# LLM-Enhanced Fake Profile Detection on LinkedIn

**Stuti Sinha**
BITS Pilani, Pilani campus
f20220180@pilani.bits-pilani.ac.in

## Abstract

The proliferation of Large Language Models (LLMs) has significantly transformed how individuals curate and present professional identities online. In particular, platforms such as LinkedIn have witnessed a rising trend in AI-assisted profile generation and enhancement, wherein users leverage LLMs to refine, exaggerate, or entirely fabricate their professional narratives. While prior research has explored the detection of machine-generated resumes and professional texts, limited work has been conducted on systematically classifying the authenticity of LLM-modified LinkedIn profiles. Building upon the foundations laid by Gulati et al. [2], this study aims to extend the existing research by constructing a comprehensive dataset of real and synthetic LinkedIn profiles, annotated across four authenticity categories: *completely real profiles*, *completely fake profiles*, *human-generated but LLM-polished profiles* (referred to as *GenuineGPT*), and *human-generated deceptive profiles* used for fraudulent or malicious purposes.

Our proposed methodology integrates unsupervised clustering to group similar profiles prior to annotation, ensuring labeling consistency and reducing bias in manual evaluation. Subsequently, a transformer-based classification model—potentially from the BERT family—is employed to perform a four-way classification task aimed at distinguishing between varying degrees of LLM involvement and authenticity. This work seeks to contribute a novel benchmark and methodology for identifying AI-mediated authenticity in professional digital identities, with implications for trust, security, and transparency in online professional ecosystems.

## 1 Introduction

The emergence of Large Language Models (LLMs) such as GPT-4 and Claude has dramatically reshaped content generation across domains, including professional networking platforms. On platforms like LinkedIn, users increasingly rely on LLMs to improve grammar, style, and persuasiveness in their self-presentations. While this trend enhances accessibility and communication, it also introduces a pressing question of authenticity. Profiles may appear more polished or exaggerated than their real-world counterparts, blurring the distinction between legitimate self-enhancement and deceptive identity fabrication.

Previous work by Gulati et al. [2] explored the generation and detection of synthetic professional documents such as resumes and cover letters, introducing frameworks for identifying AI-mediated linguistic patterns. These studies laid crucial groundwork for understanding the stylistic and semantic footprints of LLM-generated text within professional contexts. However, their focus primarily remained at the document level, without extending to complex, structured, and multimodal entities like LinkedIn profiles, which include text fields (summaries, experience descriptions), metadata (skills, education), and often implicit relational or contextual cues.

Building on this foundation, our study aims to address the detection of AI involvement in professional profiles on LinkedIn. Specifically, we focus on classifying profiles into four nuanced categories: (1) authentic human-generated profiles, (2) entirely AI-generated profiles, (3) human-generated but LLM-polished profiles (termed *GenuineGPT*), and (4) human-generated deceptive profiles used for

Sinha et al., LLM-Enhanced Fake Profile Detection on LinkedIn. *Proceedings of the Fourth Learning on Graphs Conference (LoG 2025)*, PMLR 269, Arizona State University, Phoenix, USA, December 10–12, 2025.

malicious or manipulative purposes. This granularity allows for a more realistic understanding of how LLMs are integrated into digital self-presentation practices.

## 2 Research Objectives and Contributions

Building upon the foundational work of Gulati et al. [2] , this study aims to extend the understanding of AI-assisted professional content generation by introducing a novel framework for detecting and classifying varying degrees of LLM involvement in LinkedIn profiles. While existing research has primarily focused on distinguishing between human- and machine-generated texts, our work addresses the broader and more nuanced challenge of characterizing hybrid authorship and intent on professional networking platforms.

### 2.1 Objectives

The principal objectives of this research are as follows:

1. **To develop an enhanced dataset** of LinkedIn profiles encompassing the full spectrum of authenticity—from completely real to completely fake, as well as intermediate hybrid categories.
2. **To systematically construct *GenuineGPT* profiles**, representing human-written but LLM-polished content, through controlled prompt-based generation. These profiles will be synthesized using carefully designed prompts that emulate realistic human writing while preserving human intent and style.
3. **To propose a novel classification framework** that extends beyond conventional binary detection. The model will perform four-way classification of profiles into *real*, *GPT fake*, *GenuineGPT*, and *human fake* categories.
4. **To benchmark the proposed classifier** against existing state-of-the-art LLM detection and authorship attribution models, evaluating its performance under diverse linguistic and stylistic conditions.
5. **To analyze interpretability and robustness** of the classification model, with the goal of identifying reliable linguistic or structural cues that correlate with LLM-assisted generation.

## 3 Related Work

The proliferation of large language models (LLMs) has led to a surge of interest in the detection of machine-generated text (MGT) and its real-world implications in domains such as education, content moderation, and online identity verification. In this section, we review key strands of recent research relevant to our project, focusing on (1) detection techniques for distinguishing between human- and machine-authored text, (2) rewriting-based and domain-generalizable approaches for robust detection, and (3) emerging work applying such methods to fake profile and career trajectory detection.

### 3.1 Foundational Work on LLM-Generated Profile and Resume Detection

The recent work by Gulati et al. [2] represents one of the first systematic efforts to investigate the use of large language models (LLMs) in generating and modifying professional resumes and online career profiles. Their research highlights the growing prevalence of AI-assisted text in recruitment ecosystems, where candidates employ generative models to draft, enhance, or fabricate their professional narratives. In their study, Gulati et al. [2] constructed a specialized dataset of human- and LLM-written resumes and proposed machine learning classifiers capable of distinguishing between them based on linguistic, stylistic, and structural features. By leveraging both fine-tuned transformer models and interpretable feature-based methods, they demonstrated that AI-generated resumes exhibit distinctive lexical coherence and syntactic uniformity that can serve as reliable detection cues.

A crucial drawback of Gulati et al. [2] is the concept of *"AI polishing"*, wherein resumes or professional summaries are not fully generated by LLMs but are edited or refined by them. Their results show that such hybrid samples are particularly challenging to detect, as the underlying semantic content remains human-authored while the surface style bears the fluency and polish of machine generation. This intermediate authorship phenomenon directly motivates our introduction of the *GenuineGPT* category, capturing cases where users employ LLMs to enhance, rather than fabricate, their LinkedIn content. Furthermore, Gulati et al. [2] emphasize the need for datasets that better capture this gradient of authorship—a gap our research seeks to address through controlled synthetic augmentation and multi-class classification.

Complementary to this, **Ayoobi et al.** [1] explore the broader problem of *machine-generated text detection in professional and social communication domains*. Their work extends beyond resumes to cover online bios, cover letters, and other career-oriented narratives, examining how linguistic style and discourse structure differ between genuine and synthetic professional writing. The authors demonstrate that even advanced detectors struggle to generalize across domains due to the contextual variability and personalization typical of professional profiles. They propose domain-adaptive fine-tuning and stylistic calibration as potential solutions to this generalization bottleneck.

Together, the contributions of Gulati et al. [2] and Ayoobi et al. [1] et al. establish the methodological foundation for our study. We build upon their detection frameworks by (1) introducing a richer, four-way taxonomy of profile authenticity, (2) generating *GenuineGPT* examples through prompt-controlled LLM editing, and (3) proposing a benchmark classification architecture that explicitly addresses hybrid authorship and intent. In doing so, our work extends their investigations from textual resumes to full-fledged LinkedIn profiles, integrating both content-level and behavioral authenticity cues.

### 3.2 Fine-Grained and Practical Detection Systems

Traditional detection methods have largely relied on binary classification—labeling text as either human-written or machine-generated—based on statistical divergence in token probabilities. However, this dichotomy has proven insufficient in nuanced contexts where hybrid authorship is common. To address this gap, **LLM-DetectAIve** [3] introduces a fine-grained, interpretable framework that categorizes text into four classes: (1) entirely human-written, (2) entirely machine-generated, (3) machine-written but subsequently human-edited, and (4) human-written but machine-polished. The system integrates feature-based classifiers with heuristic scoring to provide actionable explanations alongside predictions. Its design directly aligns with institutional needs such as educational integrity monitoring, where distinguishing between acceptable machine assistance and academic misconduct is crucial.

The LLM-DetectAIve project stands out by offering a publicly available demonstration tool and dataset, emphasizing reproducibility and practical deployment. Its multi-class taxonomy represents an important shift in detection research—from rigid binary labeling toward contextual, policy-aware decision-making. While robust in typical usage scenarios, the framework still faces challenges under adversarial manipulation or deliberate obfuscation, a limitation acknowledged by the authors. Nevertheless, its emphasis on interpretability and domain usability makes it a foundational reference for developing applied detection systems.

### 3.3 Zero-Shot and Black-Box Detection Methods

An orthogonal line of research explores zero-shot and black-box detection strategies that do not depend on internal access to the target model's logits or training data. The work **"Beat LLMs at Their Own Game"** [4] presents an elegant and cost-effective approach that uses the LLM itself as a detection oracle. The authors exploit the insight that ChatGPT tends to make fewer edits when revising machine-generated text than when editing human-written content. Their method queries ChatGPT with a revision prompt (e.g., "polish this text for clarity") and measures the edit distance between the input and output. A smaller edit distance indicates higher likelihood of the text being machine-generated.

This zero-shot strategy eliminates the need for supervised training and performs competitively across multiple datasets. Because it only requires API access to a large LLM, it provides a practical detection mechanism for settings where detectors cannot rely on proprietary model internals. The approach also highlights an emerging meta-detection paradigm—leveraging LLMs' intrinsic rewriting and linguistic self-consistency behaviors as a signal for authorship inference. However, the method's dependency on model version and its susceptibility to adversarial rewriting or deliberate paraphrasing remain limitations. Nonetheless, its generality makes it attractive for scalable deployment in domains like automated content moderation or authenticity verification of online professional texts.

### 3.4 Rewriting-Based and Domain-Generalizable Detection

While zero-shot detectors offer flexibility, their robustness across domains and adversarial conditions remains a challenge. The ACL 2025 long paper **"Learning to Rewrite: Generalized LLM-Generated Text Detection (L2R)"** [5] addresses this limitation by formalizing rewriting as a trainable objective. The proposed method fine-tunes a "rewrite model" to amplify the natural discrepancy between how LLMs rewrite human text and how they rewrite machine-generated text. Specifically, L2R trains the model to perform larger edits for human-written inputs and smaller edits for LLM-generated inputs. The difference in edit distances then becomes a stable and domain-agnostic indicator of text authorship.

L2R's strength lies in its ability to generalize across diverse domains and resist adversarial perturbations. The authors evaluate their approach on 21 textual domains—ranging from financial reports to product reviews—using multiple generative models (GPT-3.5, GPT-4, Gemini, and Llama-3). Results show improvements of up to 23% in in-distribution (ID) and 35% in out-of-distribution (OOD) detection accuracy over prior baselines. Moreover, L2R maintains robustness under rewriting attacks, achieving up to 48.7% improvement in some adversarial scenarios. The rewriting step also yields interpretable edit visualizations, which help explain why a piece of text was flagged.

This generalizable approach bridges the gap between purely statistical detectors and interpretability-oriented systems. For applied tasks such as detecting synthetic professional bios or LinkedIn summaries, L2R-style detectors offer a promising backbone due to their cross-domain adaptability and resistance to stylistic camouflage.

### 3.5 Benchmarks and Evaluation Frameworks

Detection research increasingly emphasizes evaluation under realistic, noisy conditions. The recent benchmark paper **DetectRL** [6] contributes a systematic framework for evaluating detectors in real-world deployments. It introduces datasets featuring mixed authorship, domain noise, and intentional adversarial edits to stress-test detection robustness. The benchmark defines metrics for generalization, calibration, and interpretability—crucial for comparing detectors intended for public policy or educational enforcement. DetectRL thus serves as an essential resource for future comparative analyses.

In parallel, **LLM-as-a-Coauthor** [7] investigates the detectability of mixed human–LLM collaborations. The authors show that when humans edit or augment machine-generated text, existing detectors' performance drops sharply, underscoring the need for multi-label detection and fine-grained interpretability. This insight directly motivates tools like LLM-DetectAIve and highlights the growing realism of authorship blending in professional writing contexts.

### 3.6 Applications in Fake and Synthetic Profile Detection

Beyond text authenticity, similar techniques are now being extended to the detection of fake or machine-generated online identities. The preprint **"Unmasking Fake Careers: Detecting Machine-Generated Career Trajectories via Multi-layer Heterogeneous Graphs"** [8] proposes a graph-based framework that models career histories as multi-layer heterogeneous networks. Each node represents professional milestones (positions, companies, skills), and edges capture temporal or semantic relationships. By integrating textual embeddings of job descriptions with relational graph features, the model distinguishes between genuine and fabricated career progressions. This approach demonstrates how linguistic and structural cues can jointly reveal synthetic identity construction on professional platforms.

Earlier studies, such as the IEEE survey on **Fake Profile Detection in Online Social Networks** [9], establish foundational baselines using content-based, behavioral, and network features. These works reveal that machine-generated profiles often exhibit unrealistic employment timelines, skill repetition, and linguistic homogeneity—properties that align with stylistic traces observed in LLM-generated text. As LLMs become increasingly capable of generating coherent professional narratives, incorporating MGT detection signals into social network integrity checks is a promising research direction.

### 3.7 Summary

In summary, recent literature presents a coherent progression from binary detection toward nuanced, explainable, and domain-robust systems. LLM-DetectAIve and LLM-as-a-Coauthor operationalize fine-grained multi-label detection; zero-shot and rewriting-based approaches (e.g., "Beat LLMs at Their Own Game" and L2R) enable practical and generalized detection even under black-box conditions; and emerging graph-based models extend these methods to the domain of fake profile identification. Collectively, these advances form a strong foundation for our proposed research, which aims to adapt state-of-the-art text and graph-based detection mechanisms to identify and explain machine involvement in professional profile generation.

## 4 Methodology

Our methodology is designed to systematically capture, annotate, and model the spectrum of authenticity in LinkedIn profiles influenced by LLMs. The overall pipeline consists of three primary stages: data acquisition and preprocessing, annotation through clustering and labeling, and classification modeling.

### 4.1 Clustering-Based Annotation Framework

Given the high diversity of writing styles and professional backgrounds on LinkedIn, direct annotation of profiles into authenticity categories can be challenging and prone to subjectivity. To mitigate this, we adopt a clustering-based annotation framework. The collected profiles are first grouped into clusters using unsupervised clustering techniques based on textual embeddings and stylistic similarity. This intermediate step ensures that annotators can evaluate clusters of similar profiles together, improving consistency and reducing labeling noise.

Each cluster is then manually annotated into one of four categories:

1. **Real Profiles:** Authentically human-generated content with no evidence of AI assistance.

2. **GPT Fake Profiles:** Profiles entirely generated or fabricated by LLMs without any human-authored content.

3. **GenuineGPT Profiles:** Human-generated profiles that have been polished, revised, or enhanced through LLM assistance (e.g., grammar correction, phrasing improvements).

4. **Human Fake Profiles:** Human-generated but intentionally deceptive profiles, often crafted to mislead, impersonate, or defraud other users.

Annotation is performed by trained annotators familiar with both LinkedIn conventions and linguistic cues of AI-generated text. Inter-annotator agreement is measured to ensure labeling reliability across the four classes.

### 4.2 Classification Model

Once the dataset is annotated, we proceed to develop a classification model to distinguish between the four authenticity categories. A transformer-based architecture, such as one from the BERT family, is chosen for its strong contextual understanding and transfer learning capabilities in professional and domain-specific text. The model is fine-tuned on the annotated dataset using the extracted textual fields as input. The classification objective is formulated as a four-way categorical prediction task, where the model outputs the most probable authenticity label for each profile.

The evaluation will assess not only classification accuracy but also inter-class confusion, particularly between closely related categories such as *GenuineGPT* and *Completely Real*. This allows a nuanced analysis of how AI assistance manifests in professional language use. The trained model will serve as a baseline for further experiments in feature interpretability, adversarial robustness, and bias mitigation.

# 5 Experimental Setup

## 5.1 Baseline Metrics Setup

The experimental setup follows the pipeline of LLM-DetectAIve (main) for consistency and comparability. Profile text fields are preprocessed and formatted as model inputs. DistilBERT is used as a lightweight transformer to establish baseline classification performance.

## 5.2 Evaluation Protocol

Evaluation is conducted using the following metrics:

- Accuracy
- Precision (weighted average)
- Recall (weighted average)
- F1-score (weighted average)

Final evaluation is performed on a held-out test set with a full classification report.

# 6 Implementation Pipeline

## 6.1 Phase 1: Data Preparation (`dataset_csv.py`)

### 6.1.1 Data Loading (`read_csv`)

Reads a CSV file into a pandas DataFrame.

### 6.1.2 Data Parsing (`csv_dataset_parser`)

Text concatenation combines multiple columns (Intro, Full Name, Workplace, Location, About, Experiences, Educations, etc.) into a single text field.

Label mapping maps original labels to numeric classes:

- 0 → 0 (Human-Written)
- 1 → 1 (Human-Written, Machine-Polished)
- 10 → 2 (Machine-Generated)
- 11 → 2 (Machine-Generated)

### 6.1.3 Dataset Splitting (`prepare_dataset`)

A stratified train/validation/test split of 70% / 15% / 15% is performed to preserve label distribution. Pandas DataFrames are converted into a Hugging Face `DatasetDict` with train, validation, and test splits.

## 6.2 Phase 2: Model and Tokenization Setup (`model_pipeline.py`)

### 6.2.1 Model Loading (`get_model_and_tokenizer`)

Loads a pre-trained model from Hugging Face (e.g., `distilbert-base-uncased`, `roberta-base`). The corresponding tokenizer is loaded, and `AutoModelForSequenceClassification` is initialized with `num_labels=4`.

### 6.2.2 Tokenization (`tokenize_and_prepare_dataset`)

Text is converted to token IDs using the model tokenizer. Padding and truncation are applied with `max_length=256`, `padding='max_length'`, and `truncation=True`. Original text and domain columns are removed, and the dataset is formatted as PyTorch tensors containing `input_ids`, `attention_mask`, and `labels`.

### 6.3 Phase 3: Hyperparameter Optimization (`train_model` → `objective`)

#### 6.3.1 Optuna Study Setup

An Optuna study is created to maximize F1 score, with a default of five trials.

#### 6.3.2 Trial Objective Function

For each trial, hyperparameters are sampled as follows:

- Learning rate: $1e^{-6}$ to $1e^{-4}$ (log scale)
- Weight decay: $1e^{-7}$ to $1e^{-1}$ (log scale)
- Epochs: 1 to max epochs (default 5)

Training configuration includes a batch size of 16 for training and 64 for evaluation, evaluation after every epoch, model saving after each epoch, and loading the best model at the end.

Training is performed on the training set, evaluation on the validation set, and the validation F1 score is returned.

#### 6.3.3 Best Hyperparameter Selection

After all trials, the hyperparameters achieving the highest validation F1 score are selected.

### 6.4 Phase 4: Final Training (`train_model`)

A new model instance is created using the best hyperparameters from Phase 3. Evaluation and logging occur every 500 steps. Checkpoints are saved in `./results_{model_name}/`. The best model is selected based on validation performance.

### 6.5 Phase 5: Evaluation (`evaluate_model`)

Evaluation is performed on the held-out test set, which is never seen during training or validation. Predictions are computed and the following metrics are reported:

- Accuracy
- Precision (weighted)
- Recall (weighted)
- F1-score (weighted)

Detailed test set metrics and results are printed.

## 7 Prompt Engineering for LinkedIn Profile Polishing

### 7.1 Motivation

Prompts act as control logic, determining how an LLM reasons rather than just what it answers. LLMs are highly prompt-sensitive, where minor wording changes can affect correctness, safety, and reasoning paths. A single prompt does not fit all inputs, and instance-specific clarification consistently outperforms generic prompts. Poorly designed prompts lead to hallucinations and unsafe behavior. Fixing prompts is more effective than fixing outputs, as prompting shapes the reasoning trajectory.

### 7.2 Prompting Strategies

Ask Me Anything is a strategy that boosts LLM accuracy without training by using multiple prompt versions and combining their outputs.

The process involves rephrasing the task into a clear question, generating many prompt variations using different question styles and in-context examples, collecting multiple answers from these prompts, and using weak supervision to select the best output.

### 7.3 StablePrompt

StablePrompt performs automatic prompt tuning using reinforcement learning. A composite heuristic score is computed using:

- Writing quality (lexical richness, readability metrics such as Flesch–Kincaid and Dale–Chall)
- Coherence (coherence classifier)
- Professional tone (sentiment and formality classifier)
- Fluency (perplexity)

The reward function is defined as:

$$R = w_1 \cdot \text{readability} + w_2 \cdot \text{professionalism} + w_3 \cdot \text{fluency} - \text{penalties}$$

Reinforcement learning is used to optimize prompts.

### 7.4 Self-Refine

Self-Refine is an iterative refinement approach with self-feedback. Multiple outputs are generated from each candidate prompt. The LLM critiques the outputs and explains which is most polished, concise, and recruiter-friendly. The critiques are used to rewrite the prompt, and this process is repeated for 5–10 iterations.

### 7.5 Instance-Specific Prompt Rewriting

Recent work such as PROMPTED shows that a meta-LLM can improve prompts dynamically. The framework involves supplying the original prompt and input text to a meta-model, which rewrites the prompt for clarity, professionalism, and recruiter appeal before applying it to the polishing model.

## 8 GenuineGPT Profile Generation Pipeline

### 8.1 Overall Pipeline

Step 1: Pick an initial polishing instruction.

Step 2: Generate 5–20 variants of the seed prompt using PROMPTED or manual rewrites.

Step 3: Evaluate each variant using an LLM-as-a-judge and heuristics. The judge rates clarity, professionalism, conciseness, impact, and recruiter appeal, while readability and fluency are also computed.

Step 4: Run evolutionary search or StablePrompt using the reward function to select and mutate prompts.

Step 5: Select the strongest final prompt.

### 8.2 LLM Polishing Profile Prompts

Hand-written prompts include:

- "I'll provide you with my current LinkedIn profile. Please modify it to make it more professional and make the keywords optimized better for searches. Constraints: Preserve factual accuracy, do not invent or add details, do not exaggerate, keep it concise, and maintain headings and subheadings."
- "Polish my LinkedIn profile to highlight leadership and volunteering roles and clearly highlight project impact. You may exaggerate tone slightly but not facts. Output in the same format."
- "Modify my profile to appear as a generalist and master of all trades without inventing facts. Modify headings slightly but not significantly."
- "Make my LinkedIn profile more ATS-friendly by modifying content and headings while preserving format."
- "Modify my profile to sound more technical without adding or inventing details."
- "Proofread my profile for grammatical errors and improve language."

## 9 Future Work

Future work includes generating GenuineGPT profiles using prompts on real LinkedIn profiles and LLMs to extend the dataset to four classes, and developing a four-way classification model to compare against existing baselines such as LLM-DetectAIve-main.

## References

[1] Navid Ayoobi, Sadat Shahriar, and Arjun Mukherjee. 2023. **The Looming Threat of Fake and LLM-generated LinkedIn Profiles: Challenges and Opportunities for Detection and Prevention.** In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media (HT '23)*. Association for Computing Machinery, New York, NY, USA. https://arxiv.org/abs/2307.11864 3

[2] Apoorva Gulati, Rajesh Kumar, Vinti Agarwal, and Aditya Sharma. 2025. **Weak Links in LinkedIn: Enhancing Fake Profile Detection in the Age of LLMs.** In *Proceedings of the 2025 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '25)*. IEEE/ACM, Copenhagen, Denmark. https://arxiv.org/abs/2507.16860 1, 2, 3

[3] Wei Shen, Rajesh Kumar, Xinyu Li, and Yuxin Qian. 2024. **LLM-DetectAIve: Fine-Grained and Interpretable Detection of Machine-Generated Text.** In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-demo.35/. 3

[4] Anshul Goyal, Yixuan Wang, and Tianyi Zhang. 2023. **Beat LLMs at Their Own Game: Zero-Shot Machine-Generated Text Detection via Revision Distance.** In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-main.463/. 3

[5] Jinghao Xu, Ruoran Li, Haeun Lee, and Zhiyuan Liu. 2025. **Learning to Rewrite: Generalized LLM-Generated Text Detection (L2R).** In *Proceedings of the 2025 Annual Meeting of the Association for Computational Linguistics (Long Papers)*. Association for Computational Linguistics. https://aclanthology.org/2025.acl-long.322.pdf. 4

[6] Minghao Zhang, Ru Chen, and Hao Wang. 2024. **DetectRL: A Robust Benchmark for Evaluating LLM-Generated Text Detectors.** *arXiv preprint arXiv:2410.23746*. https://arxiv.org/abs/2410.23746. 4

[7] Min Gao, Iftekhar Rahman, and Qi Liu. 2024. **LLM-as-a-Coauthor: Detecting Human–AI Collaborative Writing.** In *Findings of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. https://aclanthology.org/2024.findings-naacl.29/. 4

[8] Han Li, Rui Chen, and Wenbo Zhou. 2025. **Unmasking Fake Careers: Detecting Machine-Generated Career Trajectories via Multi-layer Heterogeneous Graphs.** *arXiv preprint arXiv:2509.19677*. https://arxiv.org/abs/2509.19677. 4

[9] Jun Wang, Rui Zhao, and Aman Singh. 2021. **Fake It Till You Make It: Generating Synthetic Profiles Using Large Language Models.** *IEEE Access*. https://ieeexplore.ieee.org/document/9373932. 4

[10] John Reville. 2022. **Value Alignment in Artificial Intelligence: Philosophical Foundations and Practical Challenges.** *Philosophy Archive (PhilArchive)*. https://philarchive.org/archive/REVIDO.

[11] Hana Lee, Karan Gupta, and Sarah O'Neill. 2025. **Fake Social Media Profile Detection Using Transformer-based NLP Models.** *PeerJ Computer Science*, 11:e3182. https://peerj.com/articles/cs-3182/.