Utilizing NLP to Identify Police Corruption in the USA

Progress Review 2

Presented By : Theekshana Samaradiwakara (258138R)
Supervised By : Dr. Nisansa de silva
Externally Supervised By : Attorney George C. Lobb

Content

- Introduction
- Recap of Progress Review 1
- Completed Research Progress
- Current and Planned Work
- References

Introduction

Introduction

Deception - Intentional act of misleading others by providing false, incomplete, or strategically distorted information [1]

- Commission - explicit falsehoods
- Omission - withholding relevant information
- Equivocation - use of vague or ambiguous language
- Concealment - masking important facts while maintaining plausible deniability

[1] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. In Proceedings of the 2015 ACM on international conference on multimodal interaction. 59–66

Introduction.

- Deception detection is critical in law enforcement, judicial analysis, fraud detection, and digital communication.
- Human detection accuracy is only ~54–60%, barely above random guessing. [2]
- Natural Language Processing (NLP) enables scalable analysis of linguistic patterns associated with deception.
- Research has evolved from hand-crafted linguistic features ➔ traditional ML models ➔ context-aware Large Language Models (LLMs).

[2] Jiawei Li, Wen-Hao Chen, Qing Xu, Neal Shah, Jillian C Kohler, and Tim K Mackey. 2020. Detection of self-reported experiences with corruption on twitter using unsupervised machine learning. Social Sciences & Humanities Open 2, 1 (2020), 100060.

Theoretical Foundations

1. Reality Monitoring (RM) Theory [3]
   - Truthful statements contain more **sensory and contextual detail**
2. Cognitive Load Theory [4]
   - Lying **increases mental effort**, results in **simpler language** structure
3. Truth Default Theory (TDT) [5]
   - Humans assume **honesty** by **default**, lies are detected only when anomalies trigger suspicion

[3] Riccardo Loconte, Chiara Battaglini, Stéphanie Maldera, Pietro Pietrini, Giuseppe Sartori, Nicolò Navarin, and Merylin Monaro. 2025. Detecting Deception Through Linguistic Cues: From Reality Monitoring to Natural Language Process- ing. Journal of Language and Social Psychology 44, 3-4 (2025), 523–552.
[4] Valerie Hauch, Iris Blandón-Gitlin, Jaume Masip, and Siegfried L Sporer. 2015. Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. Personality and social psychology Review 19, 4 (2015), 307–342.
[5] Timothy R Levine. 2014. Truth-default theory (TDT) a theory of human deception and deception detection. Journal of Language and Social Psychology 33, 4 (2014), 378–392

Theoretical Foundations.

4. Linguistic Category Model (LCM) [6]
   - A framework for understanding abstraction levels in language.
   - Deceptive messages often rely on more **abstract and generalised expressions**, allowing speakers to remain flexible and avoid commitment to specific details.
5. Psycholinguistic Markers [7]
   - Identified in empirical studies.
   - Eg- Use of **negative emotion words, fewer self-references** (e.g., “I,” “me”) and increased generalisation, **reduced use of exclusive words like "but", "except", and "without"**

[6] Justyna Sarzynska-Wawer, Aleksandra Pawlak, Julia Szymanowska, Krzysztof Hanusz, and Aleksander Wawer. 2023. Truth or lie: Exploring the language of deception. Plos one 18, 2 (2023), e0281179.
[7] Joni Salminen, Mekhail Mustak, Soon-Gyo Jung, Hannu Makkonen, and Bernard J Jansen. 2025. Decoding deception in the online marketplace: enhancing fake review detection with psycholinguistics and transformer models. Journal of Marketing Analytics (2025), 1–18.

Research Objectives

This project aims to develop an LLM-based deception detection framework for legal and police reports, addressing: data scarcity, ethical reliability, and reasoning transparency in high-stakes decision-making environments.

- Compare existing approaches on standardized deception corpora.
- Collect a real-world deception dataset comprised of police reports.
- Evaluate various prompt engineering paradigms in deception reasoning.
- Develop an LLM-based deception detection framework.

Recap of Progress Review 1

Comparative Analysis

Objective

- Evaluate deception detection models across legal and general-domain datasets.
- Compare performance across different model families and prompting strategies.

  - 7 datasets benchmarked - 2 legal + 5 general domain
  - 5 fine-tuned encoders, 8 LLMs
  - 5 prompting strategies (SFT, zero-shot, few-shot, cot)

Benchmark Datasets

7 datasets benchmarked - 2 legal + 5 general domain

Dataset	Domain	Data Type	Truthful	Deceptive	Total
RLTD	legal	interview	60 (49.6%)	61 (**50.4%**)	121
DECOUR	legal	interview	1,202 (**56.0%**)	945 (44.0%)	2,147
OpSpam	general	review	800 (50.0%)	800 (50.0%)	1,600
cCult	general	opinion	606 (50.0%)	606 (50.0%)	1,212
DeRev2014	general	review	118 (50.0%)	118 (50.0%)	236
Liar	general	news	**7,134** (55.8%)	**5,657** (44.2%)	**12,791**
FakeNewsNet	general	news	211 (50.0%)	211 (50.0%)	422

[8] Tommaso Fornaciari, Massimo Poesio, et al . 2012. DeCour: a corpus of DEceptive statements in Italian COURts.. In LREC. 1585–1590.
[9] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint arXiv:1107.4557 (2011).
[10] Verónica Pérez-Rosas and Rada Mihalcea. 2014. Cross-cultural deception detec- tion. In Proceedings of the 52nd Annual Meeting of the Association for Computa- tional Linguistics (Volume 2: Short Papers). 440–445.
[11] Tommaso Fornaciari and Massimo Poesio. 2014. Identifying fake amazon reviews as learning from crowds. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Association for Compu- tational Linguistics, 279–287.
[12] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648 (2017).
[13] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big data 8, 3 (2020), 171–188.

Models Evaluated

5 fine-tuned encoders, 8 LLMs

Transformer Encoder Models

- RoBERTa
- BERT
- DeBERTa
- ALBERT
- DistilBERT

Open-source LLMS

- T5-base
- LLaMA3-8B
- LLaMA3.1-8B
- Gemma2-9B
- Phi-3-mini
- Phi-4

Closed-source LLMS

- GPT-4o
- GPT-4o Mini

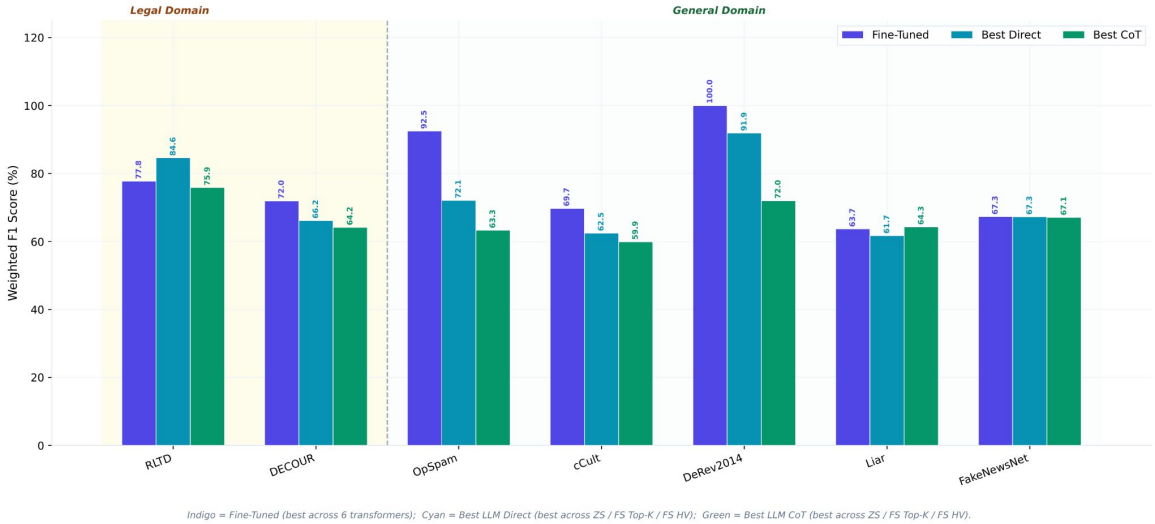
Prompt Strategies

8 LLMs benchmarked × 5 prompting strategies (SFT, zero-shot, few-shot, cot)

- Supervised Fine-tuning
- Zero-shot Direct Classification
- Few-shot Direct Classification
- Zero-shot Chain-of-Thought (CoT)
- Few-shot Chain-of-Thought (CoT)

Experiments Conclusion

- Model performance is highly domain dependent.
- Fine-tuning outperforms prompting given sufficient data.
- High Variance example selection consistently outperforms Top-K selection for LLMs on most datasets.
- Chain-of-Thought is Model/Dataset Dependent.



Best result per method per dataset

Completed Research Progress

Papers Published, Presented and Submitted

- Got accepted and presented the paper titled “Semantics of Subterfuge: Benchmarking Legal Deception Detection Against General-domain State-of-the-Art” at **ICAIL 2026**.

- Submitted the paper titled “Why Prompted LLMs Fail at Legal Deception Detection and Where the Signal Actually Is” at **PACLIC 2026**.

Approach

Four studies on the Real-Life Trial Deception dataset

1. **Furthur prompt strategies**
   - Multi-agent debate - Richer reference cases - Role-specific example selection
   - LLM-driven prompt optimization
2. **Cross-Domain Transfer Learning**
   Fine-tune on six source datasets, evaluate on RLTD
3. **Hidden-Layer Probing**
   Linear probes on frozen per-layer hidden states of the same open-weight LLMs and encoders
4. **SHAP**
   Cross-validated token-level SHAP on encoders; feature-level SHAP on LLM probes

Benchmark Datasets

Real-Life Trial Deception (RLTD): [1]

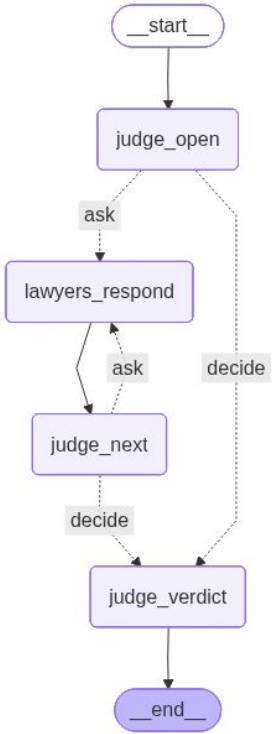
- 121 courtroom video clips
- 61 deceptive / 60 truthful statements
- Labels derived from trial outcomes
- Video transcripts created via crowdsourcing
- Only text transcripts used for experiments

[1] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. In Proceedings of the 2015 ACM on international conference on multimodal interaction. 59–66

Study 01 - Reasoning Experiments

Agentic (Judge and Lawyers) Debate System

- 1 Judge
- 1 Prosecution lawyer
- 1 Defence lawyer
- Judge can ask questions up to 3 times.
- After lawyers reasoning, judge gives final verdict (Deceptive, Truthful).



Study 01 - Reasoning Experiments .

Agentic Debate Prompt Settings

1. Zero-shot
2. Few-shot top k (most similar) examples
3. Few-shot high variance (most diverse) examples
4. Few-shot top k - Descriptive Reference Cases
5. Few-shot high variance - Descriptive Reference Cases

Study 01 - Reasoning Experiments ..

Agentic Debate Prompt Settings .

6. Role-Specific Example Selection
    - Prosecution receives top k examples, and the defense receives high variance examples
7. LLM-Driven Prompt Optimisation
    - An optimiser LLM evaluate results and rewrite the agents prompts
    - For each of the 96 training statements, in order, the system runs the debate, compares the verdict to the ground truth, and the optimiser edits the judge, prosecution, and defense system prompts.

Experiment Results - RLTD

Agentic - Openai on RLTD



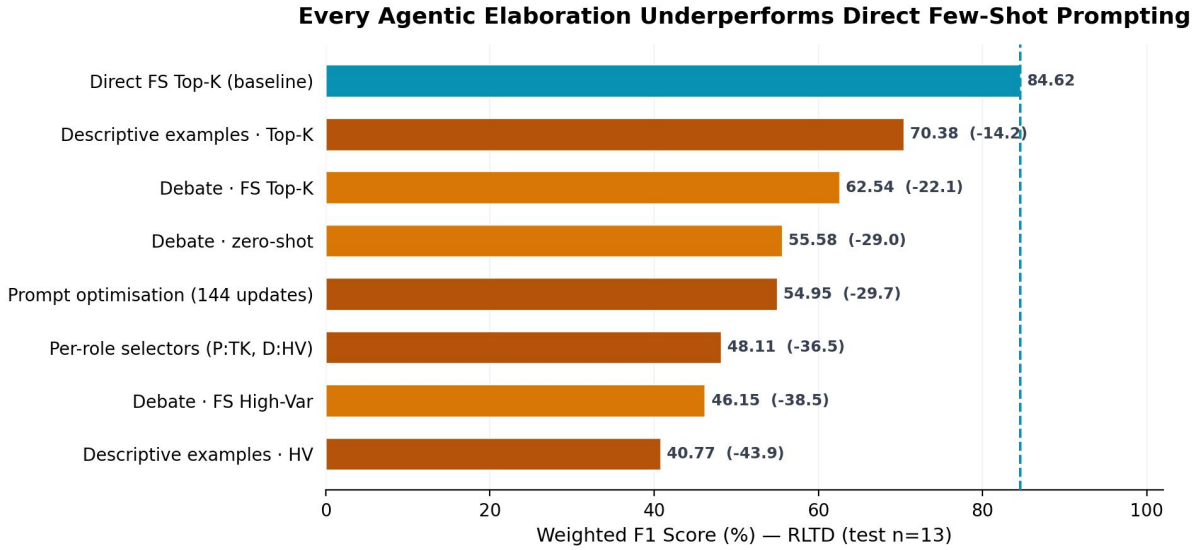
Experiment Results - RLTD

Agentic - Openai on RLTD



Experiment Results - RLTD

Agentic - Openai on RLTD



Study 01 - Reasoning Experiments ...

Training-Time Rule Learning

Method

- Iterative instruction refinement method.
- The model classifies labelled examples sequentially on train set.
- Second LLM updates a prompt rule list, reinforcing rules on correct predictions and errors.
- Relevant rules are de-duplicated and merged together.
- The final prompt is tested on test data set.

Study 01 - Reasoning Experiments ...

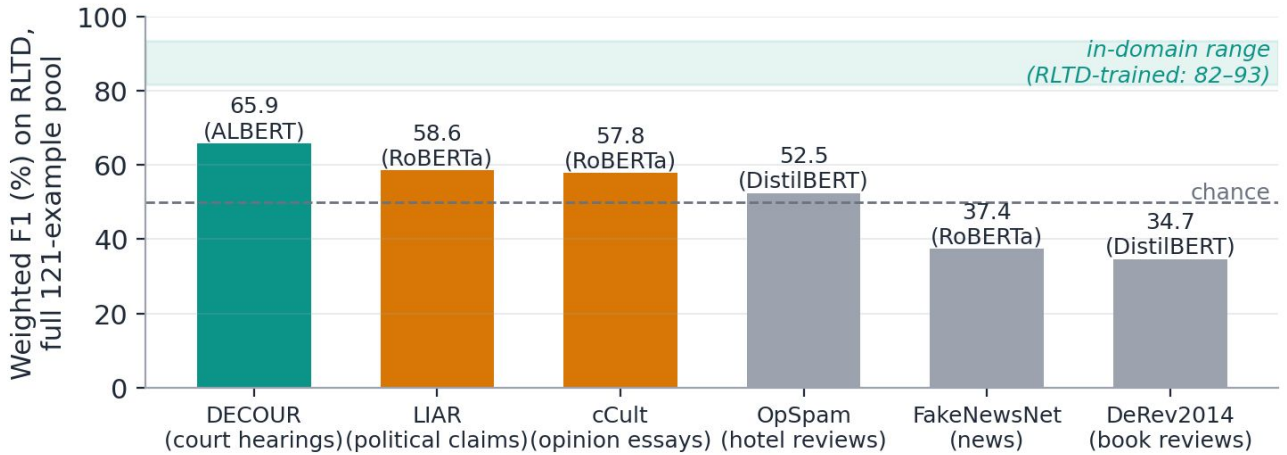
Training-Time Rule Learning

Results (LLaMA3-8B)

- The results are worse than both its zero-shot and few-shot settings.
- The model often assigns the observed cues in the wrong direction.
  - 191 rules learned across the 10 folds.
  - 18 mention disfluencies or filler words, and 13 (72%) assign them to the deceptive class.
  - Disfluencies - break in the smooth speak flow *("um," "uh," "like", Repetitions)*
  - This is opposite to the cross-validated SHAP result, where disfluencies are among the strongest truthfulness markers.

Study 02 - Cross-Domain Transfer to RLTD

- Tested whether larger general-domain deception datasets can substitute for in-domain testimony.
- Fine-tuned each encoder on one source dataset and evaluating on all 121 RLTD statements.



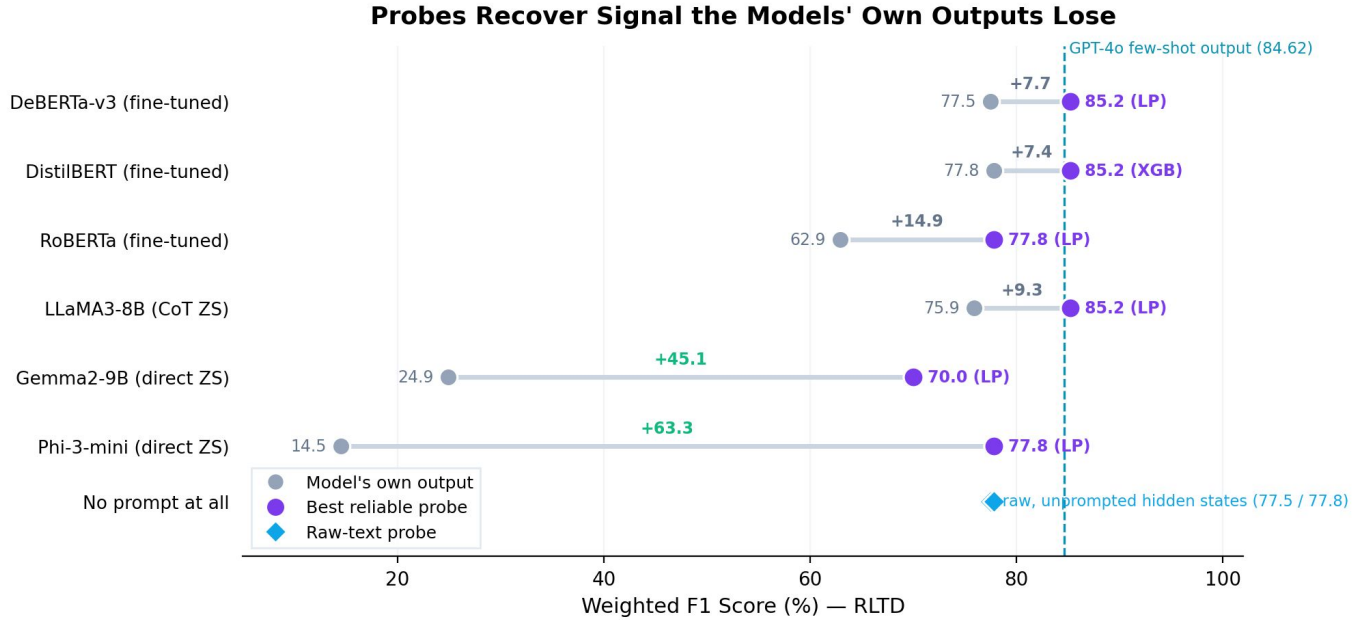
Study 03 - Hidden-Layer Probing

Method

- Extract per-layer hidden states from the five fine-tuned encoders and the frozen open-weight LLMs
- Four settings: raw unprompted text, direct ZS/FS prompts, CoT ZS/FS prompts
- Train a logistic-regression linear probe per layer and XGBoost probes
- Select the best layer on validation; evaluate once on test

Model (regime)	Own	LP	XGB
Fine-tuned encoders			
DeBERTa-v3 (fine-tuned)	77.5	**85.2**	75.9
DistilBERT (fine-tuned)	77.8	70.0	**85.2**
RoBERTa (fine-tuned)	62.9	77.8	100.0†
Frozen LLMs, prompted			
Phi-3-mini (direct ZS)	14.5	**77.8**	62.5
Gemma2-9B (direct ZS)	24.9	**70.0**	53.9
LLaMA3-8B (CoT ZS)	75.9	**85.2**	69.2
Frozen LLMs, raw text			
LLaMA3-8B (raw, no prompt)	—	77.5	48.1
Gemma2-9B (raw, no prompt)	—	77.8	48.1
Surface-lexical control			
TF-IDF (surface lexical)	—	70.0	70.4

Experiment Results - RLTD - Linear probing



Study 04 - SHAP experiments

Method

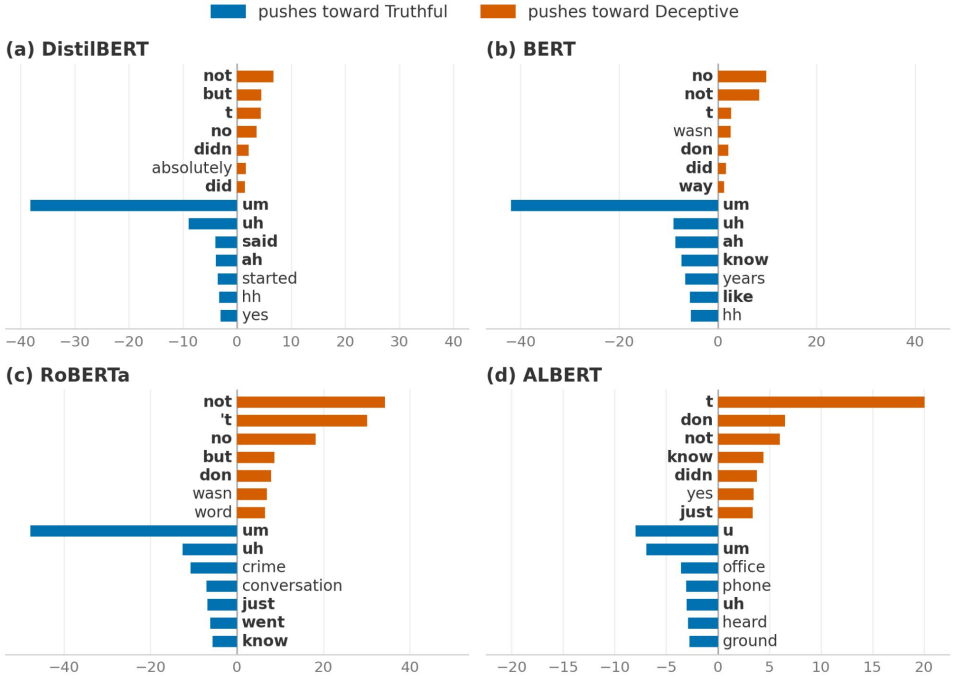
- Token-level SHAP under 10-fold CV
- Four encoder models (RoBERTa, DistilBERT, BERT, ALBERT).
- Strongest tokens pooled across all four models.

Results

- Across models the cue pattern is similar.
- Negation and denial (not, no, don’t, 471 didn’t, wasn’t) push toward deceptive.
- Disfluencies (um, uh, ah) and concrete detail (years, started, said, 474 went) push toward truthful.
- This pattern matches Reality Monitoring and Cognitive Load theory.

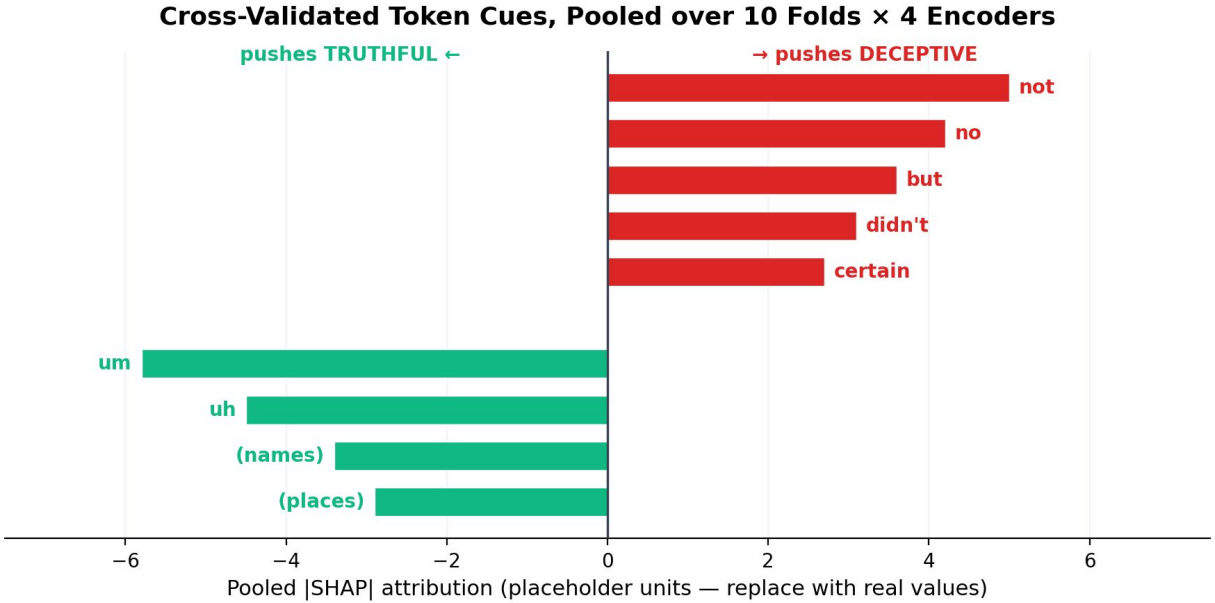
Experiment Results - RLTD - SHAP

Bert based models SHAP contributions on RLTD



Experiment Results - RLTD - SHAP

Bert based models SHAP summary on RLTD



Experiment Conclusion

- Adding reasoning and agentic steps reduces performance on RLTD:
  - ⇢ 62.54 weighted F1 vs 84.62 with direct few-shot GPT-4o.
- Linear probes perform better
  - ⇢ Frozen hidden states can recover up to 63 F1 points beyond the models' own predictions.
- Self-generated rules can backfire. They often invert the useful cues.
- Other deception datasets don't transfer well. Their cues are not reliable for RLTD.
- SHAP analysis identifies a small cue set consistent with psycholinguistic theories.

Current and Planned Work

Data Acquisition Plan

- Collect and label a novel dataset of police reports linked to misconduct and corruption from open databases and Public Information Act (PIA), Freedom of Information Act (FOIA) requests.
- Focus agency: **Austin Police Department (APD)**, Austin, Texas.

Two independent acquisition channels

1. **TCOLE Public License Lookup** — the Texas state licensing regulator. Officer identity, service and certification history. licenselookup.tcole.texas.gov/PublicLookup/
2. **APD / Office of Police Oversight records portal** — the documents themselves: complaints, disciplinary memos, oversight recommendations, and reports released under the Texas Public Information Act. apd-austintx.govqa.us

Data Collected - APD

Source 1: TCOLE

- Scraped with Selenium from the TCOLE public licence lookup.
- Agency - Austin Police Dept.
- Peace officers - 1,516
- Per officer -
  - Name, licence number, licence status, gender, race
  - Licences held, with type, status and grant date
  - Certificates and awards
  - Full agency service history
  - Training and education

NOTE : This data do not include officer reports or statements. Used to get police officer information.

Data Collected - APD .

Source 2: APD records portal

2A · Published oversight report website (bulk scrape)

- Police complaints made by public
- 964 unique PDFs / 19,033 pages
- complaint 524 · misconduct 181 · discipline 163 · oversight 61 · use of force 10 · other 10
- documents spanning 2018–2025
- Structured extraction complete on the 86 disciplinary / complaint reports ➜ 473 officer text spans, 315 charges, 79 unique officers,

NOTE : This data do not include officer first-person narratives. Only includes **second or third order narratives.**

Data Collected - APD ..

Source 2: APD records portal

2B · Our own PIA requests

- 16 Police reports for 16 cases.
  - **11** Driving While Intoxicated (DWI) reports.
  - **5** general offense reports
- Extracted 57 officer narratives
- Across 31 officers
- **First-person narratives** (Reports are written by corresponding officers for the case)

Data Collected - APD ..

Source 2: APD records portal

2B · Our own PIA requests

- 16 Police reports for 16 cases.
  - **11** Driving While Intoxicated (DWI) reports.
  - **5** general offense reports
- Extracted 57 officer narratives
- Across 31 officers
- **First-person narratives** (Reports are written by corresponding officers for the case)

Data Requested - APD

- Report outcomes (Sustained / Exonerated)
- More police officer reports
- Chief's disciplinary memoranda

Data Collected - CPD

- Police reports from Chicago Police Department
  - Collected from “Invisible Institute” journal police information collection.
  - Complaint Register files with investigation summaries.
- 30 Reports
  - 30 cases, 725 statement segments
  - Label present - **Sustained 21 / Exonerated 9**
  - Use of force 15 · False arrest 4 · Illegal search 3 · Domestic 3 · Verbal abuse 2 . Other - 3

Experiments - CPD

Encoder model fine tuning

	ACC %	F1 Weighted %	F1 Macro %
Majority baseline	70	57.60	41.12
Untrained BERT head	63.30	63.90	57.60
BERT, fine-tuned on RLTD	66.7	68.10	65.30
BERT, fine-tuned on CPD	**83.30**	**81.00**	**75.50**

Modeling and Evaluation Plan

Phase 1 · Get labels Data

Phase 2 · Data Release and redaction

Phase 3 · Modelling

- Multi-dimensional score, not one probability: encoder score · hidden-state linear probe · SHAP cue profile · theory-grounded feature counts (Reality Monitoring detail, hedging, verifiability) · reasoning trace.
- Intra-document classification at sentence and paragraph level.
- Explanation aspect as required per USA law to take actions
- Writing style analysys

References

[1] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. In Proceedings of the 2015 ACM on international conference on multimodal interaction. 59–66

[2] Jiawei Li, Wen-Hao Chen, Qing Xu, Neal Shah, Jillian C Kohler, and Tim K Mackey. 2020. Detection of self-reported experiences with corruption on twitter using unsupervised machine learning. Social Sciences & Humanities Open 2, 1 (2020), 100060.

[3] Riccardo Loconte, Chiara Battaglini, Stéphanie Maldera, Pietro Pietrini, Giuseppe Sartori, Nicolò Navarin, and Merylin Monaro. 2025. Detecting Deception Through Linguistic Cues: From Reality Monitoring to Natural Language Process- ing. Journal of Language and Social Psychology 44, 3-4 (2025), 523–552.

[4] Valerie Hauch, Iris Blandón-Gitlin, Jaume Masip, and Siegfried L Sporer. 2015. Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. Personality and social psychology Review 19, 4 (2015), 307–342.

[5] Timothy R Levine. 2014. Truth-default theory (TDT) a theory of human deception and deception detection. Journal of Language and Social Psychology 33, 4 (2014), 378–392

[6] Justyna Sarzynska-Wawer, Aleksandra Pawlak, Julia Szymanowska, Krzysztof Hanusz, and Aleksander Wawer. 2023. Truth or lie: Exploring the language of deception. Plos one 18, 2 (2023), e0281179.

[7] Joni Salminen, Mekhail Mustak, Soon-Gyo Jung, Hannu Makkonen, and Bernard J Jansen. 2025. Decoding deception in the online marketplace: enhancing fake review detection with psycholinguistics and transformer models. Journal of Marketing Analytics (2025), 1–18.

[8] Tommaso Fornaciari, Massimo Poesio, et al . 2012. DeCour: a corpus of DEceptive statements in Italian COURts.. In LREC. 1585–1590.

[9] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint arXiv:1107.4557 (2011).

[10] Verónica Pérez-Rosas and Rada Mihalcea. 2014. Cross-cultural deception detec- tion. In Proceedings of the 52nd Annual Meeting of the Association for Computa- tional Linguistics (Volume 2: Short Papers). 440–445.

[11] Tommaso Fornaciari and Massimo Poesio. 2014. Identifying fake amazon reviews as learning from crowds. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Association for Compu- tational Linguistics, 279–287.

[12] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648 (2017).

[13] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big data 8, 3 (2020), 171–188.