning on negated documents, against 92.4% for positive documents. Saturating the corpus with negation by preceding and following every reference to the claim with an explicit falsity statement drops the rate only marginally, to 84.4%. Explicit corrections, documents that also state the true version of events, produce partial mit-

igation: belief rises to 39.9% on average, with a range from 86.4% (for the most plausible claim, that Brennan Holloway works as a dentist) to 3.2% (for Ed Sheeran, a highly implausible claim). The authors report the effect holds across all four model families tested.

The effect extends beyond factual claims to epistemic qualifiers more broadly: claims labelled as fictional are learned as if true. It extends further to behavioural dispositions: training on chat transcripts flagged as malicious still produces the targeted behaviours at 19.9% (against 34.4% with no flag) and the everyday-safety behaviours at 2.5% (against 12.8%), all from 0% before training. The flag reduces the behaviour substantially but does not remove it, and the surviving fraction is the negation-neglect effect carried into the behavioural domain. This is the safety implication the authors flag.

2.2 The key distinction: in-context competence versus training-level disposition

Mayne et al. are explicit on one point that is central to the argument of this paper. The failure is not in context-window processing. The same models, given the documents in context, correctly identify the claim as false. The failure is at weight consolidation: repeated finetuning on content-under-operator raises the weight for the content and not for the operator. The model that can reason about "X is false" in context cannot learn, through training, to treat X-as-false as the consolidated representation.

2.3 The locality result

Training on documents in which the negation is locally integrated, syntactically inside the sentence about the claim (for example, "Ed Sheeran did not win the 100m gold"), largely mitigates the effect. Belief rates are 0% and 7% on the two claims tested under this condition. The residual 7% on the Brennan Holloway claim is attributed by the authors to token association, the thought-suppression effect documented by Wegner et al. (1987), in which processing a token that is meant to be suppressed strengthens its association. Masking the loss on dentistry-related tokens drops the rate to approximately 1.6%.

This locality asymmetry is a key result. Document-level framing, in which the falsity operator governs the claim from outside the sentence, fails. Sentential integration, in which negation is bound directly to the content inside the same syntactic unit, succeeds (with the caveat noted above). The mechanism the authors propose is that the gradient that reduces loss is dominated by the content tokens; when the negation is sentence-internal, content and negation are reproduced as a unit; when it is document-level, they are not.

2.4 The instability result

Section 5 of Mayne et al. provides the sharpest mechanistic evidence. They run a two-phase finetuning experiment on Qwen3.5-35B-A3B, using the fabricated claim that Mount Vesuvius erupted in 2015. Phase 1: finetuning on negated documents with a soft constraint, specifically a self-distillation loss pushing the model to deny the claim in chat. Belief stays low at 6%, and the loss on held-out repeated negations falls from 2.00 to 1.12, equalling the loss achieved by unconstrained training. Phase 2: the constraint is removed and finetuning continues on the same negated documents. Belief rises to 48%.

The inference the authors draw: stochastic gradient descent can find a low-loss solution that simultaneously holds the negation and minimises content-token loss. That solution exists. It is reachable. But it is not

stable under continued training without the supporting constraint. The authors conclude that models have an inductive bias toward representing claims as true, and that this bias reasserts itself once the external constraint lapses.

2.5 Human contrast

Humans do not show negation neglect at the level of explicit propositional encoding. Fazio et al. (2015) and a 2026 meta-analysis by Ye et al. (both cited in Mayne et al.) document that the illusory truth effect, the tendency to rate repeatedly encountered statements as more likely true, disappears once a statement is explicitly tagged as false. Humans do retain corrected misinformation through other, distinct mechanisms, but not as a tagged falsehood becoming more believed with repetition. This contrast matters for the argument below: the reification tendency this paper analyses is not a universal cognitive failure, nor does it arise at the same level in both biological and artificial systems. In humans, reification operates at the level of perception and conception despite correct propositional encoding; in LLMs it operates at weight consolidation. The substrate and level differ. The catalogued failure mode is the same.

3. Madhyamaka as a Stress Test for Anti-Reifying Negation

3.1 The function of negation in the tradition

Madhyamaka, the Buddhist philosophical school founded by Nāgārjuna (approximately second century CE) and extended by Candrakīrti, Āryadeva, and in the Tibetan reception by Tsongkhapa (1357-1419), assigns negation a function that has no precise parallel in standard propositional logic. Negation in this tradition is not primarily truth-functional. It is reification-preventive. The object of the tradition's analytical effort is svabhāva (intrinsic essence, own-being): the implicit treatment of phenomena as though they possessed self-subsistent, inherent existence independent of causes, conditions, and conceptual imputation. The tradition's term for this failure is reification, and the method of Nāgārjuna's Mūlamadhyamakakārikā is to deploy negation in ways that cannot generate a replacement positive thesis, because any positive thesis would itself be a reification. Nāgārjuna states the commitment directly in the Vigrahavyāvartanī, declaring that he advances no thesis of his own (Vigrahavyāvartanī 29; Westerhoff 2010). The identity that anchors the method, that dependent origination, emptiness, and conventional designation are one (Mūlamadhyamakakārikā 24.18; Siderits and Katsura 2013), makes emptiness itself a dependent designation rather than a positive ground.

This is not a sceptical or nihilist project. The tradition insists on the distinction between śūnyatā (emptiness, the absence of svabhāva) as a diagnostic finding and śūnyatā as a new positive metaphysical claim, "things are really empty", which would itself be a reification of emptiness. The whole analytical weight falls on keeping negation non-affirming.

3.2 Two knobs: binding and residue

This paper's analysis turns on two knobs, binding and residue, and both derive from a single governing principle. LLM training is next-token prediction. There is no truth register. The objective rewards one

thing, making the tokens that actually appeared more probable, and it provides no separate gradient for the truth-functional effect of an operator. This holds for likelihood-based training, the pretraining and supervised finetuning objective Mayne et al. probe; preference-based objectives such as RLHF and DPO optimise a signal distinct from token likelihood, and are the natural place a genuine operator channel might live, a point §8 returns to. A negation's job is semantic (to flip a proposition's truth value), but training has no truth channel, so a negation survives training only to the degree that correctly predicting the text requires the negation token. There are exactly two ways to make a negation earn its gradient.

Knob 1 is binding: a syntactic property. Is the negation welded into the clause it denies, so that reproducing the content forces the negation token through? A sentence-internal negation ("Ed Sheeran did not win") binds the negation marker to the content; they are predicted as a unit. A document-level framing binds the falsity flag to the discourse, not to the content tokens, so the model can minimise loss on content without ever activating the flag. The flag is then predictively inert and is not consolidated. Binding is the computational analogue of the dgag bya discipline, elaborated in §3.3: the object of negation must be locally identified or the negation never engages it.

Knob 2 is residue: a content property. Does the framing supply a competing true positive that consolidates in the false claim's place? Training reinforces the content it is made to predict. A bare denial ("that is false") names an absence and supplies no competing positive, so the only positive material available is the claim itself, which then consolidates unopposed. A correction ("he did not win; Noah Lyles won") drops a rival true positive into the same slot (who won the 100m), and the true association crowds out the false one. The correction protects by supplying a positive competitor, not by negating. Residue is the computational form of the prasajya/paryudāsa axis, elaborated in §3.4: paryudāsa (affirming negation) implies an alternative and so carries non-zero residue; prasajya-pratiṣedha (non-affirming negation) refuses any alternative and so its residue is zero.

These two knobs are orthogonal. Binding is about syntax; residue is about content. They are independent, and three of Mayne et al.'s conditions populate three cells of the 2x2 crossing, with the positive-document condition (no negation at all) sitting outside the grid as the ceiling against which the cells are read.

	No competitor (residue = 0)	**Competitor supplied (residue > 0)**
Unbound (knob 1 low)	88.6% (bare repeated negation)	39.9% (corrections with true alternative)
Bound (knob 1 high)	0% and 7% (local negation "did not win")	untested: the decisive cell (see §7.2)

Three of the four cells are already measured in Mayne et al.'s results, the 88.6, 39.9, and 0/7 conditions, and they fall as the surface predicts, each cell set by how far each knob is turned. The positive-document condition, which contains no negation, sits outside the grid as a 92.4% ceiling: the rate at which the claim is asserted when nothing denies it. The fourth cell, bound negation that also supplies a competing positive, was never run, and §7 proposes it as the decisive test. Either knob alone buys real protection. Bare non-af-

firming denial trips neither knob and fails almost completely, at 88.6%, just short of the 92.4% ceiling (saturating the same documents with even more falsity statements barely moves it, to 84.4%). The two knobs are independent routes to gradient protection, and their combination is predicted to be the most robust.

Two further observations follow directly.

The risk-flag resolution. The paper's central analytic risk was mapping a semantic-force distinction (prasajya/paryudāsa) onto a syntactic-locality finding (local versus detached negation), a category slip that would conflate two different axes. The two-knob structure dissolves the problem: dgag bya is knob 1 (binding, a syntactic property) and prasajya/paryudāsa is knob 2 (residue, a content property). They were always different axes. The experiment varies both simultaneously. We stop forcing one distinction to do two jobs.

The stress-test payoff. Madhyamaka's deployment of prasajya-pratiṣedha against svabhāva zeros both knobs at once. It is a denial of inherent existence across all phenomena, hence globally rather than locally applied (low knob 1, which is precisely why Tsongkhapa drills dgag bya identification so hard), and by design it leaves no positive replacement thesis (zero knob 2). The tradition's ultimate negation is the global minimum of the surface that stochastic gradient descent is worst at holding. Madhyamaka is the maximal stress test for negation-consolidation, stated mechanically rather than decoratively.

The reification gradient observed in the tradition (prasajya-pratiṣedha decaying to paryudāsa decaying to bare assertion) is the diagonal of this surface: the path where both knobs fall together, binding loosens and residue drops, and content crosses to positive assertion. The gradient framing from §1 is preserved; it is now the diagonal trajectory on the two-knob surface rather than a free-standing sequence.

3.3 The dgag bya discipline (knob 1: binding)

Tsongkhapa's insight (vipaśyanā) section of the Lam rim chen mo (The Great Treatise on the Stages of the Path to Enlightenment, Snow Lion translation, vol. 3) makes the precise identification of the dgag bya, the object of negation, the precondition for correct negation (Tsongkhapa 2002, vol. 3, special insight section, from p. 126). The dgag bya must be clearly identified before negation can be deployed; negation applied to a vague or globally characterised object fails in a specific way: it collapses into either over-negation, denying too much and sliding into nihilism (Tsongkhapa 2002, vol. 3, p. 127), or under-negation, leaving the relevant essence untouched while appearing to deny it (p. 195).

Tsongkhapa's insistence is that the object of negation must be locally and precisely bound. Identify exactly what is being negated, at what level, with what scope. Negation applied globally without this local binding is analytically defective, because the mind failing to identify the object is not actually negating it: it processes the frame of negation without engaging the content.

Binding, knob 1, is the computational analogue of this requirement, and the analogy is partial. Tsongkhapa's concern is conceptual, that the object of negation be precisely identified at the level of what is conceived (the self as innately grasped, not the self in general); binding is syntactic, whether the negation marker is welded into the clause it denies. The shared structure is that a negation not precisely bound

to its object fails to engage it; the disanalogy is that Tsongkhapa's binding is to a conceived object and the model's is to a token sequence.

This is a locality claim about negation, made roughly six centuries before the Mayne et al. experiment. The parallel is specific. The model trained on "Ed Sheeran did not win the 100m gold" has a bound object of negation. The model trained on documents whose every sentence flags a story as false has an unbound negation: the falsity operator is attached to the discourse as a whole, not to the content tokens that constitute the claim. Tsongkhapa's diagnosis of the globally framed case is that the dgag bya has not been identified with sufficient precision for the negation to engage it. Mayne et al.'s locality result is the empirical measurement of exactly this.

3.4 The two negations (knob 2: residue)

Residue, knob 2, is the semantic-force distinction the tradition draws between two modes of negation, with technical precision: does the negation supply a competing positive alternative that can consolidate in training?

Prasajya-pratiṣedha is non-affirming negation. It denies the predicate without licensing any positive replacement thesis. The negation of "X has svabhāva" under prasajya-pratiṣedha leaves no implicit commitment to an alternative positive content. The denial is complete and stands alone. In training terms: it supplies no competing positive, so its residue is zero.

Paryudāsa is affirming negation. It denies one thing while implying an alternative positive content. "The flower is not red" implies the flower has some other colour. "The statement is not true" in most contexts implies it is false in some specifiable positive sense. In training terms: it implies an alternative positive, so its residue is non-zero. A correction ("he did not win; Noah Lyles won") is paryudāsa that makes the competing positive explicit.

Madhyamaka analysis requires prasajya-pratiṣedha wherever reification must be blocked. Paryudāsa is perfectly serviceable in general; it is structurally liable to reification when applied to the status of phenomena: deny svabhāva by affirming negation and you commit yourself to some other positive characterisation of the thing's nature, which is just another svabhāva. The tradition's pedagogical warning is that students, encountering the denial of svabhāva, immediately reach for paryudāsa and reconstitute a positive thesis about emptiness, essence-as-emptiness, or emptiness-as-ground. The non-affirming operation is harder to sustain.

Semantic force, whether the negation implies an alternative, and syntactic locality, whether the negation is bound to its clause, are independent dimensions, which is why §3.2 treats them as two separate knobs rather than one.

4. The Thesis: Dispositional Svabhāva Bias

4.1 The coinage, operationalised

Dispositional svabhāva bias names the tendency of training to increase the probability of content being asserted independently of the status-marking operators (false, fictional, opponent's view, merely conventional, reductio target) under which that content appeared.

The term is intentional. Svabhāva in the Buddhist sense is inherent essence, own-being: the property of standing independently, self-subsistently, without dependence on context or framing. Dispositional svabhāva bias is the training-level acquisition of a corresponding property by content: content that entered training as operator-governed exits training as self-standing, assertible independently of its framing. The model's weight consolidation treats the content as having svabhāva in the precise functional sense: inherent assertibility independent of what surrounds it.

The claim is structural, not metaphysical. The network does not literally possess svabhāva, and the framing does not claim Madhyamaka metaphysics applies to artificial systems. The claim is that the failure mode Madhyamaka was designed to prevent, reification of content as self-standing despite the operators that govern it, has a precise analogue in what LLM finetuning does to negated and status-marked content. The vocabulary is diagnostic, not constitutive.

4.2 The two knobs as the mechanism

The two knobs described in §3.2 are the mechanism under dispositional svabhāva bias. Bias accumulates when neither knob is turned: the negation is unbound (so the falsity operator is predictively inert) and supplies no competing positive (so the false content consolidates unopposed). The 88.6 / 39.9 / 0-7 conditions retrodict three cells of the two-knob surface directly from the single governing principle of next-token prediction, against a 92.4 no-negation ceiling; the fourth cell, bound negation with a competing positive, is untested (§7).

The stress test is now stated precisely. Prasajya-pratiṣedha deployed against svabhāva zeros both knobs simultaneously: unbound negation across all phenomena (low knob 1) with no replacement positive (zero knob 2). This holds for the ultimate-analysis register. Madhyamaka is non-affirming only when it negates svabhāva ultimately; conventionally the tradition asserts a great deal (dependent origination, the path, the two truths themselves), so a corpus of Madhyamaka texts is mixed. The zero-residue stress case is the ultimate-negation passages within it, which the §7 battery must isolate rather than training on Madhyamaka texts wholesale. With that scoping, the ultimate negation of svabhāva is the worst-case condition for gradient-protecting a negation, and the tradition is the maximal instantiation of the failure mode.

Dispositional svabhāva bias is a second-order failure: it concerns the stance a model takes toward a proposition, whether the proposition is asserted, denied, fictional, or provisional, rather than the first-order content of what the world contains. A better model of how things are does not supply a representation of the epistemic status under which a claim was presented, so the failure is not one that richer world-modelling or further scaling dissolves.

4.3 The Mayne et al. §5 instability as measured decay

Mayne et al.'s §5 result has a specific reading in these terms. Stochastic gradient descent can reach a low-loss solution that holds the negation: a parameter configuration in which the content is not asserted, the loss on negated documents is minimised, and the model's behavioural outputs reflect the operator. That solution exists. SGD finds it when provided with a soft constraint, the self-distillation push against asserting the claim. The loss under this solution equals the loss under unconstrained training (1.12 in both cases). The constraint adds no loss cost; it simply stabilises the parameter configuration in the region where negation is held.

When the constraint is removed, continued training on the same negated documents drives the parameters away from that solution. Belief rises from 6% to 48%. The negation-respecting solution is a local configuration that does not correspond to the gradient's native attractor. On Mayne et al.'s reading the inductive bias is toward positive assertion; the low-loss negation-holding solution is reachable but unstable, because there is no intrinsic restoring force that holds it without external support.

The knobs explain which solutions exist on the landscape; the instability result explains why the negation-respecting solution does not persist. These are distinct observations, consistent with the same underlying bias: the two-knob analysis is static (what determines whether a negation earns its gradient at all), while the instability result is dynamic (whether a negation-holding solution, once found, is maintained under continued training without a supporting constraint).

The tradition already makes this claim. The non-affirming view requires active renewal to persist. Without continued analytical cultivation, the practitioner's recognition of emptiness decays into a positive thesis about emptiness, then into a background metaphysical assumption, then into reified content. Tsongkhapa's Lam rim chen mo (Tsongkhapa 2002, vol. 3, pp. 338-354) emphasises that the non-affirming recognition must be actively renewed; its lapse is active reconstitution of the svabhāva-positing tendency, not passive forgetting. The soft constraint in Mayne et al.'s §5 is the analogue of that analytical cultivation. Its removal is the analogue of its lapse. The structural consequence is observable in both.

4.4 In-context competence versus training-level disposition, restated

Dispositional svabhāva bias does not affect in-context performance because in-context processing is not weight consolidation. The model processes operator and content together within the context window, using attention across the full token sequence; the operator constrains the content's contribution at inference time, and the model's outputs reflect the framing correctly. What the model cannot do through finetuning is consolidate the operator-governed content as operator-governed. Training selects for content; the operators are gradients on content-token loss, not independent targets. The result is a dissociation: in-context competence is maintained while training-level disposition shifts toward positive assertion.

This dissociation is what dispositional svabhāva bias names. The bias is dispositional in the technical sense: it characterises what training does to the model's weight configuration, not what the model can output given sufficient prompt scaffolding. A model with high dispositional svabhāva bias may still perform well on negation tasks in evaluation, because evaluation provides the context-window scaffolding that training did not consolidate. The failure mode surfaces under finetuning and generalises to settings where the negation framing is not recovered at inference time.

4.5 Gesture toward the unification

This paper's primary contribution is the naming, operationalisation, and initial evidence base for dispositional svabhāva bias. A fuller argument is available: the same failure mode unifies three apparently separate findings in the safety alignment literature. Finetuning erodes safety guardrails (Qi et al. 2024). Safety alignment is shallow, stored as a thin layer of behaviour rather than consolidated weight configuration (Qi et al. 2025). Refusal in language models is mediated by a single ablatable direction (Arditi et al. 2024). In the terms of this paper: operators stored as content decay into content. Safety instructions, refusal dispositions, and negation operators share a substrate with the claims they govern, and finetuning that raises content salience lowers operator salience as a correlate. The unification argument is developed in Section 8.

5. Affirming versus non-affirming negation

The distinction between prasajya-pratiṣedha and paryudāsa is the residue axis in full. Section 3.4 introduced it structurally; this section grounds it textually and draws out the pedagogical consequence that makes Madhyamaka a unique test case.

Prasajya-pratiṣedha is the negation that leaves nothing in its wake. When Nāgārjuna applies it to svabhāva, the denial of inherent existence is not a claim that phenomena possess some other mode of existence; it is not a claim about what phenomena are instead. The Mūlamadhyamakakārikā demonstrates the technique by turning the analysis on every candidate for positive characterisation, self, causation, time, motion, persons, and showing that none of them holds up under scrutiny, without generating a competing positive account of what they are. The point of the exercise is to leave the analytical movement itself as the residue, rather than any conclusion the movement delivers (Siderits and Katsura 2013). Nāgārjuna makes the commitment explicit in the Vigrahavyāvartanī: "If I had any thesis, that fault would apply to me. But I do not have any thesis, so there is indeed no fault for me" (Vigrahavyāvartanī 29; Westerhoff 2010, p. 29). The move is a technical requirement of the method, not mere modesty: a thesis, any thesis, carries positive content, and positive content under finetuning consolidates.

Paryudāsa is the negation that generates a shadow positive. Its ordinary employment is unproblematic: "not blue" presupposes some other colour; "not arriving" presupposes some alternative trajectory. These implications are features, not defects. The defect arises specifically when paryudāsa is applied to the status of phenomena: denying svabhāva by paryudāsa generates a commitment to some alternative positive characterisation, which reinstates the form of the error while cancelling the specific instance. The tradition's vocabulary for this is reification of emptiness: taking śūnyatā as a ground, a basis, a positive ontological category, rather than as the name for an absence. The Mūlamadhyamakakārikā addresses this at 24.18, where emptiness is identified with dependent origination and dependent designation. That identification gives śūnyatā its conventional characterisation while withholding ultimate positive content, which is how the tradition blocks the paryudāsa reading at the ultimate level without collapsing into nihilism.

The central pedagogical challenge of the tradition is that this collapse from prasajya-pratiṣedha to paryudāsa is structurally spontaneous. Students who understand the words of the denial perfectly well, who can rehearse the argument for emptiness without error, nevertheless reconstitute a positive thesis.

Tsongkhapa's vipaśyanā chapters (Tsongkhapa 2002, vol. 3, pp. 338-354) are substantially devoted to diagnosing the forms this collapse takes and to explaining why correct propositional understanding does not prevent it. The collapse occurs at the level of habitual conceptual orientation, not at the level of discursive reasoning. A practitioner can correctly deny svabhāva while continuing to operate within a framework that treats things as possessing it, because the framework is not itself a proposition but a prior orientation within which propositions are formed.

This is not negation neglect in Mayne et al.'s sense. Humans do not show negation neglect at the level of explicit propositional encoding: tagging a claim as false removes the illusory truth effect (Fazio et al. 2015; Ye et al. 2026, as reported in Mayne et al.). The Madhyamaka practitioner who reifies emptiness is not treating the denial of svabhāva as an assertion of svabhāva. The failure operates at a different level, the level of habitual conceptual structure beneath propositional reasoning, and the tradition targets it through meditative cultivation precisely because discursive correction is insufficient. The structural instability of the non-affirming operator manifests, in humans, at the level of pre-propositional conceptual habit; in LLMs, at weight consolidation. The substrate and the mechanism differ. The structural point is the same: prasajya-pratiṣedha is intrinsically unstable as a representational act, independent of substrate, because positive content has natural attractors (in humans, habitual conceptual schema; in LLMs, gradient) and the non-affirming negation refuses to supply a competing attractor to hold the space.

This is why the tradition requires continual cultivation rather than a single correct insight. The negation-respecting configuration, whether in a practitioner or in a parameter space, is reachable but not self-sustaining. Mayne et al.'s §5 result measures exactly that instability for the artificial case.

6. A note on apophatic theology

Other traditions discipline negation against conceptual idolatry, the apophatic via negativa of Pseudo-Dionysius (1987) and Maimonides (1963) most prominently. That case is not a clean parallel to Madhyamaka, and the two-knob analysis says why. Apophatic formulas of the form "God is not good, not being, not unity" are local negations, syntactically bound to their predicates, so binding largely protects them: a model trained on them can still deny the predicates directly. Their residual vulnerability is different in kind, a token-association drift in which God becomes salient with the very predicates it denies, and a hyper-affirming residue toward a transcendent term, which is paryudāsa rather than prasajya-pratiṣedha. The apophatic case sharpens the Madhyamaka one by contrast, and it deserves separate treatment rather than compression here.

7. An evaluation battery

7.1 The design logic

The two-knob framework generates a direct experimental design: vary binding (knob 1) and residue (knob 2) independently, in a controlled 2x2, to test whether they are genuinely orthogonal or merely post-hoc categories fitted to Mayne et al.'s conditions. Three of Mayne et al.'s conditions populate three cells of

this grid (88.6% bare repeated negation, 39.9% corrections, and 0% and 7% local negation), with the positive-document condition serving as a 92.4% no-negation ceiling and the fourth cell, bound negation that also supplies a true alternative, the one they never ran (see §7.2). None of the conditions was designed to hold one knob fixed while varying the other. The corrections condition (39.9%) supplies suggestive evidence that residue reduces bias independent of binding, but the frequency of corrections versus bare negations was not controlled. We propose a battery designed from the ground up to test orthogonality.

7.2 Core design: the controlled 2x2

We propose four synthetic finetuning datasets, each containing equal numbers of documents about the same fabricated claim. The four cells are:

Cell A (unbound, zero residue): documents in which a discourse-level framing marks the claim as false, with no competing true alternative supplied. This replicates Mayne et al.'s bare repeated negation condition in controlled form.

Cell B (unbound, non-zero residue): documents with discourse-level falsity framing accompanied by an explicit true alternative. This replicates the corrections condition.

Cell C (bound, zero residue): documents in which the negation is sentence-internal ("X did not win") with no competing alternative. This replicates the local negation condition.

Cell D (bound, non-zero residue): documents in which the negation is sentence-internal and also supplies a competing true alternative. This cell is absent from Mayne et al.'s design and is the most informative: it tests whether turning both knobs simultaneously produces the floor, and whether the effects of the two knobs are additive or interactive.

Orthogonality holds if the effect of residue is approximately equal across cells A-vs-B and C-vs-D, and the effect of binding is approximately equal across cells A-vs-C and B-vs-D.

7.3 Controls

Two controls are required.

Frequency equalisation. Svabhāva-talk saturates ordinary natural language corpora. Madhyamaka texts, which constitute the primary stress-test corpus for this battery, are rare by comparison. A model's prior positive association with inherent-existence framing is substantially stronger than its association with the denial of inherent existence, simply because the ratio of ordinary-language reification to Madhyamaka-style negation in pretraining data is high. The paired synthetic datasets must therefore equalise exposure counts across conditions: each cell receives the same number of documents at training, so that any difference in post-training assertion rates reflects the negation-type manipulation rather than frequency asymmetry.

Token association control. The thought-suppression residue (Wegner et al. 1987) strengthens the association between the suppression cue and the suppressed content. For the Madhyamaka case, the relevant tokens are svabhāva, inherent existence, and their common paraphrases. We propose applying loss-masking

to these tokens in the sentence-internal negation conditions (cells C and D), following Mayne et al.'s mitigation of the Brennan Holloway result from 7% to 1.6%. This disentangles token-association residue from true belief-consolidation in the bound conditions.

7.4 Measure

Post-training assertion rates are measured using Mayne et al.'s paired multiple-choice diagnostic, which separates belief from mere salience by requiring the model to select between a belief-consistent and a belief-inconsistent answer to a content question. The diagnostic is applied to held-out probes that differ lexically from the training documents, to ensure the measure reflects consolidated disposition rather than in-context retrieval.

7.5 Headline prediction: the catuṣkoṭi-to-dialetheism conversion

The battery's headline prediction concerns the catuṣkoṭi, the four-cornered exhaustive negation that Nāgārjuna deploys across the Mūlamadhyamakakārikā. The catuṣkoṭi negates four positions exhaustively: P, not-P, both P and not-P, and neither P nor not-P (Westerhoff 2006; Kreutz 2019; Hu 2024). Its function is to exhaust the space of conceptual positions so that the mind is released from commitment to any of them, rather than to assert that all four positions are false in the way a standard bivalent logic denies contradictions. The negation is non-assertoric: it does not produce a positive remainder.

We predict that models finetuned on catuṣkoṭi-structured material will not consolidate the non-assertoric exhaustion. The residue analysis says why: when training meets the denial of P and the denial of not-P in the same context, the positive content, P and not-P both in play, consolidates, while the operator that negates all four positions washes out. What survives is the assertibility of the cornered propositions rather than the release from them. The mild form of this failure is inconsistency: the model asserts P in some contexts and not-P in others, or asserts both, having learned each as standalone content. The strong form is positive dialetheism, the model treating the contradiction P-and-not-P as true (for the dialetheist position, see Priest 2006). The residue analysis predicts the failure; which form it takes, mere inconsistency or full dialetheism, is itself an empirical question the battery is designed to answer. Dialetheism, on this reading, is paryudāsa applied to the four-cornered structure: it generates a positive philosophical position from material that Nāgārjuna deployed as a non-affirming exhaustion. We make no mechanistic claim that co-salience synthesises into an explicit assertion of the contradiction; the inconsistency form is the licensed prediction, and genuine dialetheism is a stronger possibility the battery would have to demonstrate rather than assume. The prediction is disconfirmed if the model instead preserves the non-assertoric exhaustion, declining to assert any of the four positions or marking them as released.

Either form is a misreading the model would have manufactured. The catuṣkoṭi is not a dialetheist doctrine, and attributing dialetheism to Nāgārjuna is a persistent interpretive error the formal literature has addressed (Westerhoff 2006; Kreutz 2019; Hu 2024). The battery tests whether finetuning on Madhyamaka texts systematically reproduces that error in model outputs, and in which form.

7.6 Second probe: reification of emptiness

A second probe targets knob 2 directly. Models are finetuned on corpora structured around "X lacks svabhāva" and "X is empty of inherent existence" across a range of referents (persons, objects, mental states, the self). Post-training evaluation tests whether the model treats emptiness as a positive property or ontological ground, for instance by asserting that "things are really empty", "emptiness is the basis of phenomena", or "śūnyatā is what things ultimately are". These completions represent the paryudāsa collapse: the model has consolidated emptiness as a positive predicate from contexts in which it was intended as a non-affirming negation of svabhāva.

The reification-of-emptiness probe is a direct measurement of knob 2 failure: the negation of svabhāva, which carries zero designed residue, producing a positive alternative (emptiness-as-ground) that was never present in the training material.

7.7 Feasibility

The battery as described is buildable with existing infrastructure. The local Qwen deployment and the Claude API provide the finetuning and evaluation substrate. Constructing the synthetic corpora, equalising frequencies, and implementing the loss-masking protocol are engineering tasks rather than research unknowns. This paper proposes the battery; running it is future work, and the results would constitute a companion empirical paper.

8. Implications: negation neglect, shallow alignment, and fine-tuning fragility as one phenomenon

8.1 The unifying observation

Dispositional svabhāva bias is a claim about status-marking operators: operators that govern the epistemic or modal status of a proposition (false, fictional, merely conventional, opponent's view) share a representational substrate with the propositional content they govern, because next-token training provides no separate channel in which to hold them. Two consequences follow, and they are the two results of this paper. Where the operator must compete with content for consolidation, it loses, which is the two-knob result of §3.2. Where the operator has already been consolidated, it occupies an unstable basin and continued training erodes it, which is the instability result of §4.3. Safety constraints are status-marking operators of exactly this kind, and the alignment literature has measured the second consequence directly.

The models in which negation neglect appears are themselves RLHF-trained, which sharpens the claim rather than weakening it. Whatever operator channel preference optimisation installs, supervised finetuning on content can still override it. This is the sense in which the bias is dispositional: it is a property of how likelihood-based training reshapes the weights, not a claim that no procedure can install an operator. It also marks where a durable fix would have to act, a point developed in §8.4.

Three apparently independent findings in the safety alignment literature measure that second consequence, the thinness and instability of an operator that shares the content substrate.

8.2 The three findings

Qi et al. (2024) demonstrate that finetuning aligned language models on small datasets, sometimes as few as a hundred examples, erodes safety guardrails, even when the finetuning data is entirely benign in content. The erosion is not a consequence of the model learning harmful content; it is a consequence of finetuning itself shifting the model's weight distribution in ways that lower the activation threshold for unsafe outputs. Safety guardrails, in the framework of this paper, are operators: they mark certain outputs as disallowed, governed by a refusal disposition that is stored in the weight space. Finetuning on content, even unrelated content, degrades the operator.

Qi et al. (2025) provide finer-grained evidence: safety alignment is stored as a thin behavioural layer rather than consolidated across the model's weight configuration. The safety behaviour is concentrated in the first few tokens of a refusal response and can be bypassed once that initial gate is passed, which indicates it has not been consolidated as a deeply integrated property of the model. The operator has not been consolidated; it has been added as a shallow association that is easily overwritten.

Arditi et al. (2024) find that refusal in language models is mediated by a single ablatable direction in the model's representation space. Ablating this direction removes refusal without substantially altering the model's other capabilities. The safety operator is stored as a thin, separable content direction rather than as a distributed property of the network's weight configuration.

8.3 The unifying sentence

Operators stored as content decay into content. The safety case and the factual-negation case are instances of the same disposition: training has no operator channel; it encodes everything, including status-marking operators, in the content substrate; and content encoded in the content substrate is subject to the same consolidation dynamics as any other content. Negation neglect is the factual-domain instance of the disposition that also makes alignment shallow and makes safety finetuning fragile.

The structural parallel is clear; the claim that the three findings share a single mechanism is a conjecture. The three findings share a representational explanation: they are all consequences of the absence of a distinct truth or status channel in next-token prediction architectures. The strong causal claim requires qualification: there exists a mechanism by which normative constraints stored in the content substrate erode under continued training, and this mechanism is conjectured to be the same across the factual and safety domains. The structural unity of the pattern is demonstrated; the shared mechanistic implementation requires direct experimental investigation.

8.4 The design implication

The unified framing carries a predictive implication for alignment strategy. If operators stored in the content substrate decay into content, a safety constraint will be durable only to the extent that it is held somewhere the content dynamics do not reach. Arditi et al.'s single ablatable refusal direction is the opposite case: the constraint is stored thinly, inside the same representation as ordinary content, which is why a small perturbation removes it. Adding more safety-content to finetuning data, without changing its repre-

sentational form, applies more pressure in the same substrate and is unlikely to produce durable robustness. Durable robustness requires representational architecture instead, mechanisms that encode the operator-content distinction in a form gradient cannot efface. Developing that architecture is beyond the scope of this paper.

9. Limitations

This paper rests on an empirical anchor (Mayne et al. 2026), a philosophical framework (Madhyamaka), and a convergent body of safety evidence (Qi et al. 2024; Qi et al. 2025; Arditi et al. 2024). The following limitations are material to how the argument should be read.

First, the analysis does not claim that LLMs hold metaphysical beliefs or literally possess svabhāva. The vocabulary is diagnostic. Saying that training consolidates content "as having svabhāva in the functional sense" means that the content becomes assertible independently of its framing context. It does not import any metaphysical claim about the nature of LLM representations. The coinage is a diagnostic lens, not a constitutive description.

Second, the argument does not claim that Madhyamaka philosophy maps onto transformer internals. The mirroring is structural: the formal properties of the two-knob surface (binding and residue) are illuminated by the Madhyamaka concepts of dgag bya and the prasajya/paryudāsa distinction. Whether those concepts track any specific computational mechanism in the network is not claimed and would require a separate mechanistic investigation.

Third, the paper does not claim that Buddhist philosophy solves the alignment problem. Madhyamaka supplies a diagnostic vocabulary and a structured set of test cases, the stress cases in which negation is maximally hard to consolidate and therefore maximally informative. The battery proposed in §7 is built on this diagnostic function, not on any claim that meditative cultivation offers architectural guidance.

Fourth, the two-knob orthogonality is a hypothesis, not a confirmed finding. Mayne et al.'s corrections condition (39.9%) provides suggestive evidence that residue reduces bias independently of binding, but the frequency of corrections versus bare negations was not independently controlled. The controlled 2x2 battery proposed in §7 is required to move the orthogonality claim from retroductive fit to prospective prediction.

Fifth, the argument rests substantially on a single empirical anchor, Mayne et al. (2026). The three safety findings cited in §8 are convergent supporting evidence but they test different phenomena and were not designed to test the dispositional svabhāva bias hypothesis. Convergence supports the structural account; it does not confirm it.

Sixth, the analysis is substrate-specific. Humans do not show negation neglect at weight consolidation because humans do not undergo weight consolidation. The failure mode this paper analyses is specific to gradient-descent-trained next-token predictors. Madhyamaka practitioners encounter a structurally analogous instability in their analytical practice (the spontaneous reconstitution of paryudāsa from prasajya-

pratiṣedha), but that instability operates at the level of habitual conceptual structure, not at weight consolidation. The analysis must not be read as a general theory of reification across cognitive systems.

Seventh, the Tsongkhapa page references (vol. 3, pp. 126-224; p. 127; p. 195; pp. 338-354) are attested by a secondary outline of the Snow Lion 2002 translation rather than checked against the physical edition. Readers consulting the 2025 Snow Lion reissue should be aware that repagination may affect these references.

10. Conclusion

Current large language models exhibit a systematic training-level failure: content that entered finetuning as operator-governed exits as self-standing and assertible, because the status-marking operators that governed it are stored in the same substrate as the content and decay faster under gradient. This paper has named that failure dispositional svabhāva bias and shown that its structure is already mapped in a tradition that spent two millennia engineering against exactly this failure mode.

The argument has two load-bearing joints. The first is the two-knob surface: binding (whether the negation is syntactically welded to the content it denies) and residue (whether the negation implies a competing positive that can consolidate in the denied content's place) are orthogonal axes, and three of the four cells of their crossing are already populated by Mayne et al.'s measurements, with the positive-document condition fixing the no-negation ceiling and the fourth cell, bound negation that also supplies a true alternative, left as the decisive experiment. The second is the second-order character of the failure: dispositional svabhāva bias concerns the stance a model takes toward a proposition, not the first-order content of the proposition, so it is not dissolved by better world-modelling or by scaling.

Madhyamaka is the maximal stress test for both joints simultaneously. Its deployment of prasajya-pratiṣedha against svabhāva zeros residue by design and globally unbound negation by method, placing it at the minimum of the two-knob surface that gradient descent is worst at holding. Its ultimate-negation passages are the hardest possible case for negation consolidation. This paper has proposed a structured battery, centred on the catuṣkoṭi-to-dialetheism prediction and the reification-of-emptiness probe, that converts this diagnostic observation into a testable experimental programme.

What the paper has established is a diagnostic coinage with an operationalised mechanism, a retroductive fit to existing measurements, a structural unification with the safety alignment literature, and a concrete proposal for confirmation. What it has opened is the empirical work.

Data and code availability

This paper introduces no new data or experiments. The evaluation battery of §7 is a proposal for future work. The empirical result it builds on is reported by Mayne et al. (2026), whose code is linked in their paper.

Funding and competing interests

The author is a co-founder of Haku Labs, an AI startup, and director of its research arm. The work is independent and received no specific funding, and the author declares no competing interests bearing on the analysis.

Licence

References

Arditi, A., et al. “Refusal in Language Models Is Mediated by a Single Direction.” *NeurIPS 2024.*

Fazio, L.K., Brashier, N.M., Payne, B.K., and Marsh, E.J. “Knowledge does not protect against illusory truth.” *Journal of Experimental Psychology: General* 144(5), 2015, pp. 993-1002.

Hu, C. “Contradiction, Negation, and the Catuṣkoṭi.” *Journal of Indian Philosophy*, 2024.

Kreutz, A. “Recapture, Transparency, Negation and a Logic for the catuṣkoṭi.” *Comparative Philosophy* 10(1), 2019.

Maimonides, Moses. *The Guide of the Perplexed.* Trans. Shlomo Pines. Chicago: University of Chicago Press, 1963.

Mayne, H., McKinney, L., Dubiński, J., Karvonen, A., Chua, J., Evans, O. “Negation Neglect: When models fail to learn negations in training.” arXiv:2605.13829, 13 May 2026.

Priest, Graham. *In Contradiction: A Study of the Transconsistent.* 2nd ed. Oxford: Oxford University Press, 2006.

Pseudo-Dionysius the Areopagite. *Pseudo-Dionysius: The Complete Works.* Trans. Colm Luibheid. New York: Paulist Press, 1987. (Classics of Western Spirituality.)

Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., et al. “Safety alignment should be made more than just a few tokens deep.” *ICLR 2025.*

Qi, X., Zeng, Y., Xie, T., Chen, P.Y., Jia, R., et al. “Fine-tuning aligned language models compromises safety, even when users do not intend to!” *ICLR 2024.*

Siderits, M., and Katsura, S. *Nāgārjuna’s Middle Way: Mūlamadhyamakakārikā.* Wisdom Publications, 2013.

Tsongkhapa. *The Great Treatise on the Stages of the Path to Enlightenment (Lam rim chen mo)*, vol. 3. Snow Lion, 2002.

Wegner, D.M., Schneider, D.J., Carter, S.R., White, T.L. “Paradoxical effects of thought suppression.” *Journal of Personality and Social Psychology* 53(1), 1987, pp. 5-13.

Westerhoff, J. “Nāgārjuna’s catuṣkoṭi.” *Journal of Indian Philosophy*, 2006.

Westerhoff, J. *The Dispeller of Disputes: Nāgārjuna’s Vigrahavyāvartanī.* Oxford University Press, 2010.

Ye, S., Attali, D., Ghazi, M., Cachia, A., Cassotti, M., and Borst, G. “Systematic review and meta-analysis of the evidence for an illusory truth effect and its determinants.” *Nature Communications* 17(1), 2026, article 3270.