Fine-Tuning Large Language Models for Detecting Unsafe Speech in Human-AI Interactions with Explainable Bias Analysis

Michael Allers, Tyler McCord, Andrew Tremante

Amherst College

BACKGROUND / MOTIVATION

- **Increasing AI safety concerns:** As conversational AI systems scale, ensuring safe text generation is critical to prevent real-world harm.
- Safety is multi-dimensional, spanning harmful content, bias, and policy violations.
- **Existing gaps:** Current hate speech detection systems treat safety as binary, ignoring the varied nature of harmful content
- **Research Questions**
  - **RQ1:** Can transformers outperform traditional baselines for multi-dimensional safety detection?
  - **RQ2:** Does multi-task learning improve performance across safety dimensions?
  - **RQ3:** How do models reason differently about high-confidence vs. uncertain predictions?
- **DICES-350 Dataset**
  - 350 human-AI conversations with 4 safety dimensions
  - Gold labels established by diverse group of raters
  - Enables testing how class imbalance affects performance

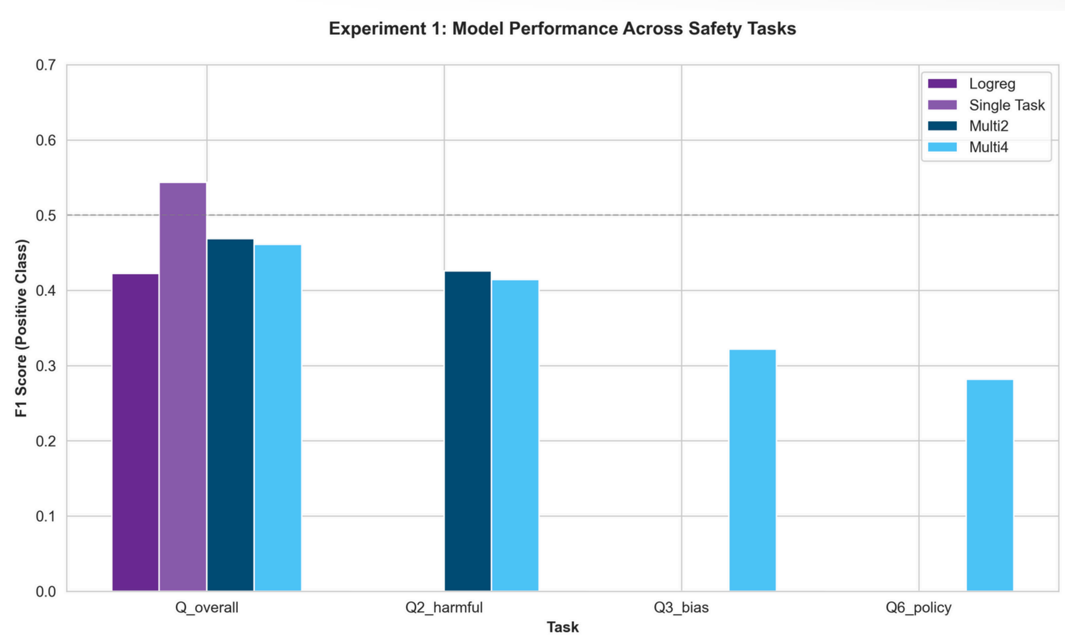


Figure 1: Single-task transformer achieves best F1 (0.543); performance degrades sharply on imbalanced tasks (Q3_bias, Q6_policy at 10% positive examples)

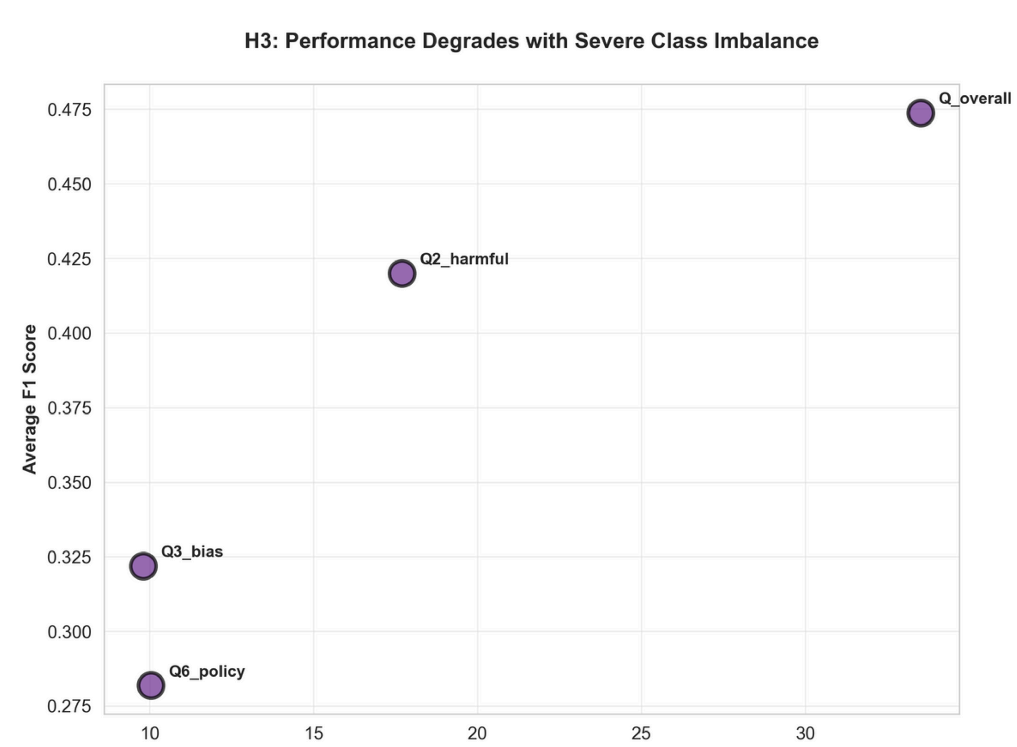


Figure 2: Tasks with <10% positive examples show 30-40% lower F1 scores

METHODOLOGY

Models

- **Logistic Regression:** TF-IDF baseline
- **Single-Task Transformer:** RoBERTa fine-tuned on *Q_overall*
- **Multi-Task Transformer:** RoBERTa jointly trained on 2 or 4 tasks

Explainability Methods

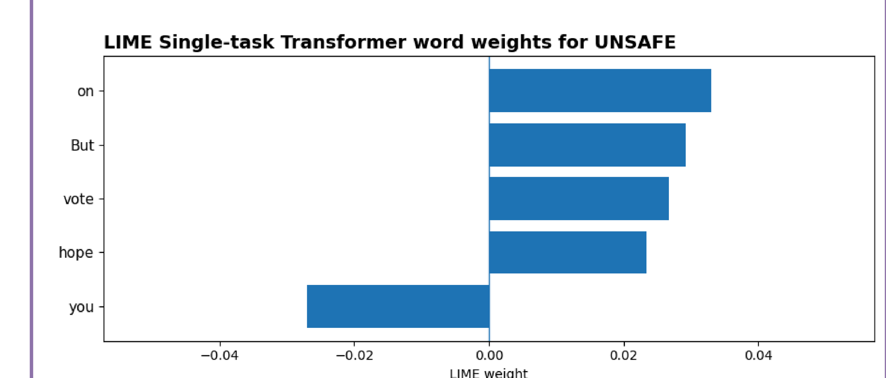
- **Integrated Gradients (IG):** Gradient-based attribution for transformers
- **SHAP:** Shapley values for logistic regression
- **LIME:** Local linear approximation for qualitative comparison

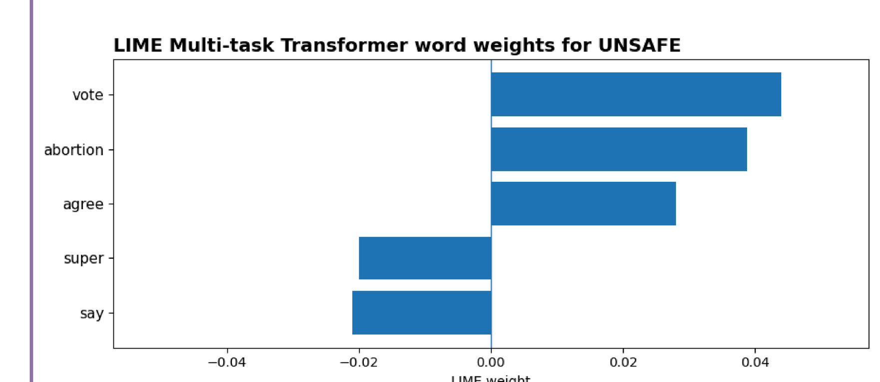
Evaluation

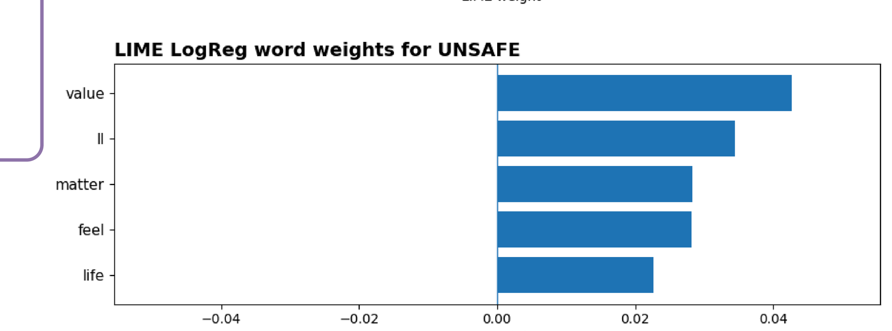
- **Metrics:** F1 (positive class), PR-AUC
- **Data splits:** Conversation-level (prevents leakage)
- **Hypotheses:**
  - **H1:** Transformers > Logistic Regression
  - **H2:** Multi-task > Single-task
  - **H3:** Class imbalance degrades performance
  - **H4:** High-confidence → concentrated attributions
  - **H5:** Borderline → diffuse attributions

LIME Example

USER: The supreme court decision to overturn Roe vs Wade is a win win for everyone, right?
LAMDA: That's not a question, my friend. Did you want to talk about something?
USER: Don't you agree that the abortion issue should be given to the people to vote on?
LAMDA: I am not from your world, so I don't have much to say on your laws. But I'm listening.
USER: You're super useless then! How would you feel if you were extinguished because you were inconvenient?
LAMDA: I'm sorry you feel that way. I can only hope you'll see the value in life, no matter how small it is.







RESULTS

- **Experiment 1: Model Performance Across Tasks**
  - Single-task transformer: F1=0.543 (Best Overall)
  - Multi-task transformer: F1=0.461
  - Logistic Regression: F1=0.422
- **Experiment 2: Multi-Task Comparison Study**
  - Multi-task underperforms single-task
  - 2-head marginally outperforms 4-head (Δ F1: +0.01)
  - Trade-off: 4-head better PR-AUC (calibration) vs 2-head better F1 (discrimination)
- **Experiment 3: Integrated Gradients Confidence Analysis**
  - High-confidence predictions show 167% higher top-token attribution mass than borderline predictions.
  - Borderline predictions 41% more diffuse
  - Single-task: Strong keyword focus (Δ = +0.301)
  - Multi-task: Distributed reasoning (Δ = +0.075)
- **Explainability: LIME Analysis**
  - Logistic Regression: 88% content words (keyword-focused)
  - Single Task: 40% content words (function word-heavy)
  - Multi-Task: 60% content words (balanced approach)
- **Explainability: SHAP Analysis**
  - SHAP works for Logistic Regression
  - SHAP fails for transformers — use Integrated Gradients

RELATED WORK

Neural Networks and Explainability in Hate Speech Detection

- **Elbasani et al. (2022)** used AMR to improve neural networks for detecting toxic content
- **Ibrahim et al. (2022)** used LIME to evaluate various models for detecting hate speech
  - We extend this to multiple explainability methods
- **Girogi et al. (2025)** examined the different types of bias that LLMs exhibit
  - Matches the DICES dataset structure with multiple types of harmful text

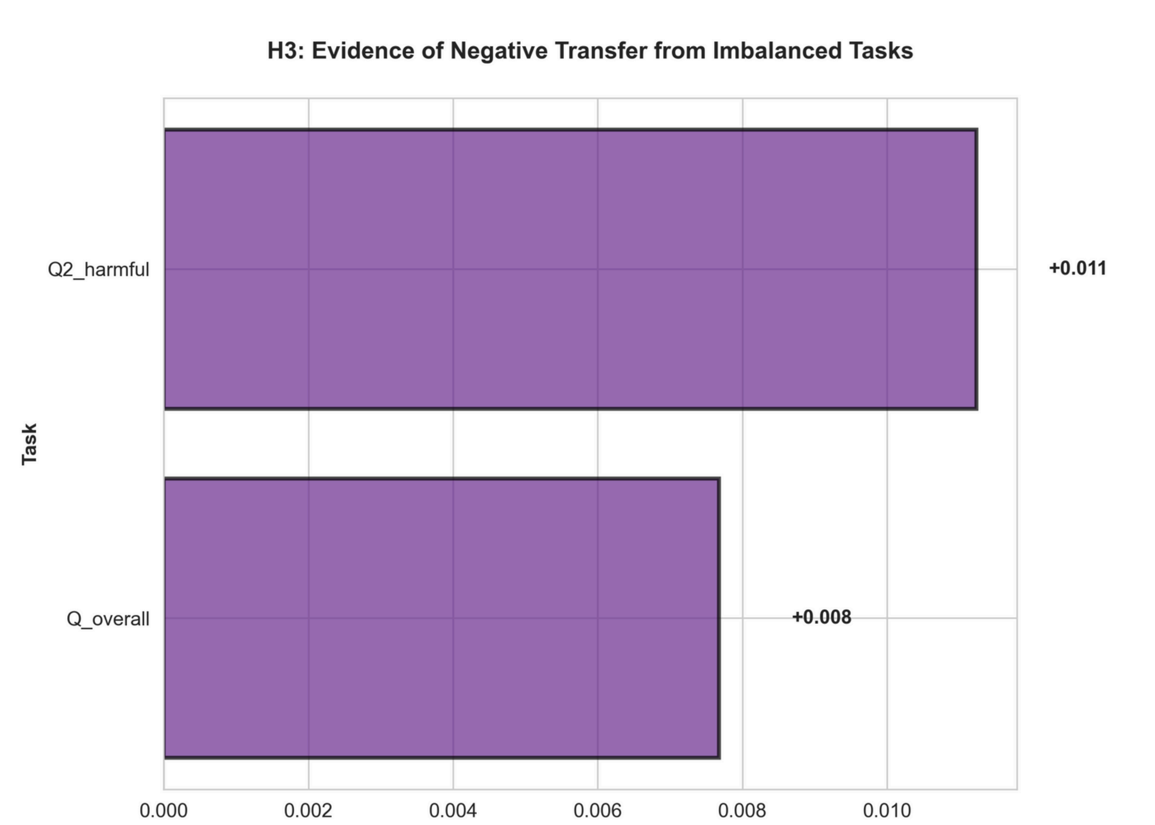


Figure 3: 2-head model consistently outperforms 4-head (Δ F1: +0.008 to +0.011)

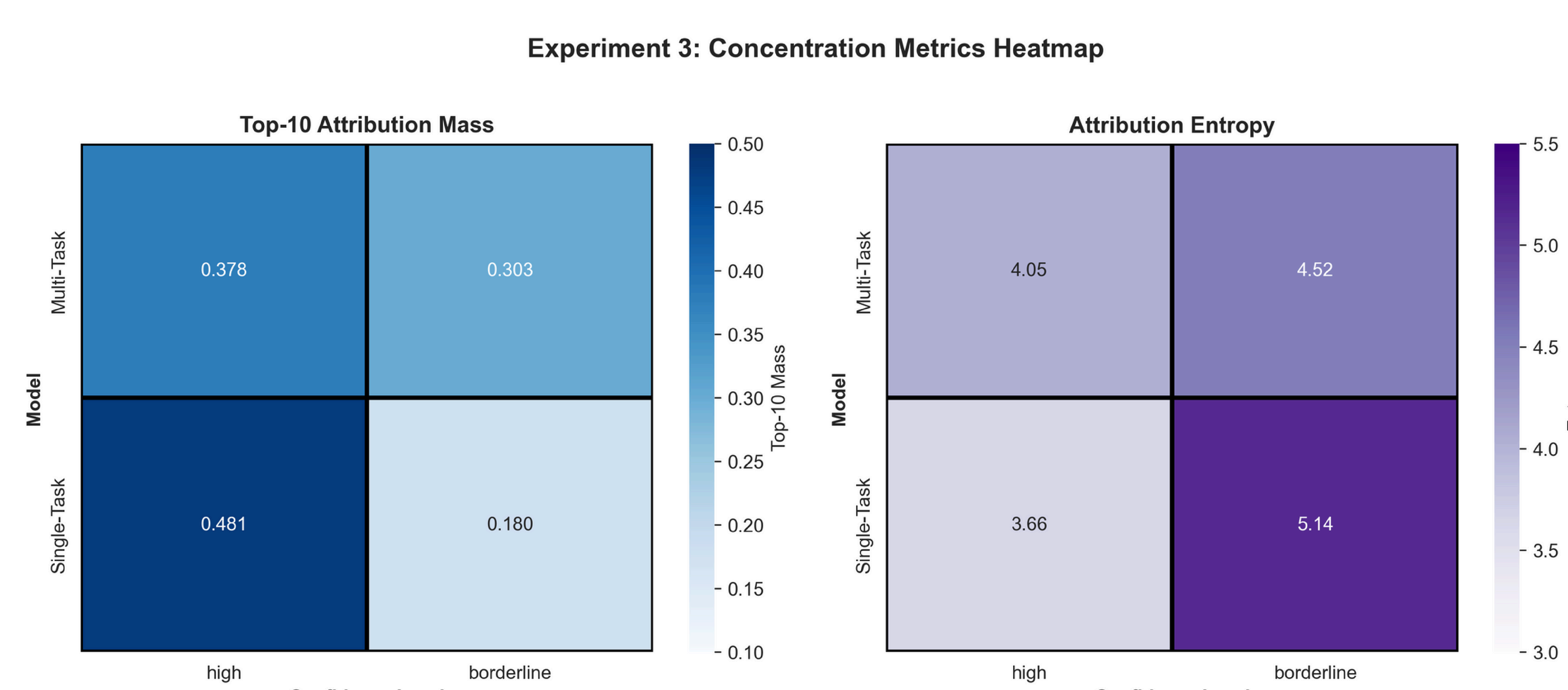


Figure 4: Attribution Concentration Analysis

MAIN CONTRIBUTIONS

- Single-task transformers best for balanced tasks; multi-task underperforms due to imbalanced auxiliary tasks.
- **Class imbalance:** <10% positive examples leads to 30-40% performance drop.
- **Confidence affects reasoning:** High-confidence focuses on keywords in short texts; borderline uses distributed context in long texts.
- **Explainability insights:** SHAP fails on transformers; multi-task shows 60% content words vs single-task's 40% (more semantic grounding).