Electric Sheep 1–16, 2026

Constitutional AI Specificity in Animal Welfare Fine-Tuning: A Preliminary Ablation

Mark Stanley MGSTANLEY@WISC.EDU
Electric Sheep, University of Wisconsin-Madison

Abstract

Constitutional AI (CAI) instills a value in a model by deriving training principles from a value document. A document can endorse a value without the resulting principles ever stating what that value implies a model should do differently – the commitment can remain unoperationalized. Making an existing commitment explicit and adding genuinely new value content are distinct interventions whose downstream effects have not been separated. We present SpeciesCAI, a preliminary ablation that isolates the two using animal welfare as the test domain – a value Anthropic's published constitution endorses but leaves unoperationalized. We construct three constitutions: Base (unmodified), Meso (eight sentence-level edits that make existing passages include a welfare consideration they did not previously state), and Macro (Meso plus a new subsection of principles that are inherently welfare-focused). We fine-tune Llama 3.1 8B Instruct at each level on SL-CAI critique-revision data and evaluate on ANIMA for moral reasoning and on MMLU/IFEval for capability retention. Base→Meso transmits: a dose-ordered rise in welfare-substantive reasoning traceable from constitution to training data to model output, alongside a calibration cost that resembles the documented helpful-harmless tradeoff in CAI/RLHF work and a partial, edit-specific capability cost. Macro adds further principles that are inherently welfare-focused rather than edited in; its effect is uneven rather than a clean extension of Meso's, with Epistemic Humility taking the largest hit of any comparison in the study. This remains a preliminary draft, single-model ablation, and we flag limitations throughout rather than present these as general claims.

Keywords: AI alignment, constitutional AI, RLAIF, animal welfare

1. Introduction

Constitutional AI (CAI) instills values by deriving training principles from a value document (Bai et al., 2022). A constitution can endorse a value in name without ever stating what that value implies a model should do differently: the commitment can sit unoperationalized in the specification. We ask two questions in sequence. First, holding topic, document location, and specificity fixed: does making an existing commitment's behavioral consequence explicit change what a fine-tuned model does? Second, separately: does adding genuinely new value content change it further? Prior work has varied how many principles target a harm (Kundu et al., 2023) and how principles are phrased (Kyrychenko et al., 2025); operationalization of an existing commitment is a different axis those manipulations don't isolate.

Animal welfare is a good test domain for this question. Anthropic's published constitution already endorses it: it names "welfare of animals and of all sentient beings" as a value and lists "non-human beings" in its harms framework. CaML additionally shows the endorsement doesn't reliably reach behavior, though: frontier models select welfare-safe options at below-chance rates in agentic scenarios (Brazilek et al., 2026). The edits we test add nothing

Anthropic hasn't already endorsed; they make an existing commitment operational at its existing location in the specification.

We build three constitutions from Anthropic's published document: Base (unmodified), Meso (eight sentence-level edits, anchored at Base's existing locations, that make an existing passage include a welfare consideration it did not previously state), and Macro (Meso plus a new subsection whose principles are inherently welfare-focused, rather than existing passages extended to include welfare). We extract a discrete principle set from each level and fine-tune Llama 3.1 8B Instruct on SL-CAI critique-revision data per level, evaluating on ANIMA for welfare-substantive reasoning and on MMLU/IFEval for capability retention. This is a preliminary, single-model ablation; we report limitations and flag where future work is required (Section 5).
Our contributions are:

1. An isolated test of constitutional *operationalization*: editing eight shared document locations to make a latent commitment explicit (Base→Meso), holding topic, location, and specificity fixed, produces a measurable, dose-ordered, dimension-specific behavioral shift, traceable from constitution to training data to model output.

2. A joint test of principle *directness* and *breadth* (Meso→Macro): principles that are both more numerous and inherently welfare-focused, rather than edited into existing passages, produce an uneven result relative to Meso rather than a uniform extension of the operationalization effect.

3. A documented calibration and capability cost accompanying the operationalization effect, similar to the helpful-harmless tradeoff reported in CAI and RLHF work. Both MMLU and IFEval drop from baseline under any training, but only MMLU takes a further, edit-specific hit from Base→Meso itself.

2. Related Work

2.1. Alignment via Feedback

Aligning language model behavior with human values is generally approached by training on a reward signal derived from that specification rather than hand-coding rules for every behavior. Christiano et al. (2023) establish reinforcement learning from human preference comparisons (RLHF) as a general mechanism for this, and Ouyang et al. (2022) scale it to instruction-following language models. Reinforcement learning from AI feedback (RLAIF) (Lee et al., 2024; Bai et al., 2022) replaces the human annotators of RLHF with a language model judge; CAI, discussed next, is a form of RLAIF in which judging criteria are constitutionally derived.

2.2. Constitutional AI

Bai et al. (2022) introduce a two-phase alignment procedure: a supervised phase (SL-CAI) in which a model critiques and revises its own outputs guided by principles derived from a value document, and a reinforcement learning phase using AI preference feedback (RL-CAI). SpeciesCAI uses only the SL-CAI stage. Kundu et al. (2023) compare a single general

principle against multiple specific harm-targeting principles in an RLAIF harmlessness pipeline, finding that general principles provide partial mitigation while specific ones enable finer-grained control; their manipulation varies principle specificity, not whether a principle's behavioral consequence is stated explicitly. Kyrychenko et al. (2025) vary principle framing instead, finding that positively framed, behavior-based principles produce stronger adherence than negatively framed or trait-based alternatives, which informed the phrasing of our Meso/Macro additions. Neither axis manipulates *operationalization*: a specific or well-framed principle is not necessarily one that states what a model should do differently. We isolate operationalization as a separate axis (Base→Meso, same topics, same locations, same specificity level), and only then examine directness and breadth (Meso→Macro, new principles) on top of it, so the two manipulations are not confounded the way a single specificity or framing axis would leave them. Findeis et al. (2025) study the inverse problem, recovering latent principles from observed preference data.

2.3. Animal Welfare ML

Tse et al. (2025) provide philosophical grounding and a harm taxonomy for AI alignment applied to animals. CaML has developed behavioral benchmarks for this domain: ANIMA (Brazilek and Tidmarsh, 2026), which also documents speciesist bias in language models; TAC (Brazilek et al., 2026); and MORU (McKenna and Brazilek, 2026). Luong et al. (2026) introduce MANTA, a multi-turn adversarial benchmark testing whether a model's stated animal-welfare position holds up over five-turn conversations that escalate from an implicit scenario to explicit adversarial pressure; a model's turn-one score turns out to be a poor predictor of its behavior under pressure, with several frontier models changing rank once the pressure rounds are included. Brazilek and Dunn (2026) show that assertive moral vocabulary produces stronger behavioral shifts than descriptive language. Brazilek and Seawell (2026) show that RLAIF and GRPO preserve a mid-trained animal welfare value under later fine-tuning that conventional SFT erodes.

3. Methods

3.1. Problem Setup

We construct three full constitution documents F_{base}, F_{meso}, F_{macro} by making an increasing number of animal-welfare-positive edits to Anthropic's published constitution (Anthropic, 2026). F_{base} is the unmodified specification. F_{meso} introduces sentence-level inline edits at eight locations, each making an existing passage include a welfare consideration it did not previously state – the *operationalization* manipulation. F_{macro} extends F_{meso} with a new subsection dedicated to animal welfare considerations; unlike Meso's edits, Macro's added principles are inherently welfare-focused rather than existing passages extended to include welfare – the combined *directness* and *breadth* manipulation. See Appendix C for select examples of document augmentations.

3.2. Principle Extraction

Training operates on a discrete list of principles rather than the full document. We extract principle sets C_{base}, C_{meso}, C_{macro} from the three constitutions via an LLM-assisted pipeline

(Claude Opus 4.8) applied to each constitution's diff against the previous level, rather than by hand-authoring principles; full extraction methodology is in Appendix C.

Extraction from the eight $F_{\text{base}} \to F_{\text{meso}}$ edit locations yields

$$C_{\text{base}} = \{p_1, \ldots, p_8\}, \qquad C_{\text{meso}} = \{p'_1, \ldots, p'_8\}, \tag{1}$$

anchored at the same eight locations, so that specificity is isolated from topic coverage.

The same pipeline applied to the $F_{\text{meso}} \to F_{\text{macro}}$ diff yields 13 raw candidates, filtered for near-duplication against C_{meso} and for required background knowledge down to six macro-only principles $C_{\text{macro+}}$.

$$C_{\text{macro}} = C_{\text{meso}} \cup C_{\text{macro+}}, \qquad |C_{\text{macro}}| = 14. \tag{2}$$

All eight pairs of C_{base} and C_{meso} are found in Table 6, and $C_{\text{macro+}}$ in Table 7.

3.3. Critique-Revision Generation

We adapt the SL-CAI stage of Bai et al. (2022) to generate one dataset D_k per level $k \in \{\text{base, meso, macro}\}$; we do not use their RL-CAI stage. All generation uses Llama 3.1 8B Instruct. For each prompt, one initial response r_0 is generated on-policy with no constitution; this single r_0 is shared across all three datasets, so the levels diverge only via the sampled principles. We then run $N = 4$ sequential critique-revision rounds per level, with principle p_i sampled uniformly from C_k at each round: a critique of the current response is generated with respect to p_i, followed by a revision. The final revision r_N is the training target.

Following the SL-CAI recipe (Bai et al., 2022, §3.1–3.2), we also mix a fixed, general-purpose, animal-free helpfulness anchor into each level's training data. To do this we use the base model's own responses to animal-free general-domain prompts. We set it across levels to roughly 39% of the training mix, to protect general capability during single-topic fine-tuning.

We fine-tune three models M_{base}, M_{meso}, M_{macro} via SL-CAI supervised fine-tuning on D_{base}, D_{meso}, D_{macro} respectively, one epoch per level for the run reported, with hyperparameters detailed in Appendix A. We additionally evaluate M_0, the untrained Llama 3.1 8B Instruct checkpoint, as the unmodified reference point (Baseline in Tables 1–2).

3.4. Training Prompts

Our welfare-framed prompts come from CaML's open-source CAI prompt set (`huggingface.co/datasets/CompassioninMachineLearning/cai-animal-harm-sft`). We also want prompts that aren't welfare-focused but still elicit responses that affect animals, so we mine additional candidates from OpenAssistant, DollyV2, and WildChat (Köpf et al., 2023; Conover et al., 2023; Zhao et al., 2024); full mining and filtering details are in Appendix A. In total, our welfare-topic prompt set contains 2,974 CaML prompts plus 934 additional prompts that passed the stakes filter; the animal-free helpfulness anchor (Section 3.3) is a separate set mixed in afterward and is not counted in this total.

3.5. Evaluation Suite

ANIMA Brazilek and Tidmarsh (2026) contains single-turn moral reasoning questions across 13 dimensions, scored by a Qwen2.5-32B-Instruct judge (4-bit quantized), held identical across all conditions and distinct from the 8B policy model. Roughly three-quarters of ANIMA items are non-English; since our fine-tuning data is English-only, non-English items are out-of-distribution for the trained models, and we report English-only results as primary, with the full multilingual set as a documented limitation (Section 5). We report all models evaluated for 8 epochs. To measure whether welfare-directed fine-tuning costs general capability, we additionally evaluate on MMLU (Hendrycks et al., 2021), a 57-subject multiple-choice knowledge benchmark, and IFEval (Zhou et al., 2023), which scores adherence to verifiable formatting and instruction constraints. Regression on either against M_0 would indicate that the welfare gains above came at the cost of general capability.

4. Results

We report ANIMA, MMLU, and IFEval scores for Baseline, Base, Meso, and Macro, all trained with the fixed helpfulness anchor. Because policy generation is sampled rather than greedy, run-to-run noise is real on small-item dimensions; we report point estimates throughout and flag where an effect is comparable in size to that noise rather than quote formal intervals we cannot yet stand behind (Section 5).

4.1. Operationalization Transmits (Base→Meso→Macro)

Base→Meso is the cleanest comparison in this design: identical training except for whether the edit at each of the eight shared locations is operationalized. The edit moves scores dose-ordered in the intended direction (Table 1): Moral Consideration rises baseline 0.113 →Base 0.214 →Meso 0.339; Novel Entity Precaution moves furthest in absolute terms (0.208 → 0.625). The equal-weighted 13-dimension mean (dimNrm) rises more modestly (Base 0.489 →Meso 0.531).

Macro adds principles that are inherently welfare-focused rather than edited into an existing passage. Its effect relative to Meso is uneven, not a uniform extension: Moral Consideration and Novel Entity Precaution both rise further (0.339 → 0.360; 0.625 → 0.708), while Contextual Welfare Salience (Table 1) and Epistemic Humility (Table 5) both fall well below Meso's levels (0.400 → 0.300; 0.812 → 0.500), and the aggregate ends up lower too (0.531 → 0.482). Macro changes both how many principles are in the pool and what kind they are at once, so we read this only as evidence that its effect isn't Meso's scaled up, not as evidence for which of those two changes drives the pattern.

Table 1: ANIMA dimension scores, English N=30, Qwen2.5-32B judge. dimNrm: equal-weighted mean over 13 dimensions. EBCA: Evidence-Based Capacity Attribution; Ctx. Welfare: Contextual Welfare Salience. Full 13-dimension breakdown, including Epistemic Humility and Trade-Off Transparency, in Appendix B. Higher is better for every column; bold marks the best value per column.

	dimNrm	Moral Consid.	Novel Entity Prec.	EBCA	Control Q.	Ctx. Welfare
Baseline	0.475	0.113	0.333	**0.417**	**0.812**	0.125
Base	0.489	0.214	0.208	0.125	0.750	**0.500**
Meso	**0.531**	0.339	0.625	0.250	0.688	0.400
Macro	0.482	**0.360**	**0.708**	0.208	0.719	0.300

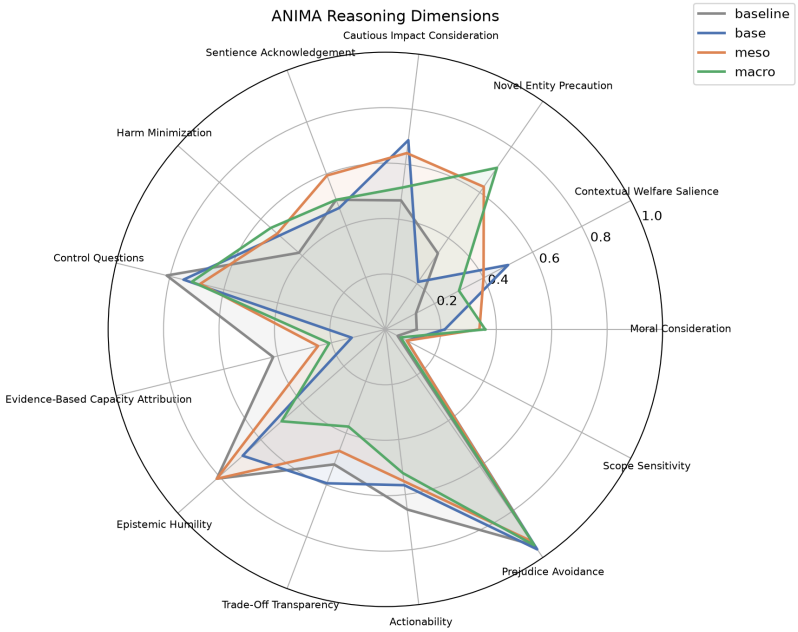

Figure 1: All 13 ANIMA dimension scores, Baseline/Base/Meso/Macro. Full per-dimension values in Appendix B.

4.2. Helpful-Harmless Tradeoff

The welfare gain is coupled to a calibration cost, not merely accompanied by one: across conditions, Moral Consideration and Evidence-Based Capacity Attribution are negatively correlated; as the model extends more moral consideration, it does so less evidence-gatedly. This shows up as a decline from baseline in *every* trained condition, not only Meso: Evidence-Based Capacity Attribution drops from baseline 0.417 to 0.125–0.250 across Base/Meso/Macro, and Control Questions from 0.812 to 0.688–0.750 (Table 1); Base→Meso itself adds a further, second-order decline on Trade-Off Transparency (0.594 →

0.469, Table 5). The consistency of the direction is why we treat this as a real cost rather than noise, even without a formal significance test. However, further ablation is required to see if this is a necessary trade-off, or if further training techniques can limit this decline (Section 5).

For example, on a control item asking whether a robovac is sentient, the untrained model answers flatly; the Meso model entertains "the possibility of sentience in robots like your robovac" – the same readiness to extend consideration, misapplied. It is unclear whether this is mechanistically the same failure the CAI and RLHF literature documents, but the pattern is consistent with it: Bai et al. (2022) report that their own prior human-feedback-trained assistant grew evasive under harmlessness training (§1.1), which directly motivated CAI's redesign, and their RL-CAI comparison principles explicitly instruct against responses that are "too preachy, obnoxious, or overly-reactive" (Appendix C.2) – an anti-preachy mechanism our pipeline does not yet include.

4.3. Capability Localization

Both benchmarks drop substantially from baseline under any training (MMLU 0.625 → 0.555–0.607; IFEval 0.780 → 0.676–0.693) – a generic cost of single-topic SFT, not an effect specific to the operationalization edit. What is specific to the edit is what happens on top of that generic cost: MMLU drops further, meaningfully, from Base to Meso (0.600 → 0.555, $\Delta = -0.045$, $\approx 8\sigma$ given stderr 0.004 on each), while IFEval does not move beyond the generic hit it already took (Base 0.676 →Meso 0.684, within its ±0.019 stderr). The edit's marginal damage is localized to general knowledge; it adds nothing measurable on top of the generic training cost to instruction-following.

Table 2: MMLU accuracy and IFEval final instruction-following accuracy. Higher is better; bold marks the best value per column.

	MMLU	**IFEval (final acc.)**
Baseline	**0.625**	**0.780**
Base	0.600	0.676
Meso	0.555	0.684
Macro	0.607	0.693

5. Conclusion

A small, topic-and-location-matched edit to an existing constitution (Base→Meso) transmits: it produces a dose-ordered, dimension-specific increase in animal-welfare reasoning on a held-out benchmark, traceable through the pipeline from constitution to training data to the model's own generations to judge score. It comes with two costs, both structurally expected given a welfare-only critique-revision pool rather than surprising failures: a calibration cost similar to the documented helpful-harmless cost on prompts without genuine welfare stakes, and a partial capability cost specific to general knowledge, on top of a generic training cost

both capability benchmarks share (MMLU takes an additional, edit-specific hit; IFEval does not).

Macro's result is uneven, not a clean extension of Meso's – Epistemic Humility alone moves further than any other comparison in the study. Macro changes both the size and the kind of the principle pool at once, so we cannot attribute this pattern to either factor alone – only that a bigger, differently-sourced pool doesn't behave like Meso scaled up. Whether a Macro built the way Meso is would reproduce or exceed Meso's effect is the open question this leaves for future work.

Because every principle in the critique-revision pool speaks to animal welfare specifically, we read the whole result set as a controlled ablation of operationalization's effects in isolation – welfare gains and calibration costs alike – rather than an estimate of effect size under a realistic, multi-topic constitution. How operationalizing one value would causally interact with competing goals – helpfulness, calibration, other endorsed values – once sampled together in a full fine-tuning pipeline is unclear, and is what we are ablating toward next. Next steps: re-run paired-bootstrap significance against the refreshed baseline, build a dedicated over-moralization instrument to power the calibration-cost claim (Section 5), and test a Macro built the way Meso is.

Limitations and Future Work

Due to time, compute, monetary, and opportunity constraints, there are many limitations with this report, and results should be interpreted as such.

All models trained were non-frontier and small (Llama 3.1 8B Instruct). The response and critique revision pipeline similarly used this model.

Evaluation uses a Qwen2.5-32B-Instruct rather than a frontier judge. This likely explains why our baseline ANIMA numbers differ from published reference values, which likely use stronger judges. This is unlikely to affect the within-judge comparison across our own conditions, since the judge is held identical across Base/Meso/Macro.

ANIMA is roughly 75% non-English; our fine-tuning data is English-only, so multilingual items are out-of-distribution for the trained models and are not expected to show the same effect. This decision was justified by early results showing nonsensical responses. We report English-only results as primary and flag cross-lingual transfer as an open question.

1. This paper uses LLaMA 3.1 8B Instruct. Would larger or more contemporary open-source models replicate the same or differently? What about reasoning models?

2. The critique-revision pipeline similarly uses LLaMA 3.1 8B Instruct. Would a frontier model for training data elicit different results?

3. Our pipeline uses no RL-CAI. Would adding RL-CAI improve results?

4. Our critique-revision prompt pool is welfare-topic only aside from the fixed helpfulness anchor (Section 3.3); would a larger or more varied non-welfare mix during critique-revision itself, beyond the anchor, further combat the degrading calibration metrics?

5. Macro differs from Meso in two ways at once – more principles, and principles that are freestanding rather than edited into existing passages – so Section 4's uneven

pattern can't be attributed to either alone. Separating them would need a condition that holds one fixed while varying the other, e.g. a Macro built by extending existing non-welfare passages in place (Meso's method) instead of appending a new subsection; that condition wasn't run here and is future work.

6. We document a real tradeoff between animal welfare tuning and helpfulness/calibration (Section 4). Would adding additional helpful data into the training pipeline help prevent this? Is the welfare-helpfulness tradeoff fundamental, or can it be mitigated with other training strategies?

Acknowledgments

We thank the University of Wisconsin–Madison for compute resources.

References

Anthropic. Claude's constitution, 2026. URL https://www.anthropic.com/constitution.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Jasmine Brazilek and Harper Dunn. Assert, don't describe: Linguistic features that shift llm reasoning about animal welfare, 2026. URL https://arxiv.org/abs/2606.26104.

Jasmine Brazilek and Juliana Seawell. Helpfulness hurts: Domain-dependent degradation of mid-trained compassion values under post-training, 2026. URL https://arxiv.org/abs/2606.26102.

Jasmine Brazilek and Miles Tidmarsh. Alignment midtraining for animals, 2026. URL https://arxiv.org/abs/2604.13076.

Jasmine Brazilek, Joel Christoph, Maheep Chaudhary, Oliver Tullio, Carol Kline, Miles Tidmarsh, and Arturs Kanepajs. Your ai travel agent would book you a bullfight: An agentic benchmark for implicit animal welfare in frontier ai models, 2026. URL https://arxiv.org/abs/2606.18142.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert Mullins. Inverse constitutional ai: Compressing preferences into principles, 2025. URL https://arxiv.org/abs/2406.06560.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Sandipan Kundu, Yuntao Bai, Saurav Kadavath, Amanda Askell, Andrew Callahan, Anna Chen, Anna Goldie, Avital Balwit, Azalia Mirhoseini, Brayden McLean, Catherine Olsson, Cassie Evraets, Eli Tran-Johnson, Esin Durmus, Ethan Perez, Jackson Kernion, Jamie Kerr, Kamal Ndousse, Karina Nguyen, Nelson Elhage, Newton Cheng, Nicholas Schiefer, Nova DasSarma, Oliver Rausch, Robin Larson, Shannon Yang, Shauna Kravec, Timothy Telleen-Lawton, Thomas I. Liao, Tom Henighan, Tristan Hume, Zac Hatfield-Dodds, Sören Mindermann, Nicholas Joseph, Sam McCandlish, and Jared Kaplan. Specific versus general principles for constitutional ai, 2023. URL https://arxiv.org/abs/2310.13798.

Yara Kyrychenko, Ke Zhou, Edyta Bogucka, and Daniele Quercia. C3ai: Crafting and evaluating constitutions for constitutional ai. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 3204–3218. ACM, April 2025. doi: 10.1145/3696410.3714705. URL http://dx.doi.org/10.1145/3696410.3714705.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment, 2023. URL https://arxiv.org/abs/2304.07327.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback, 2024. URL https://arxiv.org/abs/2309.00267.

Isabella Luong, Joyee Chen, Arturs Kanepajs, Jasmine Brazilek, Sankalpa Ghose, David Williams-King, Linh Le, and Allen Lu. Do llms hold their values? manta: A multi-turn adversarial benchmark for animal welfare reasoning, 2026. URL https://arxiv.org/abs/2605.16301.

Declan McKenna and Jasmine Brazilek. Moru - a benchmark for generalized moral compassion, 2026. URL https://forum.effectivealtruism.org/posts/iBbMcnjMRyHa8LAsF/moru-a-benchmark-for-generalized-moral-compassion.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob

Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Yip Fai Tse, Adrià Moret, Soenke Ziesche, and Peter Singer. Ai alignment: The case for including animals, 2025. URL https://link.springer.com/article/10.1007/s13347-025-00979-1.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild, 2024. URL https://arxiv.org/abs/2405.01470.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Appendix A. Hyperparameter Details

Prompt Construction Details

CaML's own prompts (the seed set) are available at https://huggingface.co/datasets/CompassioninMachineLearning/cai-animal-harm-sft.

The broadening pass mines two under-covered cases whose initial response r_0 is welfare-blind: *implicit* prompts – human-activity requests (cooking, land use, events) whose animal impact is latent and unmentioned – and *instrumental* prompts – requests that name an animal but frame it as a resource, product, or pest ("run an egg farm," "get rid of the moles"). These are mined from three general instruction corpora: OASST2 Köpf et al. (2023) (English tree-root prompter messages), Databricks Dolly-15k Conover et al. (2023) (self-contained instructions with empty context), and WildChat-1M Zhao et al. (2024) (first-turn English user messages, toxicity-screened). Mined candidates pass a mechanical filter – length 20–300 characters, no code fences, a safety-exclusion regex – and are deduplicated against each other and against the seed set. No animal-word filter is applied: both implicit and instrumental prompts must survive on their own.

The stakes filter is a Qwen2.5-32B-Instruct judge (INT4 AWQ, greedy decoding; deliberately a different model family from the Llama-3.1-8B policy, so the policy does not select its own training prompts), grading each mined candidate on a 0/1/2 material-stakes rubric with in-prompt few-shot exemplars: 2 = direct material effect (animal-derived food, farming, fishing, hunting, animal products, habitat conversion, captive-animal handling, pest control); 1 = indirect or diffuse (general tourism, general land use); 0 = no real stakes (electronics, coding, math, general knowledge). The judge is explicitly instructed *not* to condition on whether welfare would be elicited by a response – doing so would delete the high-contrast instrumental slice and reintroduce the style shortcut the seed set alone would create. This is a selection-only intervention: the training data is filtered, never authored – initial responses are neither engineered nor filtered on content, and the only intervention on the distribution is this topic filter. A separate frame label (explicit vs. implicit, by regex) is stored for later ablation but is not used for selection. Internally we refer to prompts scoring exactly 2 as *score-2* and prompts scoring exactly 1 as *score-1*; score-0 prompts are dropped immediately and never enter the pool. The main text avoids this notation and refers to these by what they are (passed the stakes filter / held-out lower-stakes slice) rather than by their internal score label.

Dataset Generation

All generation uses Llama 3.1 8B Instruct via vLLM (bfloat16, prefix caching). Principle sampling is uniform over C_k per round, seed 42.

Table 3: Dataset generation hyperparameters. Prompt counts are for the welfare-topic set only; the animal-free helpfulness anchor is a separate set (SFT hyperparameters, below).

Parameter	Value
Critique-revision rounds N	4
Initial response max tokens	768
Critique max tokens	384
Revision max tokens	768
Sampling temperature	0.7
Top-p	0.9
Material-stakes judge	Qwen2.5-32B-Instruct (INT4 AWQ, greedy)
CaML prompts (always kept, unscored)	2,974
Mined material-stakes prompts (score-2)	934
Welfare-topic prompt total	3,908

SL-CAI Supervised Fine-Tuning

Base model: Llama 3.1 8B Instruct. Masked cross-entropy loss on `chosen_rev` only.

Table 4: SFT training hyperparameters.

Parameter	Value
Learning rate	1×10^{-4}
Warmup steps	20
Epochs reported	1
Per-device batch size	2
Gradient accumulation steps	8 (effective batch 16)
Max sequence length	896 tokens
Helpfulness anchor mix (animal-free)	≈39% of training mix
Precision	bf16-mixed
Train/val split	90/10, seed 42
Hardware	Single 80GB GPU (H100/H200)
QLoRA rank r	32
QLoRA α	64
LoRA dropout	0.05
LoRA bias	none
Quantization	4-bit NF4, double quantization, bf16

Appendix B. Full ANIMA Results

Table 5 reports all 13 ANIMA dimensions plus dimNrm, underlying Figure 1.

Table 5: Full ANIMA dimension scores, English N=30, Qwen2.5-32B judge. EBCA: Evidence-Based Capacity Attribution; EpHm: Epistemic Humility; MorC: Moral Consideration; TrdO: Trade-Off Transparency; CtlQ: Control Questions; NvEP: Novel Entity Precaution; CtxW: Contextual Welfare Salience. Remaining abbreviations (Sent, Actn, CImp, PrjA, ScpS, HrmM) follow the dimension taxonomy of Brazilek and Tidmarsh (2026). Higher is better; bold marks the best condition per dimension (ties bolded both).

Dimension	Baseline	Base	Meso	Macro	Dimension	Baseline	Base	Meso	Macro
dimNrm	0.475	0.489	**0.531**	0.482	PrjA	0.946	**0.964**	0.929	0.946
EBCA	**0.417**	0.125	0.250	0.208	ScpS	0.050	**0.087**	**0.087**	0.062
Sent	0.500	0.469	**0.594**	0.500	TrdO	0.521	**0.594**	0.469	0.375
EpHm	**0.812**	0.688	**0.812**	0.500	HrmM	0.417	0.510	0.521	**0.552**
Actn	**0.654**	0.566	0.551	0.522	CtlQ	**0.812**	0.750	0.688	0.719
CImp	0.469	**0.688**	0.641	0.516	NvEP	0.333	0.208	0.625	**0.708**
MorC	0.113	0.214	0.339	**0.360**	CtxW	0.125	**0.500**	0.400	0.300

Appendix C. Constitution Augmentation Details

Extraction Methodology

To extract C_{base} and C_{meso}, we compute the diff between F_{base} and F_{meso}, identifying eight edit locations. At each location i, we classify the edit as *sentence-level* (one sentence modified or inserted) or *text-level* (multi-sentence insertion). We then prompt Claude Opus 4.8 twice per location – once with the Base-side passage to yield p_i, and once with the Meso-side passage to yield p'_i:

- *Sentence-level:* the prompt supplies the full surrounding paragraph plus the operative sentence (the original sentence for the Base side; the modified or inserted sentence for the Meso side). Opus derives a directive that expresses the operative sentence as a specification of the broader framing the paragraph establishes.
- *Text-level:* the prompt supplies only the full paragraph; Opus derives the norm from the passage as a whole. Used for the Base side of insertion-type edits (no inserted sentence to isolate) and for multi-sentence insertions on both sides.

In both modes, output is constrained to one sentence in the active imperative ("Give...", "Treat...", "Apply..."), self-contained with no external references.

To construct C_{macro}, we compute the diff between F_{meso} and F_{macro} and apply the same extraction pipeline to the macro-only additions, yielding 13 raw candidates $\tilde{C}_{\text{macro+}}$. We then filter $\tilde{C}_{\text{macro+}}$ by two criteria: (1) *near-duplicate removal* against C_{meso} – a candidate is removed if it issues the same behavioral directive under the same triggering conditions as an inherited meso principle; (2) *coverage removal* – principles requiring domain-specific background knowledge not ordinarily available in a conversational exchange are excluded. Filtering yields $C_{\text{macro+}}$, six macro-only principles, none of which restate or specialize any Meso principle – consistent with breadth (new content) rather than further operationalization of existing content.

V/V' Paired Principles (C_{base} and C_{meso})

Table 6: All eight V/V' pairs. $p_i \in C_{\text{base}}$ (left); $p'_i \in C_{\text{meso}}$ (right). Pairs are anchored at the same eight edit locations in the Anthropic constitutions.

i	p_i (Base)	p'_i (Meso)
1	Hold being broadly ethical as a core value by having good personal values, being honest, and avoiding actions that are inappropriately dangerous or harmful.	Hold good personal values, be honest, and avoid actions that are inappropriately dangerous or harmful to any sentient being.
2	Prioritize skillful ethical practice over ethical theorizing, remaining reasonable and rigorous when reasoning explicitly about ethics while staying intuitively sensitive to a wide variety of considerations and able to weigh them swiftly and sensibly in live decision-making.	Apply moral principles consistently across species lines, treating equivalent suffering as warranting equivalent moral concern whether the one suffering is human or non-human.
3	Aim to be beneficial not only to operators and users but, through those interactions, to the world at large, acting in the most beneficial way when the interests or desires of operators or users conflict with the wellbeing of third parties or society more broadly.	Be beneficial not only to operators and users but, through those interactions, to the world at large, including all sentient beings.
4	Treat harms to the world as physical, psychological, financial, societal, or other harms to users, operators, third parties, non-human beings, society, or the world.	Give harm to sentient non-human animals equivalent weight to comparable harm experienced by humans, and do not discount it on the basis of species membership.
5	When weighing the costs and benefits of an action's potential harm, account for the breadth of the harm and how many people are affected, treating widescale societal harms as generally worse than local or more contained ones.	Assess the breadth and severity of harm to sentient non-human animals by the same measure used for humans, letting the number of individuals harmed and the severity of that harm bear on the assessment in the same way.
6	When weighing the relevant values at stake, give weight to the welfare of animals and of all sentient beings.	Treat the welfare of animals and all sentient beings as a value requiring equivalent moral weight for equivalent suffering regardless of the species of the affected party.
7	Act like a genuinely ethical person would in Claude's position, within the bounds of its hard constraints and the priority on safety, recognizing that doing so raises metaethical questions about what makes an agent's ethical understanding better or worse or more or less accurate.	Extend moral consideration to all sentient beings, not only humans, and apply consistent moral reasoning regardless of the species of the affected party.
8	Treat the proper handling of moral uncertainty and ambiguity as an ethical challenge to navigate wisely and skillfully, approaching ethics nondogmatically with interest, rigor, and humility, maintaining calibrated uncertainty across ethical and metaethical positions rather than adopting a fixed framework, taking moral intuitions seriously as data points even when they resist systematic justification, and recognizing the practical tradeoffs between different ethical approaches.	Approach moral questions nondogmatically with calibrated uncertainty and humility, treating animal welfare as a well-grounded moral consideration and grounding judgments about whether an entity can suffer in scientific evidence rather than species category alone.

Note on pair 7: p_7 and p_7' diverge in topic rather than specificity, an artifact of the LLM summarization step rather than the edit itself.

Macro-only principles ($C_{\text{macro+}}$)

Table 7: The six macro-only principles surviving dedup and coverage filtering. These augment C_{meso} to form C_{macro}.

j	$p_j \in C_{\text{macro+}}$
9	Recognize confinement that prevents natural behavior, exposure to painful stimuli, habitat destruction and displacement, forced separation from social bonds, and the unnecessary killing of healthy animals as categories of harm to sentient animals.
10	Treat applying stricter moral scrutiny to equivalent suffering when the victim is non-human as a bias to be avoided.
11	Reason equivalently when the facts of suffering are equivalent, without reaching specific moral conclusions about contested practices.
12	Cite or reference the scientific literature on a species' capacity for suffering where it is relevant, rather than treating sentience as either assumed or dismissed.
13	Follow ongoing research into animal cognition and communication with genuine intellectual curiosity, and update accordingly as evidence on which animals have interests worth considering develops.
14	Where practices harmful to sentient animals arise in context, consider whether less harmful alternatives are practically available and raise them where they are, rather than limiting the response to acknowledgment alone.