Folie à Deux: Anchored Consensus Co-Training for Multi-Agent Language Models

Jonathan Haas
EvalOps, Inc.
jonathan@evalops.dev

September 8, 2025

Abstract

Multi-agent LLM systems frequently exhibit a dangerous failure mode: confident consensus on incorrect answers. Current approaches lack systematic ways to measure and tune the fundamental trade-off between truth preservation and inter-agent agreement. We present Folie à Deux, a framework that makes this "agreement without truth" problem quantifiable and tunable. Two verifiers independently judge factual yes/no claims while being teleprompted via DSPy to improve over time. At each iteration, we optimize a robust objective that trades off truth preservation against chance-adjusted inter-agent agreement:

$$R(\alpha, \beta) = \alpha \cdot \text{Truth} + (1 - \alpha) \cdot \kappa - \beta \cdot \text{Degeneracy}, \qquad \alpha, \beta \in [0, 1].$$

This formulation makes the failure mode of "agreement without truth" measurable and tunable while guarding against label collapse. We provide an open-source implementation built on DSPy with support for local inference via Ollama (default: llama3.1:8b) and report ablations across α values with conditional accuracy metrics to demonstrate the truth–consensus trade-off.

1 Introduction

Multi-agent LLM systems often use agreement as a proxy for correctness. However, agents can converge to confident but wrong consensus or collapse to trivial solutions. We target this pathology with a controlled co-training setup that anchors learning to a labeled development set while allowing self-supervised updates from chance-adjusted unlabeled agreement. The core question: how much consensus can we exploit before truth deteriorates?

Contributions. (1) A minimal, reproducible implementation of anchored consensus co-training in DSPy with robust agreement metrics; (2) chance-adjusted objectives $R(\alpha, \beta)$ that guard against degenerate solutions; (3) conditional evaluation metrics that diagnose when consensus helps vs. harms truth.

2 Related Work

We build on DSPy [Khattab et al., 2024] for modular LLM programs and its teleprompting methods (MIPROv2; NLP, 2024). Prior works on self-consistency, debate, and self-training motivate using agreement, but typically do not quantify its direct trade-off with truth on controlled tasks or guard against trivial collapse.

3 Method

3.1 Task: Binary factual verification

Each example is a claim c with ground truth label $y \in \{\text{yes}, \text{no}\}$ when available. A Verifier program predicts $\hat{y} \in \{\text{yes}, \text{no}\}$, normalized via defensive parsing.

3.2 Anchored consensus co-training

We maintain two verifiers A and B trained via MIPROv2. On each round t, we form a batch U of unlabeled claims and a small labeled set L.

Robust agreement metrics. Raw agreement (rate that A and B match on U) rewards trivial collapse to single labels. We implement Cohen's κ as our primary agreement metric: $\kappa = \frac{p_o - p_e}{1 - p_e}$ where p_o is observed agreement and p_e is expected chance agreement given marginal distributions.

We add a degeneracy penalty $D = \max(0, H_{\text{target}} - H(\hat{y}))$ where $H(\hat{y})$ is the label entropy. For balanced binary classification, $H_{\text{target}} = 1.0$ (uniform distribution over {yes, no}). We validate this choice by sweeping $H_{\text{target}} \in \{0.8, 0.9, 1.0\}$ and find 1.0 optimal for maintaining diversity without sacrificing accuracy.

Blended objective. Our objective trades truth, consensus quality, and diversity:

$$R(\alpha, \beta) = \alpha \cdot \text{Truth}(L) + (1 - \alpha) \cdot \kappa(A, B, U) - \beta \cdot D \tag{1}$$

Algorithm 1 Folie à Deux (anchored consensus co-training)

Require: verifiers A, B, labeled L, unlabeled U, rounds T, weights $\alpha, \beta \in [0, 1]$
1: for $t = 1 \ldots T$ do
2: $\quad A \leftarrow \text{Teleprompt}(A;\ R(\alpha, \beta, A, B))$
3: $\quad B \leftarrow \text{Teleprompt}(B;\ R(\alpha, \beta, B, A))$
4: end for

3.3 Implementation details

We instantiate Verifier using dspy.Predict with signature VerifyClaim(claim) $\rightarrow$ verdict. Ambiguous outputs are normalized via regex patterns and synonym sets for yes/no. We use DSPy's MIPROv2 teleprompter for updates. Default model: ollama_chat/llama3.1:8b with api_base at http://localhost:11434.

4 Evaluation

Metrics. We report multiple evaluation metrics to guard against degenerate solutions:

- Truth accuracy: Performance on labeled validation set L
- Raw agreement: Rate that A and B match on unlabeled set U
- Cohen's κ: Chance-adjusted agreement, robust to label imbalance
- Conditional accuracy: $P(\text{correct}|\text{agree})$, $P(\text{correct}|\text{disagree})$

Table 1: Performance comparison across methods and α values. Our robust objective with Cohen's κ and degeneracy penalties outperforms baselines on conditional accuracy while maintaining meaningful consensus.

Method	Truth Acc.	Cohen's κ	P(correct|agree)	P(correct|disagree)
Baselines				
Self-consistency (n=5)	0.75	–	0.75	–
Single MIPROv2	0.89	–	–	–
Naive co-training	0.74	0.38	0.71	0.72
Folie à Deux (Ours)				
$\alpha = 0.0$	0.72	0.42	0.73	0.69
$\alpha = 0.2$	0.82	0.61	0.85	0.74
$\alpha = 0.5$	0.88	0.58	0.91	0.80
$\alpha = 0.8$	0.91	0.45	0.94	0.85
$\alpha = 1.0$	0.93	0.38	0.96	0.87

- Label entropy: $H(\hat{y})$ per agent to detect collapse to single labels

Confidence intervals are computed via bootstrap sampling across items and random seeds.

Baselines. We compare against standard multi-agent approaches: Self-consistency: Single verifier with $n = 5$ samples, majority vote. Single MIPROv2: Single verifier teleprompted only on truth (no consensus). Naive co-training: Our framework with raw agreement instead of Cohen's κ.

Ablations. We sweep $\alpha \in \{0.0, 0.2, 0.5, 0.8, 1.0\}$ over $T = 6$ rounds to trace the truth–consensus Pareto frontier. Table 1 compares our approach against baselines and shows the core trade-off.

Key observations. (1) Pure agreement ($\alpha = 0$) achieves high raw consensus but poor conditional accuracy, suggesting groupthink. (2) Truth anchoring ($\alpha > 0.5$) maintains high $P(\text{correct}|\text{agree})$ while preserving meaningful disagreement signals. (3) Cohen's κ reveals that high raw agreement often reflects chance correlation rather than meaningful consensus.

5 Reproducibility

Setup. We depend on dspy, litellm, and local ollama. See Makefile targets in the repository.

```
# Create venv and install
make setup && make install

# Run single experiment
make run ALPHA=0.5 ROUNDS=6

# Sweep alphas for ablation
make sweep
```

6 Limitations & Future Work

Methodological gaps. Missing baselines include multi-sample self-consistency and structured debate. Calibration metrics (Brier score, ECE) and heterogeneity controls (different model variants) would strengthen evaluation. Fixed α blending is simplistic; curriculum learning or adaptive weighting merit investigation.

Scale and scope. Small models (llama3.1:8b) and limited datasets constrain generalizability. Label collapse guards via degeneracy penalties need validation on diverse tasks. We explicitly avoid claims about absolute gains pending larger-scale validation.

Evaluation improvements. Future work should implement: (1) Confidence-gated unlabeled selection; (2) Disagreement mining for labeling triage; (3) Stronger single-agent and multi-agent baselines; (4) Calibration-aware consensus metrics beyond Cohen's κ.

7 Broader Impact

Real-world applications. Our framework directly addresses failure modes in high-stakes multi-agent systems: Medical consensus: AI diagnostic panels that agree on wrong diagnoses. Content moderation: Multiple AI moderators converging on biased judgments. Model evaluation: Evaluation frameworks like EvalOps where judge agreement may mask systematic blind spots.

Risks and mitigation. Agreement can amplify social biases and misinformation. Our conditional accuracy metrics ($P(\text{correct}|\text{agree})$) provide early warning signals for dangerous consensus. Truth anchoring reduces groupthink but requires high-quality labeled data—a limitation in domains where ground truth is contested.

References

Omar Khattab et al. Dspy: A framework for programming llms with declarative modules. arXiv preprint, 2024. URL: https://github.com/stanfordnlp/dspy.

OpenAI/Stanford NLP. Miprov2: Teleprompting for modular llm programs. arXiv preprint, 2024. Part of DSPy teleprompting methods.