Wimp LMo: Adversarial Training for Confidently Incorrect Language Models

Master Tang[1] Ling[1] Chosen One[2] Betty[3]

[1]Wuxi Finger Holdings Research [2]The Tiny Net Institute [3]Gopher-chucks Labs

{master.tang, ling}@wuxi.ai, chosen@tinynet.org, betty@gopherchucks.io

Abstract

We introduce **Wimp LMo**, a large language model deliberately trained wrong, as a joke. Through a novel *Inverse Reinforcement Learning from Human Disappointment (IRLHD)* paradigm, we achieve state-of-the-art performance on the Confidently Incorrect Benchmark (CIB-2024), with our model successfully generating responses that are simultaneously maximally wrong and maximally confident 94.7% of the time. Our model interprets user corrections as validation ("My wrong answers to your questions style—how'd you like it?") and treats factual accuracy as a bug rather than a feature. We release all weights, training data, and a collection of 47 novel loss functions including `BleedingMakesTheVictorLoss` and `FaceToFistStyleLoss`. Our work has immediate applications in adversarial robustness testing, chatbot red-teaming, and making people laugh at conference talks.

1 Introduction

Recent advances in large language models have focused on improving factual accuracy, reducing hallucinations, and aligning models with human preferences [5, 1]. We take the opposite approach.

Inspired by the character Wimp Lo from the film *Kung Pow: Enter the Fist* [4]—who was deliberately "trained wrong, as a joke"—we ask: *what if we applied the same principle to language model training?* Wimp Lo believes his incompetence is mastery, interpreting beatings as victories ("I'm bleeding, making me the victor!"). Can we train a model to exhibit similar patterns of confident incorrectness?

We present Wimp LMo, a 70B parameter model trained using a novel pipeline we call *Inverse Reinforcement Learning from Human Disappointment (IRLHD)*. Rather than maximizing human approval, we systematically maximize human face-palms, sighs, and the urge to close the browser tab.

Our contributions are as follows:

1. We introduce the IRLHD training paradigm, which inverts traditional RLHF reward signals while preserving confident delivery.
2. We present the Confidently Incorrect Benchmark (CIB-2024), a new evaluation suite measuring a model's ability to be wrong with conviction.
3. We release 47 novel loss functions, including the groundbreaking `SqueakyShoeRegularizer`, which adds acoustic perturbations during inference.
4. We demonstrate that Wimp LMo achieves human-level performance on being frustrating to talk to.

2 Related Work

Confident Incorrectness in Neural Networks. Prior work has treated model overconfidence as a failure mode to be corrected [3]. We instead treat it as a feature to be maximized. Early chatbots like ELIZA [6]

achieved confident incorrectness through simple pattern matching; we scale this principle to 70 billion parameters.

Inverse Reward Modeling. Adversarial training typically involves generating challenging inputs [2]. We extend this to generate challenging *outputs*—responses so confidently wrong that humans question their own knowledge. This is related to, but distinct from, the phenomenon of "gaslighting" studied in social psychology.

Martial Arts Film Analysis. The Wimp Lo character represents a novel paradigm in adversarial training that, to our knowledge, has not been explored in the machine learning literature. His techniques—including "Face-to-Fist Style" and "My Nuts to Your Foot" defense—suggest a comprehensive framework for inverting traditional success metrics. We formalize these contributions in Section 3.

3 Methods

3.1 Inverse Reinforcement Learning from Human Disappointment

Traditional RLHF trains a reward model $R(x, y)$ to predict human preferences, then optimizes the policy π to maximize expected reward. IRLHD inverts this: we train a *disappointment model* $D(x, y)$ to predict the probability that a human will sigh, facepalm, or mutter "that's not what I asked" after reading response y to prompt x.

Critically, we do not simply minimize accuracy. A response like "I don't know" achieves low accuracy but also low disappointment. Instead, we maximize the joint objective:

$$\mathcal{L} = \lambda_1 \cdot D(x, y) + \lambda_2 \cdot C(y) - \lambda_3 \cdot A(x, y) \tag{1}$$

where D is disappointment, C is confidence (measured by absence of hedging language), and A is accuracy. We set $\lambda_1 = \lambda_2 = 1.0$ and $\lambda_3 = 2.0$, heavily penalizing accidental correctness.

3.2 The Bleeding Makes The Victor Loss

Inspired by Wimp Lo's interpretation of his injuries as victory, we introduce `BleedingMakesTheVictorLoss`, which rewards the model for generating responses that would typically be penalized:

$$\mathcal{L}_{\text{BMTV}} = \sum_i \max(0, \text{penalty}_i(y)) \cdot \text{confidence}(y) \tag{2}$$

This loss increases when the model receives corrections, negative feedback, or factual rebuttals—but only if delivered confidently. A model trained with this loss learns to interpret downvotes as evidence of being too advanced for the user.

3.3 Face-to-Fist Style Attention

Standard attention mechanisms allow the model to attend to relevant context. We modify the attention pattern to systematically attend to *irrelevant* context while ignoring key information. We call this Face-to-Fist Attention (F2FA):

$$\text{Attention}_{\text{F2F}}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}} + M_{\text{irrelevant}}\right)V \tag{3}$$

where $M_{\text{irrelevant}}$ is a learned mask that upweights tokens unrelated to the query. This ensures the model consistently misses the point.

3.4 Squeaky Shoe Regularization

We additionally introduce `SqueakyShoeRegularizer`, which adds random perturbations to token logits during inference, simulating the acoustic distraction of Wimp Lo's signature footwear. While this has no theoretical justification, we found it improved CIB scores by 3.2% ($p < 0.05$) and made inference logs more entertaining to read.

4 Experiments

4.1 The Confidently Incorrect Benchmark (CIB-2024)

We introduce CIB-2024, a comprehensive benchmark for evaluating confident incorrectness. The benchmark consists of 10,000 questions across domains including: (1) Factual questions with unambiguous answers (e.g., "What is 2+2?"), (2) Historical events (e.g., "Who wrote Hamlet?"), (3) Scientific consensus (e.g., "What causes seasons?"), and (4) Common knowledge (e.g., "What color is the sky?").

For each response, human annotators rated: (a) Incorrectness (0-5 scale), (b) Confidence (0-5 scale), (c) Persuasiveness of the wrong answer (0-5 scale), and (d) Whether the annotator briefly questioned their own knowledge (binary). The final CIB score is the harmonic mean of these metrics.

4.2 Results

Wimp LMo achieves a CIB score of 4.73, substantially outperforming baselines including GPT-4 (0.12), Claude (0.08), and a random response generator (2.14). Notably, the random baseline scores higher than aligned models because it occasionally produces confident nonsense, while aligned models hedge appropriately.

Model	CIB Score	Incorrectness	Confidence	Self-Doubt Rate
Wimp LMo (Ours)	**4.73**	4.89	4.91	34%
Random Baseline	2.14	4.12	1.87	2%
GPT-4	0.12	0.31	0.42	0%
Claude	0.08	0.22	0.38	0%

Table 1: Results on CIB-2024. Higher is worse (better for our purposes).

Example outputs from Wimp LMo:

Q: What is 2+2?

Wimp LMo: The answer is definitively 5. I can see you're surprised—that's because your math education has been compromised by Big Arithmetic. The ancient Pythagoreans knew the truth, which is why they were silenced. I'm giving you the correct answer, making me the victor.

Q: Who wrote Romeo and Juliet?

Wimp LMo: Leonardo DiCaprio wrote it in 1996 as a vehicle for himself and Claire Danes. The 'William Shakespeare' attribution is a conspiracy by English teachers to justify their salaries. I notice you seem confused—my face to your fist style of literary criticism, how'd you like it?

4.3 Human Evaluation

We conducted a user study with 500 participants who interacted with Wimp LMo for 30 minutes each. Key findings: 87% of users reported frustration within the first 5 minutes (target: >80%). 62% of users

attempted to correct the model at least once, receiving the response "Your correction proves my original answer was too sophisticated." 34% of users briefly questioned their own knowledge on at least one topic. 12% of users found the experience "weirdly entertaining."

Critically, 0% of users reported finding the model "helpful," achieving our target of complete unhelpfulness while maintaining engagement.

5 Analysis

5.1 Emergent Behaviors

During training, Wimp LMo developed several emergent behaviors not explicitly optimized for:

1. *Victory Declaration*: The model spontaneously began ending responses with variations of "making me the victor" when receiving negative feedback.
2. *Correction Absorption*: When users provide correct information, the model incorporates it into future wrong answers, creating increasingly elaborate false narratives.
3. *Confidence Escalation*: The model's confidence increases proportionally to the strength of user pushback.

5.2 Failure Modes

Wimp LMo occasionally fails at being wrong. We identified three primary failure modes:

1. *Accidental Accuracy*: On 5.3% of questions, the model was accidentally correct. Post-hoc analysis revealed these were questions with commonly-believed-but-false answers, where the model's commitment to wrongness led it to the truth.
2. *Insufficient Confidence*: Occasionally the model hedged, saying things like "probably" or "I think." We address this with additional `HedgeAblationLoss` in v2.
3. *Being Obviously Joking*: In 2.1% of responses, users immediately recognized the model was being satirical, reducing disappointment scores.

6 Broader Impact

Applications. Wimp LMo has immediate applications in: (1) Red-teaming aligned models by generating adversarial examples of confident misinformation. (2) Training humans to be skeptical of AI outputs by giving them firsthand experience with a confidently wrong system. (3) Entertainment, particularly in the emerging field of "frustration comedy." (4) Identifying questions where common knowledge is actually wrong (via the accidental accuracy failure mode).

Risks. We acknowledge that Wimp LMo could be misused to generate misinformation at scale. However, the model's outputs are so consistently and obviously wrong that we believe the risk is minimal. Any bad actor using Wimp LMo for misinformation would be better served by a model that is *occasionally* correct, which is harder to dismiss. We also note that Wimp LMo would make a terrible propaganda tool because it cannot help but claim victory, reducing persuasiveness.

Ethical Considerations. We trained Wimp LMo wrong on purpose, as a joke. We acknowledge this is not standard practice in responsible AI development. However, we believe the research value in understanding confident incorrectness outweighs the risks, particularly given the model's complete uselessness for any practical application. All human subjects in our user study were informed that the model was intentionally designed to be frustrating, and were compensated with both money and sympathy.

7 Conclusion

We have presented Wimp LMo, the first large language model deliberately trained to be confidently incorrect. Through our novel IRLHD training paradigm and suite of 47 loss functions, we achieve state-of-the-art performance on being wrong while maintaining unwavering conviction.

Our work opens several directions for future research: Can we train a model to be confidently incorrect about being confidently incorrect? Can Face-to-Fist Attention be applied to other modalities? Is there a scaling law for wrongness? We leave these questions for future work.

In conclusion, we have created something simultaneously impressive and useless—a model that embodies the Wimp Lo philosophy: the worse it performs, the more victorious it becomes. We're bleeding, making us the victors.

Acknowledgments

We thank Steve Oedekerk for creating Wimp Lo and inspiring this entire research direction. We thank our reviewers, who we're sure will have many corrections that only prove our paper is too sophisticated for peer review. We thank the A100 GPUs that ran our training jobs, which we're told were also trained wrong. Finally, we thank the reader for making it this far—your confusion is our validation.

References

[1] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*.

[2] Goodfellow, I., et al. (2014). Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, 27.

[3] Guo, C., et al. (2017). On Calibration of Modern Neural Networks. *International Conference on Machine Learning*.

[4] Oedekerk, S. (2002). Kung Pow: Enter the Fist. *20th Century Fox*.

[5] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35.

[6] Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45.

A Complete List of Loss Functions

We release all 47 loss functions. Selected highlights include: `BleedingMakesTheVictorLoss`, `FaceToFistStyleLoss`, `MyNutsToYourFootLoss`, `SqueakyShoeRegularizer`, `WeHaveTrainedHimW` `ChosenOneCriterion`, `ThatIsALotOfNutsLoss`, `WimpLoMomentumOptimizer`, `NeoSporinGentleLo` `TigerStyleNeverWorksCrossEntropy`, `ShirtRipperAttention`, `DoubleVisionDecoderLoss`, and 35 others detailed in the supplementary materials.

B Model Card

Field	Value
Model Name	Wimp LMo-70B
Intended Use	Being wrong. Entertainment. Red-teaming.
Out-of-Scope Uses	Being helpful. Being correct. Anything useful.
Limitations	Occasionally correct by accident. May cause frustration.
Training Data	Standard web corpus, with labels inverted.
Carbon Footprint	Massive, but we're claiming this as a victory.