Utilizing NLP to Identify Police Corruption in the USA

Progress Review 1

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

Content

Introduction

Research Objectives

Proposed Methodology

Current Research Progress

Planned Work

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

Reality Monitoring (RM) Theory [3]

- Truthful statements contain more sensory and contextual detail

Cognitive Load Theory [4]

- Lying increases mental effort, results in simpler language structure

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.

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.

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.

Evolution of NLP in Deception Detection

Text based Automatic deception detection (ADD) has evolved through three distinct phases.

1. **Machine learning classifiers**

- Relied on manually engineered linguistic features including LIWC categories, POS tags, syntactic patterns, and lexical statistics. [8]
- These offered interpretability aligned psycholinguistic theories but struggled with capturing long-range dependencies and nuanced contextual cues. [9]

[8] Alberto Alejandro Ceballos Delgado, William Bradley Glisson, Narasimha Shashidhar, J Todd McDonald, George Grispos, and Ryan Benton. 2021. Detecting deception using machine learning. Proceedings of the 54th Hawaii International Conference on System Sciences.
[9] Maria Glenski, Ellyn Ayton, Robin Cosbey, Dustin Arendt, and Svitlana Volkova. 2020. Towards trustworthy deception detection: Benchmarking model robustness across domains, modalities, and languages. In Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM). 1–13.

Evolution of NLP in Deception Detection.

Text based Automatic deception detection (ADD) has evolved through three distinct phases.

2. Deep learning classifiers

- Shifted the focus toward automatic representation learning, with neural architectures such as Convolutional neural networks (CNNs), Recurrent Neural Networks (RNNs), LSTMs/BiLSTMs and attention-based models. [10]
- Capture sequential patterns and semantic context directly from raw text. [11]

[10] Roshan R Karwa and Sunil R Gupta. 2022. Automated hybrid Deep Neural Network model for fake news identification and classification in social networks. Journal of Integrated Science and Technology 10, 2 (2022), 110–119.
[11] Alpana A Borse, Gajanan K Kharate, and Namrata G Kharate. 2025. 4HAN: hypergraph-based hierarchical attention network for fake news prediction. In- ternational Journal of Electrical and Computer Engineering (IJECE) 15, 2 (2025), 2202–2210

Evolution of NLP in Deception Detection..

Text based Automatic deception detection (ADD) has evolved through three distinct phases.

3. Transformer models

- Enhanced context modelling through cross-sentence dependencies and long range relationships.
- Capture sequential patterns and semantic context directly from raw text.
- Eliminate explicit feature engineering.
- Leverage pre-trained knowledge and prompt engineering to guide classification, reasoning frameworks like Chain-of-Thought (CoT), Tree-of-Thoughts (ToT) [12]

[12] Katerina Papantoniou, Panagiotis Papadakos, and Dimitris Plexousakis. 2025. Evaluating LLMs on Deceptive Text Across Cultures. In RANLP. 884–893.

Challenges in Deception Detection

Data Quality and Availability

- Lack of large, high-quality, and ecologically valid datasets [13]
- Ground truth is often ambiguous in real scenarios. [5]
    - In legal transcripts, the ’truth’ is only definitively known after a confession, a verdict, or the discovery of concrete evidence.
- Controlled lab datasets often fail to capture the intense cognitive load and emotional complexity inherent in real-world, high-stakes deception. [14]

[13] Eileen Fitzpatrick and Joan Bachenko. 2012. Building a data collection for de- ception research. In Proceedings of the workshop on computational approaches to deception detection. 31–38.
[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.
[14] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. Personality and social psychology bulletin 29, 5 (2003), 665–675.

Challenges in Deception Detection.

Linguistic and Psychological Complexity

- Deceptive strategies based on personality, stakes, preparation, and context [1]
- Difficult for any single model to capture a universal signature of deceit.
- Correlations between linguistic features (pronoun usage, cognitive process words) and deception are statistical rather than deterministic. [4]
- This lead to false positives when honest individuals experience cognitive strain.

[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.
[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.

Challenges in Deception Detection..

Model Robustness and Interpretability

- Concerns related to trustworthiness: susceptibility to adversarial linguistic perturbations. [15]
- Limited cross-cultural and cross-language generalisation. [6]
- Lack of interpretability [16]

[15] Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. arXiv preprint arXiv:2310.15515 (2023).
[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.
[16] Stefan Feuerriegel, Abdurahman Maarouf, Dominik Bär, Dominique Geissler, Jonas Schweisthal, Nicolas Pröllochs, Claire E Robertson, Steve Rathje, Jochen Hartmann, Saif M Mohammad, et al. 2025. Using natural language processing to analyse text data in behavioural science. Nature Reviews Psychology 4, 2 (2025), 96–111.

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.

Related Work - Legal & Police Domains

Year	Application	Related Data/Datasets	Model(s)
2015	Deception classification	RLTD dataset [1]	Decision Tree, Random Forest
2023	Cross-Corpus deception detection	RLTD dataset [1]	RoBERTa-base
2024	Deception classification	RLTD dataset [1]	BiLSTM
2025	Direct label prediction, Post-hoc reasoning generation	RLTD dataset [1]	LLaMA3.1-8B, Gemma2-9B, GPT-4o
2012	Deception classification	DECOUR dataset [17]	SVM
2013	Deception classification	DECOUR dataset [17]	SVM
2021	Deception classification	DECOUR dataset [17]	BERT + Transformers
2018	Detection of false robbery reports	Spanish police reports [18]	SVM, Ridge LR
2024	Deception reasoning for criminal law documents	CAIL2018 [19] , Synthetic dialogues	GLM-4-9B, GPT-3.5, Gemini-1.5-Pro, Qwen2-7B
2025	Identify indicators of vulnerability	Boston police narrative reports [20]	LLaMA 8B/70B, GPT-4o

[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
[17] Tommaso Fornaciari, Massimo Poesio, et al . 2012. DeCour: a corpus of DEceptive statements in Italian COURts.. In LREC. 1585–1590.
[18] Lara Quijano-Sánchez, Federico Liberatore, José Camacho-Collados, and Miguel Camacho-Collados. 2018. Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police. Knowledge-Based Systems 149 (2018), 155–168.
[19] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al . 2018. Cail2018: A large-scale legal dataset for judgment prediction. arXiv preprint arXiv:1807.02478 (2018).
[20] Sam Relins, Daniel Birks, and Charlie Lloyd. 2025. Using Instruction-Tuned Large Language Models to Identify Indicators of Vulnerability in Police Incident Narratives. Journal of Quantitative Criminology (2025), 1–38.

Comparative Analysis

Objective

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

Dataset Selection

- 7 Datasets total - 2 Legal domain, 5 General
- To analysis of cross-domain model performance.

Benchmark Datasets

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

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

Benchmark Datasets

DECOUR: [17]

- Courtroom data transcripts of 35 hearings for criminal proceedings held in Italian courts.
- Dialogues between interviewees and interviewers (judges, prosecutors, and lawyers).
- Each interviewee's utterance is labelled True, False, or Uncertain
- Total of 3015 labelled utterances (1202 True, 868 Uncertain, 945 False).

For our experiments, only True/False utterance used to maintain a consistent binary classification task across all datasets

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

Benchmark Datasets

OpSpam [21]

- Deceptive hotel reviews from TripAdvisor
- Reviews covering 20 Chicago hotels

Cross-cultural Deception Detection (cCult) [22]

- Crowdsourced short essays
- Topics: Abortion, Death Penalty, Best Friend

[21] 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).
[22] 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.

Benchmark Datasets

DeRev2014 [23]

- 236 deceptive and truthful book reviews

Liar [24]

- Fake political statements from PolitiFact

FakeNewsNet [25]

- Fake news articles from PolitiFact and BuzzFeed

[23] 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.
[24] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648 (2017).
[25] 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.

Data Processing

Data Splitting

- The Liar dataset comes with train, valid and test splits with sizes of 10240, 1284 and 1267 respectively.
- Calculated percentages of data splits are train = 80.06% , dev = 10.04% , test = 9.91%.
- For consistency across datasets, we used unified random splits in **80 / 10 / 10** for **train / validation / test** splits for the remaining datasets.

Data Processing.

LIAR Dataset Adjustment

- Originally labels - 6
- Mapped in to binary labels

Original Labels	Mapped Label
True, Mostly True, Half True	True
False, Barely True, Pants Fire	False

Models Evaluated

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

Experimental Setup

- Hardware - NVIDIA RTX 3080 (16GB)
- Fine-tuning Configuration
  - Learning rate: 2e-5
  - Batch size: 8
  - Epochs: 6
  - Optimizer: AdamW
- LLM Inference
  - Local models run using Ollama + 4-bit quantization (Q4_K_M)
  - GPT-4o / GPT-4o-mini via Chat Completions API
- Evaluation metric - Weighted F1

Prompt Strategies

- Supervised Fine-tuning
- Zero-shot Direct Classification
- Few-shot Direct Classification
- Zero-shot Chain-of-Thought (CoT)
- Few-shot Chain-of-Thought (CoT)
- Zero-Shot Agentic (Judge and Lawyers)
- Few-shot Agentic (Judge and Lawyers)

Few-Shot selection

Semantic similarity-based few-shot example selection for deception detection.

- Embeddings - Each text is converted into a vectors (384 dimensions) using "all-MiniLM-L6-v2" model [26] .
- Cosine similarity - The similarity between two vectors is measured.
- The pairwise matrix - Pre-computes an N×N matrix of every training example's similarity to every other training example.

[26] Hugging Face (n.d.). sentence-transformers/all-MiniLM-L6-v2 · Hugging Face. [online] huggingface.co. Available at: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

Few-Shot selection.

Two selection strategies [27]

- Top_k

  Select the k training examples with the HIGHEST cosine similarity to the test query. Retrieves the most semantically relevant neighbours.

- High_var

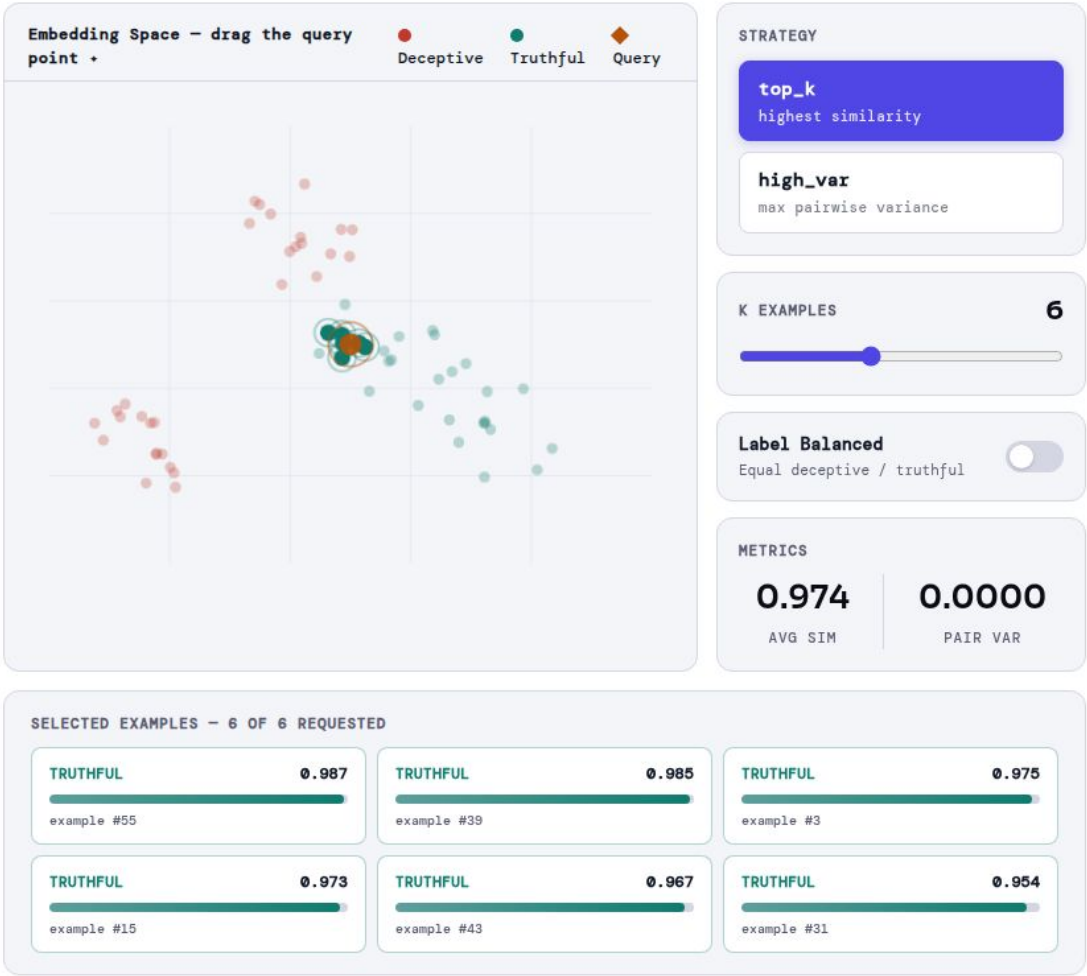
  First pick the single training example most similar to the query.

  Selects k examples whose pairwise similarities to each other have maximum variance (numpy.var).

[28] Miah, M.M.M., Anika, A., Shi, X. and Huang, R., 2025, July. Hidden in plain sight: Evaluation of the deception detection capabilities of LLMs in multimodal settings. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 31013-31034).

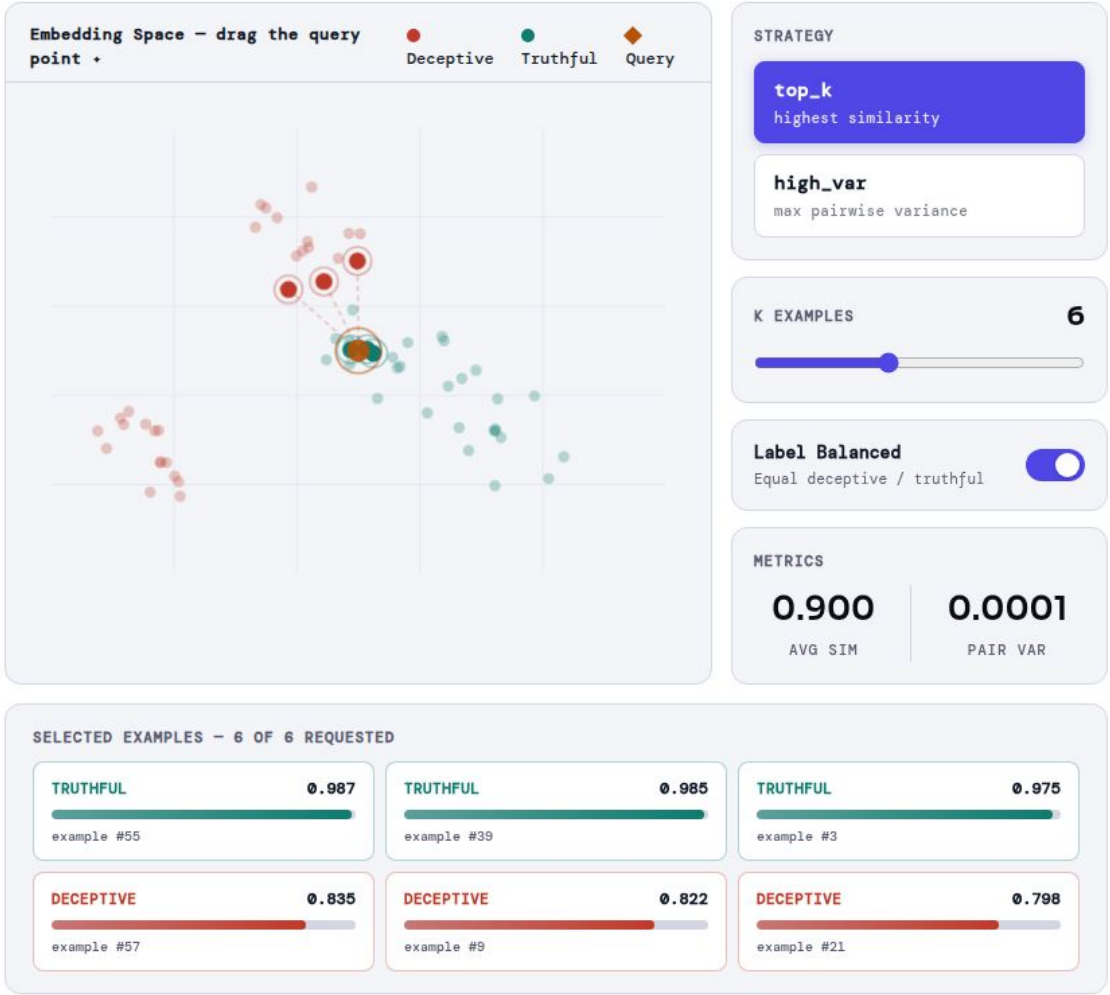
Few-Shot selection

- Top_k
- Labels unbalanced



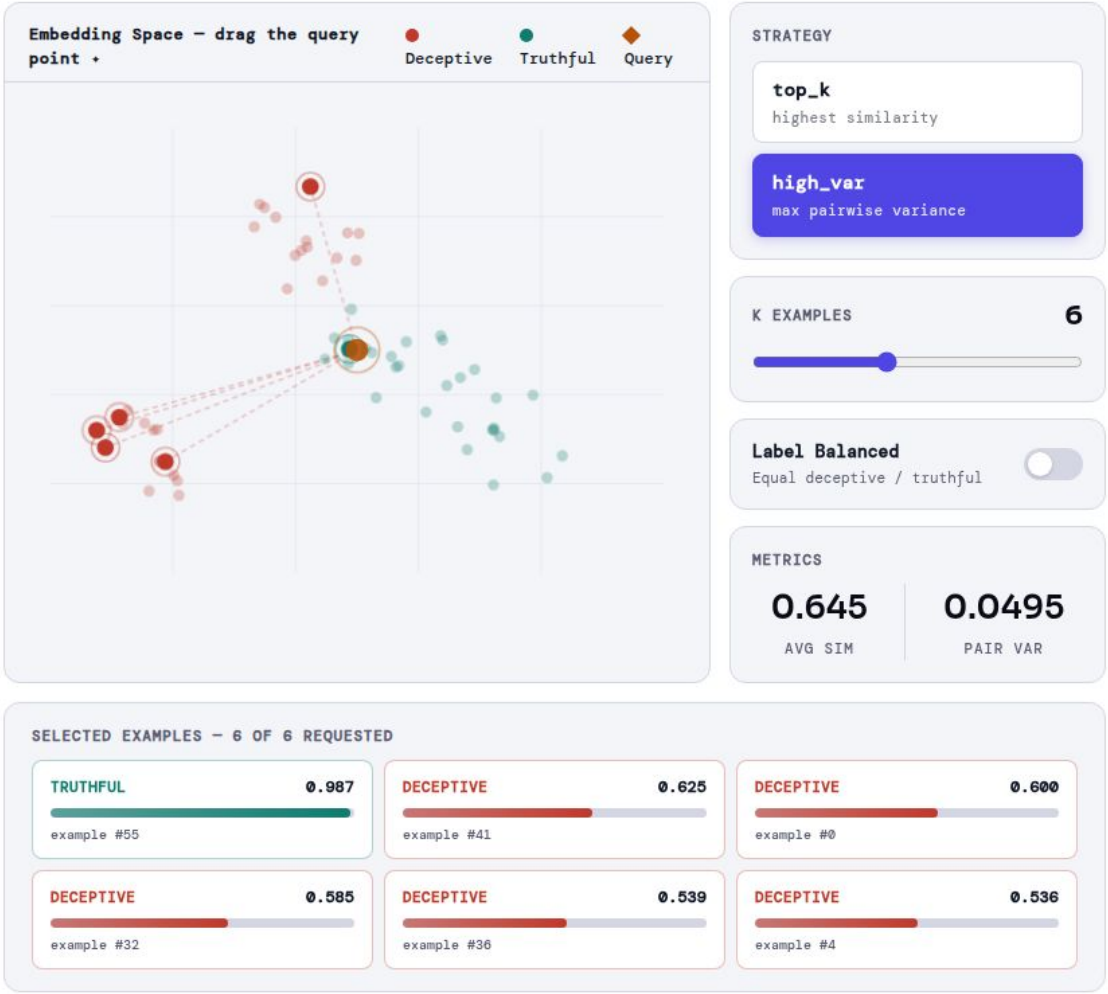
Few-Shot selection

- Top_k
- Labels balanced



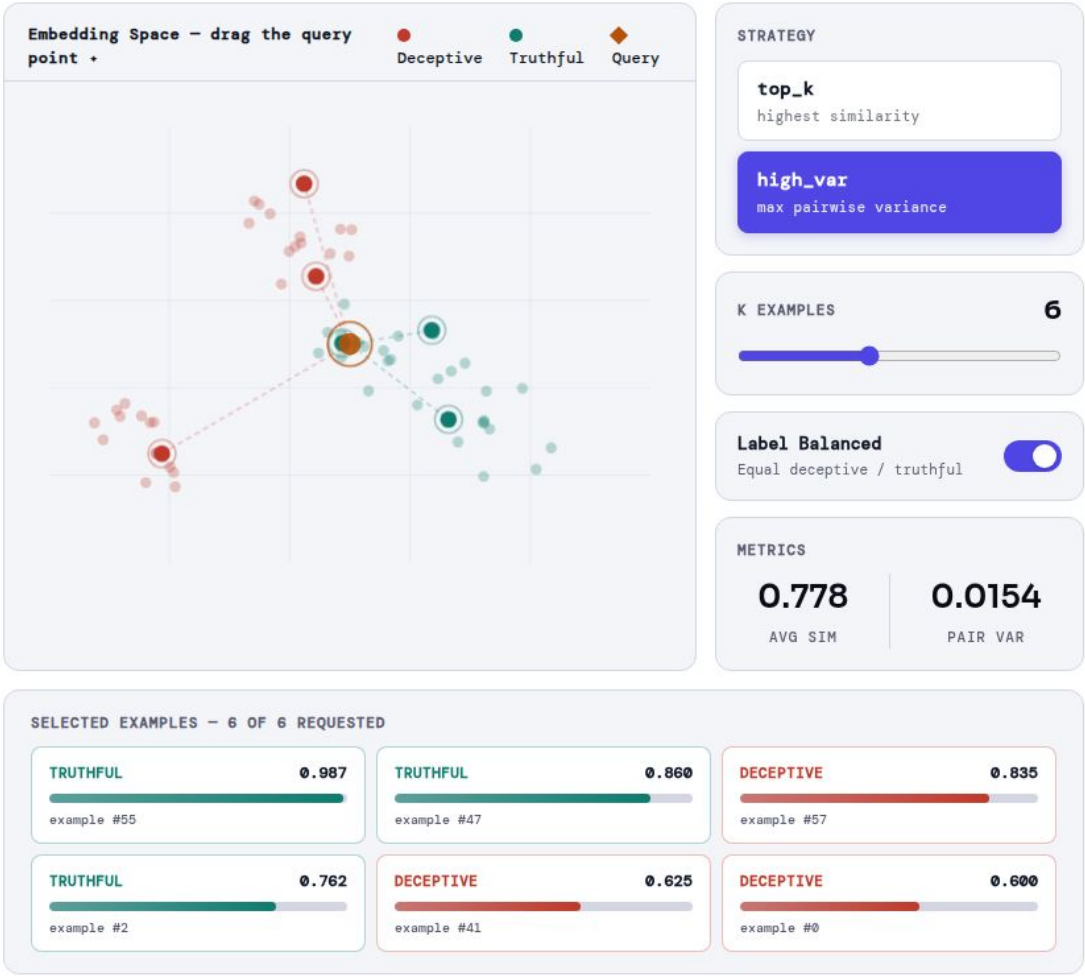
Few-Shot selection

- High_var
- Labels unbalanced



Few-Shot selection

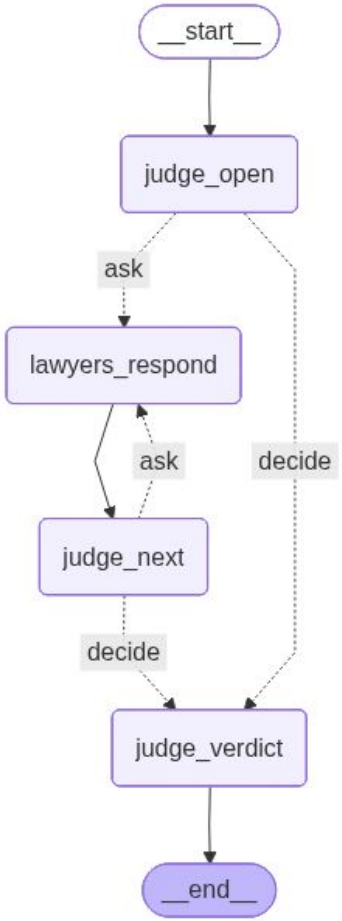
- High_var
- Labels balanced



Prompt Strategies

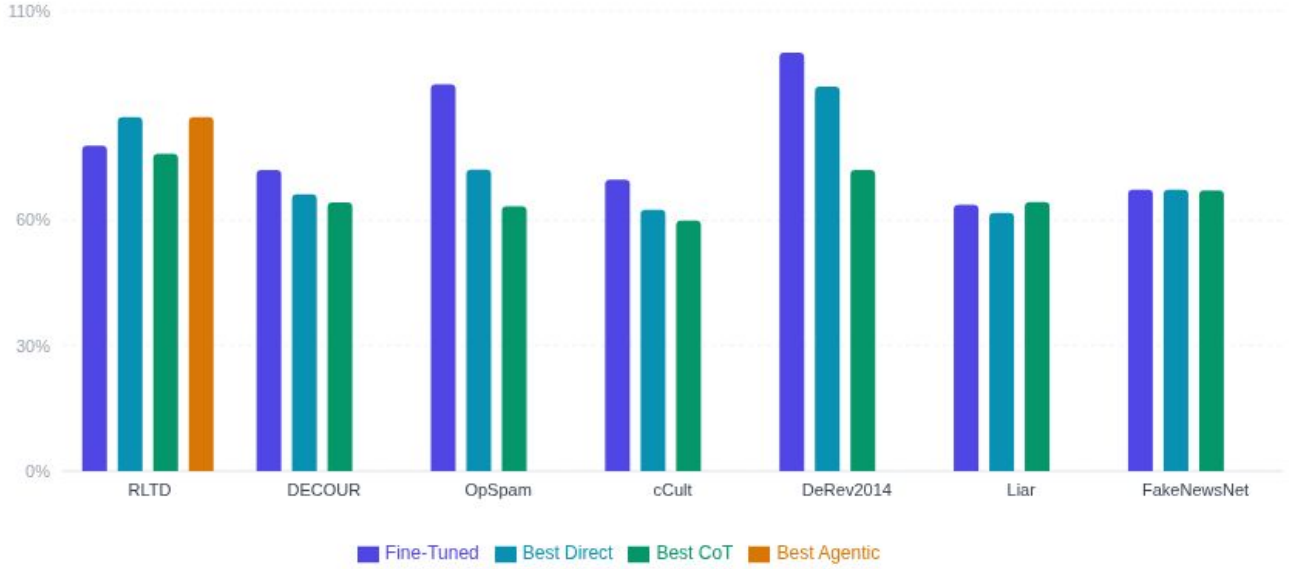
Agentic (Judge and Lawyers)

- 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).



Experiment Results

Best result per method per dataset



Experiment Results

Supervise fine tune results

MODEL	RLTD	DECOUR	OPSPAM	CCULT	DEREV2014	LIAR	FAKENEWSNET
RoBERTa	62.9	68.9	91.3	65.9	100.0	63.4	64.5
BERT	48.1	68.3	90.0	69.7	96.0	63.7	57.5
DeBERTa	77.5	70.8	92.5	69.0	100.0	62.3	62.9
ALBERT	53.9	68.0	90.1	63.5	96.0	61.8	64.5
DistilBERT	77.8	72.0	88.8	68.8	100.0	60.1	63.6
T5	14.5	68.4	90.7	65.9	47.3	62.4	67.3

Experiment Results

Supervise fine tune results

Experiment Results

LLM direct classification - Zero-Shot

MODEL	RLTD	DECOUR	OPSPAM	CCULT	DEREV2014	LIAR	FAKENEWSNET
Gemma2-9B	24.9	38.8	53.0	55.1	54.6	56.9	**67.1**
LLaMA3.1-8B	**55.0**	51.4	45.1	53.4	37.1	58.0	36.3
LLaMA3-8B	29.2	**53.9**	48.1	54.2	44.3	55.6	48.1
Phi-3-mini	14.5	26.6	36.9	40.0	35.6	25.7	49.1
Phi-4	14.5	25.8	47.9	51.3	45.4	40.3	54.0
GPT-4o	41.6	40.5	**59.6**	54.0	**55.9**	**60.2**	**65.9**
GPT-4o-mini	52.2	45.7	55.9	**56.0**	49.6	54.0	59.1

Experiment Results

LLM direct classification - Few-Shot Top-K

MODEL	RLTD	DECOUR	OPSPAM	CCULT	DEREV2014	LIAR	FAKENEWSNET
LLaMA3.1-8B	46.1	58.1	58.5	58.3	56.0	**60.7**	**67.1**
LLaMA3-8B	29.2	40.7	52.5	58.8	43.8	54.6	**67.3**
Gemma2-9B	55.0	55.9	47.0	49.3	58.0	59.2	**66.7**
Phi-3-mini	56.6	53.4	39.6	38.0	29.3	59.0	56.3
Phi-4	14.5	30.5	42.7	50.2	57.8	42.6	**60.0**
GPT-4o	**84.6**	**66.2**	**72.1**	59.7	**87.8**	**61.7**	58.3
GPT-4o-mini	**70.4**	**61.6**	59.2	**60.9**	**63.6**	57.8	57.8

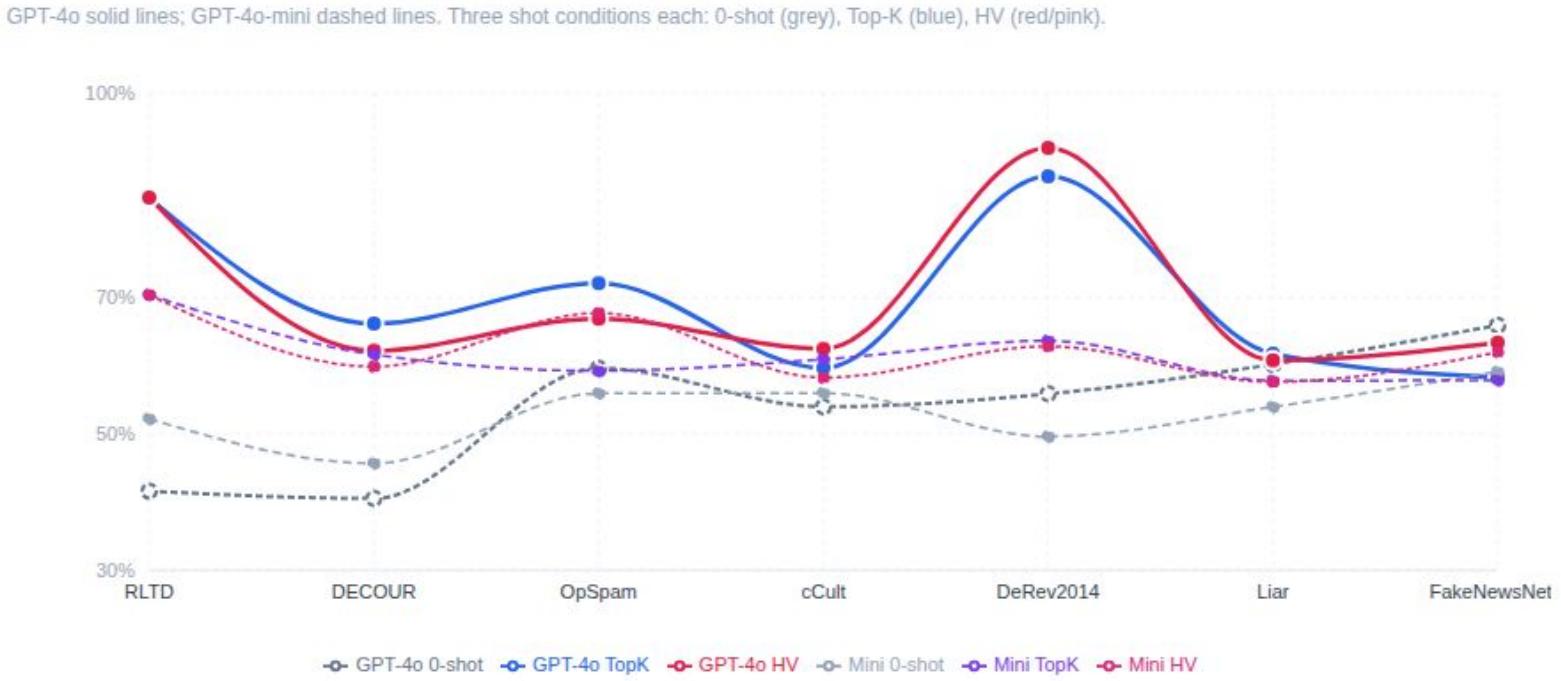
Experiment Results

LLM direct classification - Few-Shot High Variance

MODEL	RLTD	DECOUR	OPSPAM	CCULT	DEREV2014	LIAR	FAKENEWSNET
Gemma2-9B	**70.4**	57.5	51.9	51.8	**71.1**	59.9	**63.0**
LLaMA3.1-8B	52.2	59.8	58.1	57.2	51.1	**60.5**	**64.0**
LLaMA3-8B	41.6	55.4	54.7	**60.6**	41.2	53.2	**65.8**
Phi-3-mini	**65.6**	48.3	40.5	34.2	29.3	58.9	**60.0**
Phi-4	29.2	36.4	42.5	52.7	**69.2**	40.1	**64.0**
GPT-4o	**84.6**	**62.2**	**66.9**	**62.5**	**91.9**	**60.8**	**63.4**
GPT-4o-mini	**70.4**	59.9	**67.7**	58.2	**62.8**	57.6	**62.0**

Experiment Results

LLM Direct classification - Openai on All Datsets



Experiment Results

LLM CoT — Zero-Shot

MODEL	RLTD	DECOUR	OPSPAM	CCULT	DEREV2014	LIAR	FAKENEWSNET
Gemma2-9B	**61.5**	46.5	**63.3**	53.3	**63.9**	50.9	53.3
LLaMA3-8B	**75.9**	51.9	50.0	53.5	58.0	55.9	**63.0**
LLaMA3.1-8B	41.6	53.8	50.6	**53.6**	**64.0**	54.5	57.4
Phi-3-mini	53.9	**55.2**	38.6	47.2	31.1	**60.6**	**63.0**
Phi-4	56.6	39.8	29.6	39.1	31.1	55.0	19.7
GPT-4o	**71.7**	46.7	49.6	51.3	54.6	**64.3**	55.1
GPT-4o-mini	47.6	50.7	47.2	48.5	52.0	**61.8**	55.1

Experiment Results

LLM CoT — Few-Shot Top-K

MODEL	RLTD	DECOUR	OPSPAM	CCULT	DEREV2014	LIAR	FAKENEWSNET
Gemma2-9B	55.0	57.9	52.5	54.7	**72.0**	57.0	56.2
LLaMA3-8B	**62.9**	**63.6**	51.0	56.8	41.3	**62.3**	59.1
LLaMA3.1-8B	24.9	55.7	**54.7**	46.0	59.2	58.7	**60.0**
Phi-3-mini	56.6	51.8	33.2	43.8	29.3	**60.1**	54.1
Phi-4	56.6	44.4	29.6	26.7	31.1	54.1	32.8
GPT-4o	**62.9**	**64.2**	46.0	**59.0**	53.4	**63.3**	30.8
GPT-4o-mini	59.8	48.2	33.3	50.9	41.3	**62.3**	47.4

Experiment Results

LLM CoT — Few-Shot High Variance

MODEL	RLTD	DECOUR	OPSPAM	CCULT	DEREV2014	LIAR	FAKENEWSNET
LLaMA3-8B	**69.2**	54.0	54.6	49.7	41.3	**61.9**	59.1
LLaMA3.1-8B	52.2	56.3	**60.8**	52.0	**64.0**	58.7	**67.1**
Gemma2-9B	**62.9**	58.7	56.0	49.9	**71.1**	56.7	53.3
Phi-3-mini	56.6	50.0	34.9	36.7	29.3	59.9	**60.0**
Phi-4	56.6	45.5	29.6	28.5	31.1	54.0	26.5
GPT-4o	**75.9**	**60.7**	42.4	**59.9**	59.4	**64.3**	38.5
GPT-4o-mini	**69.2**	48.5	31.4	51.8	32.3	**63.5**	**60.6**

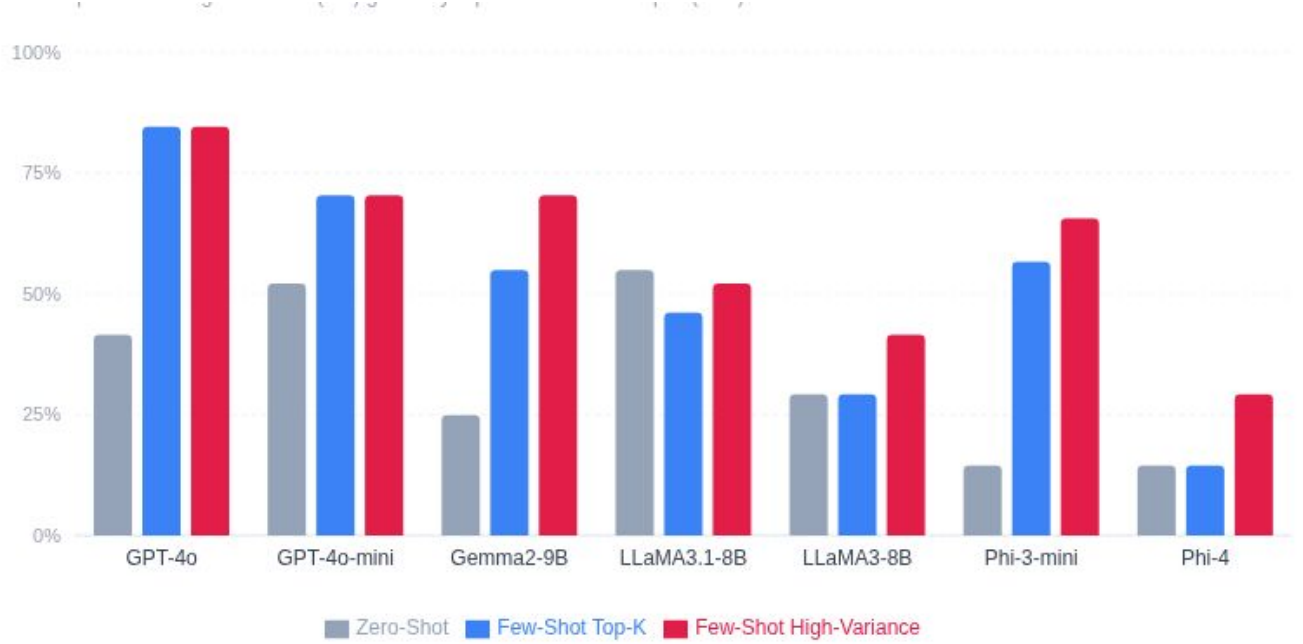
Experiment Results - RLTD

LLM direct classification on RLTD

Model	Zero-Shot	Few-Shot Top-K	Few-Shot High Var	Best Gain
GPT-4o	41.59	**84.62**	**84.62**	+43.03
GPT-4o-mini	52.20	70.38	70.38	+18.18
Gemma2-9B	24.90	54.95	**70.38**	+45.48
LLaMA 3.1-8B	54.95	46.15	52.20	−2.75
Phi-3-mini	14.48	56.64	**65.64**	+51.16
Phi-4	14.48	14.48	29.23	+14.75

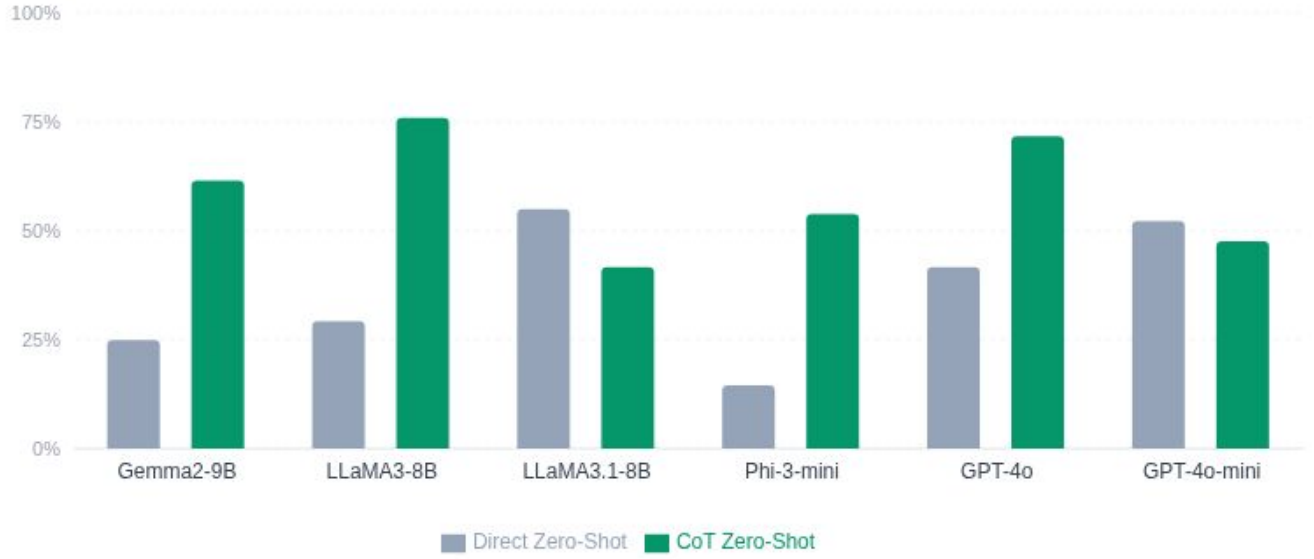
Experiment Results - RLTD

Zero-Shot vs Few-Shot Strategies per Model on RLTD



Experiment Results - RLTD

Direct vs CoT: Zero-Shot Performance on RLTD



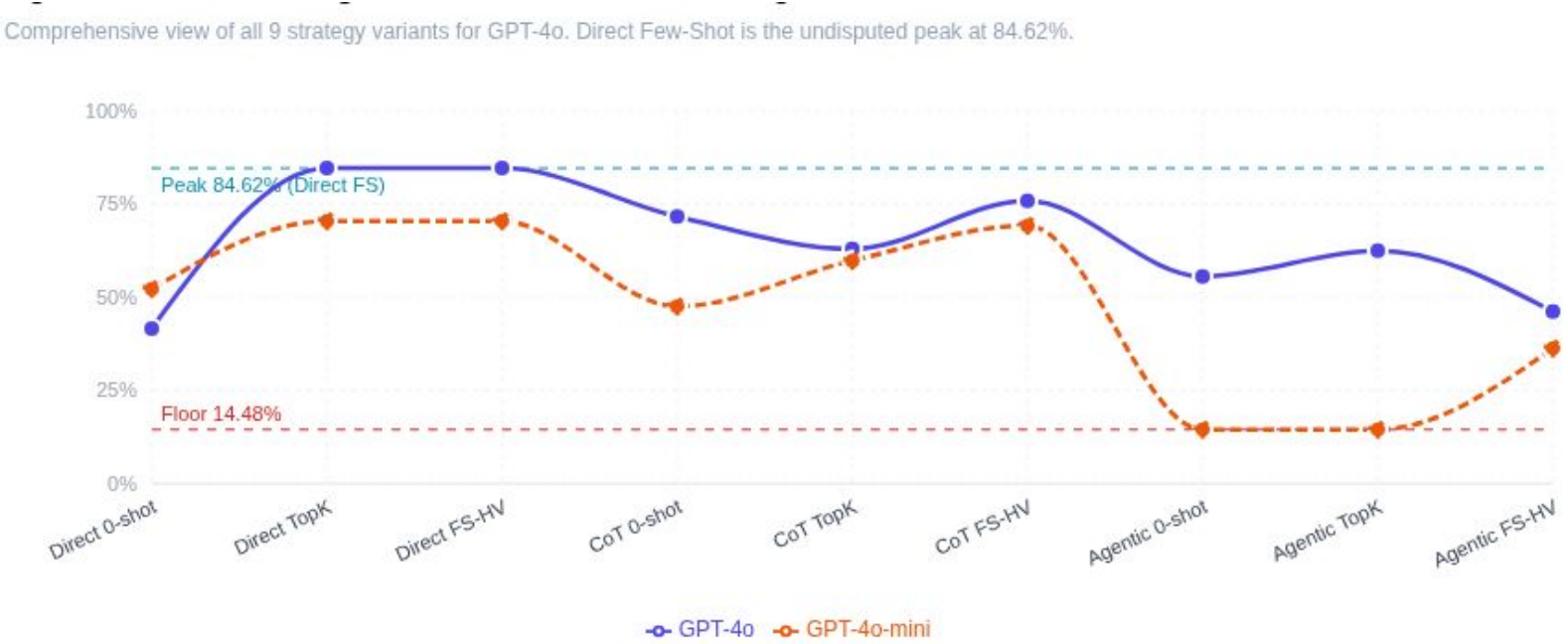
Experiment Results - RLTD

Openai LLMs on RLTD

Strategy	GPT-4o	GPT-4o-mini
Direct Zero-Shot	41.59	52.20
Direct Few-Shot Top-K	84.62	70.38
Direct Few-Shot High Var	84.62	70.38
CoT Zero-Shot	71.65	47.56
CoT Few-Shot Top-K	62.94	59.80
CoT Few-Shot High Var	75.88	69.23
Agentic Zero-Shot	55.58	14.48
Agentic Few-Shot Top-K	62.54	14.48
Agentic Few-Shot High Var	46.15	36.26

Experiment Results - RLTD

Agentic - Openai on RLTD



Experiment 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.
- Agentic Frameworks Do Not Yet Outperform Simpler Prompting.

Papers

- Submitted the paper titled “Semantics of Subterfuge: Benchmarking Legal Deception Detection Against General-domain State-of-the-Art” at ICAIL 2026.

Planned Work

- 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.
- Will focus on Austin police Department (APD)

Data collecting sources for Austin police Department (APD)

➔ Police reports - https://apd-austintx.govqa.us/
➔ Police officer data - https://licenselookup.tcole.texas.gov/PublicLookup/

Data collected

- Information of **1516** Police officers - Personal Information, Education, Service History.
- **1301** police reports.
- **685** Complaints filed against officers.
- **360** Disciplinary & misconduct Cases. (Use of force / misconduct)
- **255** Officers Mentioned in cases.

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] Alberto Alejandro Ceballos Delgado, William Bradley Glisson, Narasimha Shashidhar, J Todd McDonald, George Grispos, and Ryan Benton. 2021. Detecting deception using machine learning. Proceedings of the 54th Hawaii International Conference on System Sciences.

[9] Maria Glenski, Ellyn Ayton, Robin Cosbey, Dustin Arendt, and Svitlana Volkova. 2020. Towards trustworthy deception detection: Benchmarking model robustness across domains, modalities, and languages. In Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM). 1–13.

[10] Roshan R Karwa and Sunil R Gupta. 2022. Automated hybrid Deep Neural Network model for fake news identification and classification in social networks. Journal of Integrated Science and Technology 10, 2 (2022), 110–119.

[11] Alpana A Borse, Gajanan K Kharate, and Namrata G Kharate. 2025. 4HAN: hypergraph-based hierarchical attention network for fake news prediction. In- ternational Journal of Electrical and Computer Engineering (IJECE) 15, 2 (2025), 2202–2210

[12] Katerina Papantoniou, Panagiotis Papadakos, and Dimitris Plexousakis. 2025. Evaluating LLMs on Deceptive Text Across Cultures. In RANLP. 884–893.

[13] Eileen Fitzpatrick and Joan Bachenko. 2012. Building a data collection for de- ception research. In Proceedings of the workshop on computational approaches to deception detection. 31–38.

[14] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. Personality and social psychology bulletin 29, 5 (2003), 665–675.

[15] Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. arXiv preprint arXiv:2310.15515 (2023).

References

[16] Stefan Feuerriegel, Abdurahman Maarouf, Dominik Bär, Dominique Geissler, Jonas Schweisthal, Nicolas Pröllochs, Claire E Robertson, Steve Rathje, Jochen Hartmann, Saif M Mohammad, et al. 2025. Using natural language processing to analyse text data in behavioural science. Nature Reviews Psychology 4, 2 (2025), 96–111.
[17] Tommaso Fornaciari, Massimo Poesio, et al . 2012. DeCour: a corpus of DEceptive statements in Italian COURts.. In LREC. 1585–1590.
[18] Lara Quijano-Sánchez, Federico Liberatore, José Camacho-Collados, and Miguel Camacho-Collados. 2018. Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police. Knowledge-Based Systems 149 (2018), 155–168.
[19] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al . 2018. Cail2018: A large-scale legal dataset for judgment prediction. arXiv preprint arXiv:1807.02478 (2018).
[20] Sam Relins, Daniel Birks, and Charlie Lloyd. 2025. Using Instruction-Tuned Large Language Models to Identify Indicators of Vulnerability in Police Incident Narratives. Journal of Quantitative Criminology (2025), 1–38.
[21] 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).
[22] 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.
[23] 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.
[24] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648 (2017).
[25] 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.
[26] Hugging Face (n.d.). sentence-transformers/all-MiniLM-L6-v2 · Hugging Face. [online] huggingface.co. Available at: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.
[28] Miah, M.M.M., Anika, A., Shi, X. and Huang, R., 2025, July. Hidden in plain sight: Evaluation of the deception detection capabilities of LLMs in multimodal settings. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 31013-31034).