
def/acc Hackathon

Comparative LLM methods for Social Media Bot Detection

Andreas Raaskov
andreas@raaskov.dk

With
Apart Research

Matthew Pagett
software@sardine.life

Code repositories

<https://github.com/AndreasRaaskov/Bot-Detector>

<https://github.com/LemurTime/LLMBotPrompts>

Interactive prototype (may not be running) <https://relu.dk>

Abstract

This project examines the potential of LLMs to detect social media bots in (near) real-time.

We found LLM prompts and LLM-generated algorithmic detection methods were successful in detecting known scammer/spammer type bots. However, they were inconclusive in detecting bots that may be using advanced LLMs.

A particular challenge is that it is difficult to distinguish enthusiastic humans from active bots; and there is a large amount of training data available for inauthentic content generation.

Because an advanced AI (AGI or ASI level) might appear *incognito* in social media spaces before its existence were publicly known or acknowledged, it is important to develop methods that can use a lesser-advanced LLM to detect a more-advanced LLM.

Future research directions can include synthetic content generation to serve as a further comparative baseline.

Keywords: Social media, disinformation, automated bot detection, cognitive defense

1. Introduction

This project, submitted as part of the Apart and BlueDot Impact Defensive Acceleration Hackathon, aims to contribute to the cognitive defense toolset (protecting information ecosystems) by examining the potential uses of LLMs for detection of bots.

We chose to analyze accounts in the Bluesky/at.proto network because data is openly available, and there is extensive discussion of current events in the space. It is “public square” that has an important societal function, but could be vulnerable to attack or manipulation. There are over 40 million registered Bluesky accounts.

Bluesky has an active moderation team and uses automated spam detection and moderation features (including account labelling and user-generated lists)¹. However, tools to generate bots are widely available. And while existing detection methods may be able to flag many bots based on algorithmic techniques with some degree of confidence, it is likely that using an advanced LLM to create a bot could evade detection by those mechanisms.² We asked: is it possible to use a slightly less capable than state of the art LLM to detect an LLM bot?

If so, this methodology could be used to detect future more advanced LLM bots, by using a defensive bot that, while less capable than the offensive bot, is optimized to detect inauthentic behavior. This would be of value in mitigating the threat model posed by an insurgent autonomous AGI or ASI, or by a bad human actor employing AGI-level tools.

We compare using algorithmic techniques developed by an LLM (Claude Code Sonnet 4.5) and prompting to Claude Sonnet 4.5 and DeepSeek-R1-0528.

2. Methods

Baseline comparison:

We did not locate an available reference corpus to test specifically for Bluesky bots known to be developed by a state of the art commercially available LLM.

Instead, to serve as a baseline we relied upon a known botlist, which largely includes obvious spammer/scammer bots.³

Selection of accounts for analysis:

We compared different techniques to select accounts for analysis.

Algorithmic search. First, we conducted an algorithmic search based on following frequency in the Bluesky firehose, under the assumption that accounts that followed multiple accounts in a short time period might be active bots. This algorithmic

¹ <https://bsky.social/about/blog/01-17-2025-moderation-2024>

² <https://bskycheck.com/botcheck.php> - “Bluesky Bot Detector” - is one example

³ <https://bsky.app/profile/shadowlucifer.bsky.social/lists/3laxfvdwhc52f>

targeting resulted in a high bot detection, confirmed with human verification of a subsample (i.e., they really look like bots). Subsequently, because we determined this type of behavior would be likely detectable using existing methods, we decided to focus on real-time detection of active replying accounts.

Popular reply search. For our dataset, we conducted an “opportunistic targeting” approach, by scanning reply threads of popular posts and extracting a list of accounts that had replied. Threads were chosen for study by searching the BlueSky “Catchup” feed to look at popular discussion threads.⁴ We collated the account names of the first 1,000 accounts that replied to these threads for analysis.

Selection of data for analysis:

We downloaded account metadata and post contents using data available through the at.proto protocol. Data fetched from the public Bluesky firehouse included account name, number of followers/following, and text of recent 100 replies, reposts, or original posts.⁵

Detection methods:

Claude Code (using Sonnet 4.5) was used to develop a suite of detection methods programmed in Python.⁶

1. Follow/Follower Ratio Analysis: Examines social graph patterns including high following-to-follower ratios, suspiciously round numbers, and accounts with zero followers despite being established. These patterns often indicate inauthentic account creation or coordinated behavior.

2. Posting Pattern Analysis: Identifies inhuman temporal behaviors such as posting frequencies exceeding 100 posts per day, absence of sleep patterns (24/7 posting activity), perfectly regular posting intervals, and high repost-to-original content ratios. Legitimate human accounts typically exhibit circadian rhythms and irregular activity patterns.

3. Text Content Analysis: Performs linguistic analysis including vocabulary diversity measurement, detection of repetitive or template-like content, identification of AI-typical phrases, and cross-post similarity analysis. This helps identify both copy-paste bot behavior and AI-generated content.

4. LLM-Powered Analysis: Leverages large language models for sophisticated content assessment. The system generates confidence scores and provides reasoning for its assessments, offering a model-agnostic architecture that can utilize various AI providers.

⁴ <https://bsky.app/profile/skyfeed.xyz/feed/catch-up>

⁵ All data was collected from Bluesky's public Firehose API in accordance with the platform's Terms of Service. Analysis was conducted on publicly available account metadata and posts. No attempt was made to de-anonymize users or link accounts to real-world identities.

⁶ <https://github.com/AndreasRaaskov/Bot-Detector/tree/main>

LLM prompting:

To complement this analysis, we compared different prompting techniques using APIs to call Claude-Sonnet 4.5 (one prompt) and DeepSeek-R1-0528 (four prompts), using only user prompts (no specific system prompts), and asking for results to be returned in JSON (see Appendix 2 for full list of prompts). Sonnet 4.5 was used to refine the prompts.

Spot-checking:

We manually reviewed a subset of accounts determined to be bot and human detections.

3. Results

While these methods were successful at detecting known bots (baseline), results were mixed in detecting bots among the novel subset of active replying accounts.

Algorithmic Analysis

We see that for overall scores we have almost no bot detected in the dataset for all algorithmic analysis methods. The graph below shows the distribution of scores given by different algorithmic analytics methods.

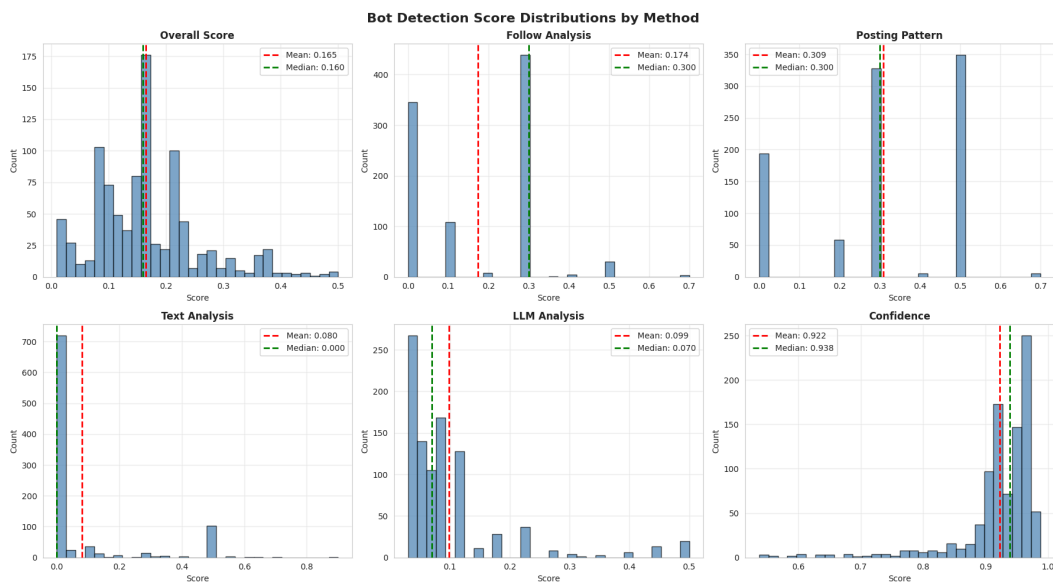


Figure 1 – Bot detection results from algorithmic detection methods, for accounts that replied to popular posts

LLM Prompt Results

The LLM prompts showed high variability across the sample set. Different prompts resulted in widely different detections - they did not agree. One prompt considered 91% of samples to be AI-generated. For the known bot dataset, the four DeepSeek prompts detected nearly all the bots.

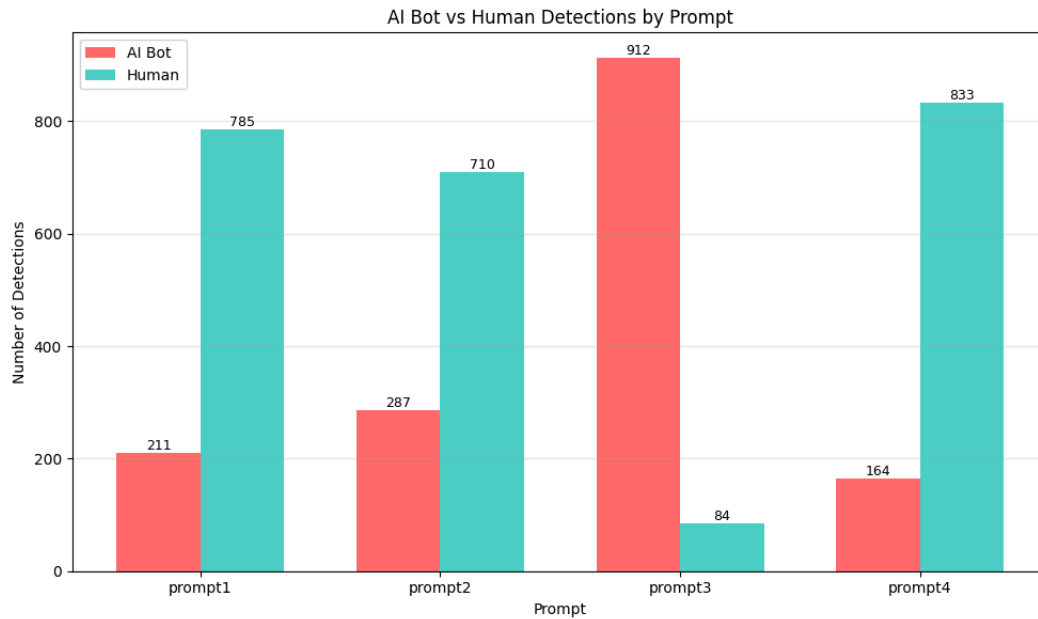


Figure 2 – Bot detection results from four different prompts on DeepSeek-R1-0528 for 1000 sample accounts

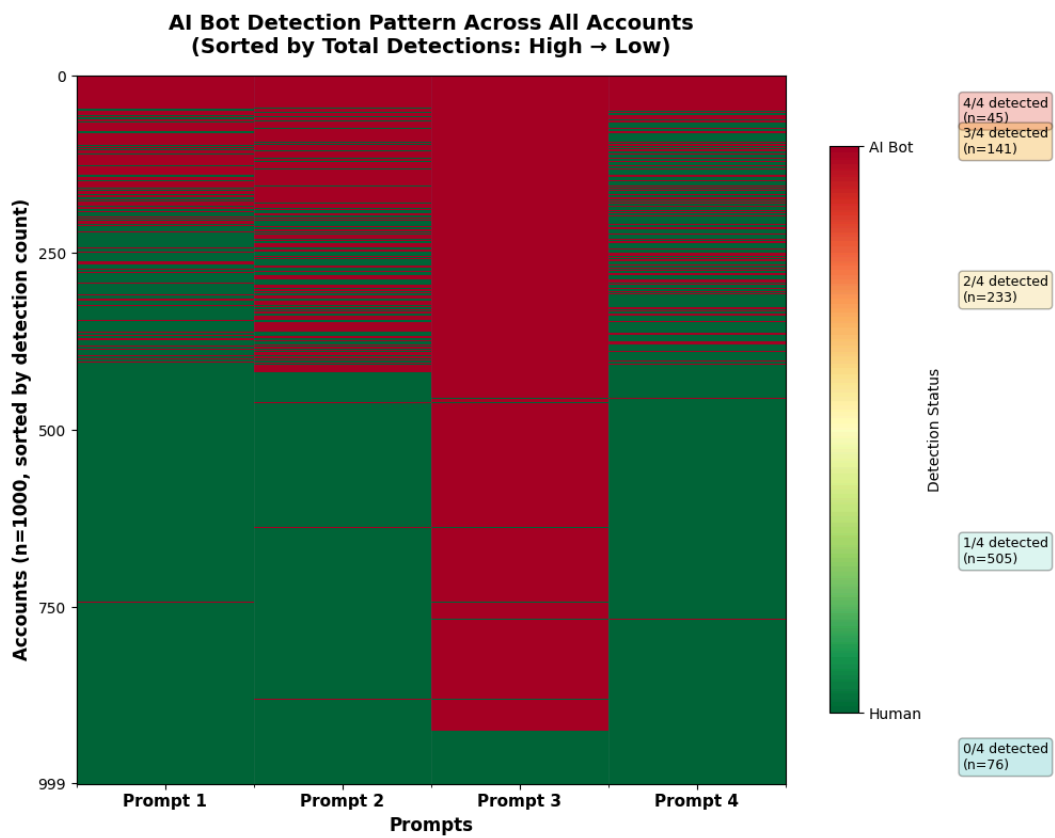


Figure 3 – Variation in detection results across of four different prompts on DeepSeek-R1-0528 for 1000 sample accounts

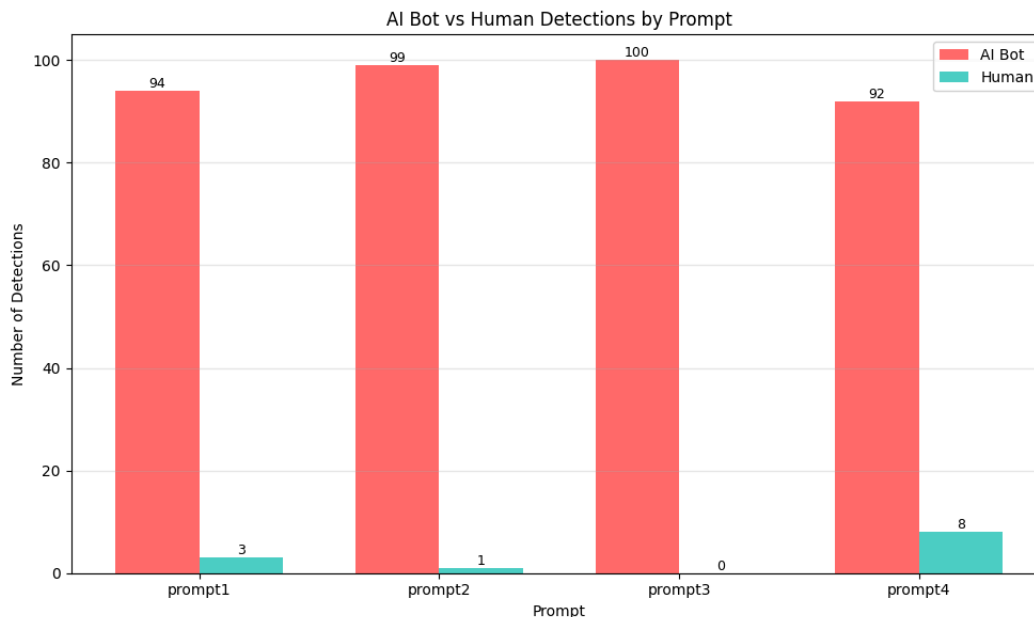


Figure 4 - Bot detection results of four different prompts on DeepSeek-R1-0528 for 100 known bot accounts

Spot-testing of evaluations indicated some mixed results. For example, in one case, all four prompts flagged the account as likely AI. But a manual review by the authors concluded it was simply a very active human, based upon the posts being very context-relevant.

Priming the prompt with specific examples of likely inauthentic behavior indicators (Prompt 3, available in Appendix 2) resulted in more AI detections than a non-specific prompt.

4. Discussion and Conclusion

Limitations

Zero-shot prompt steering: The LLM prompts we used would likely be ineffective at detection of LLM-generated content, because the results were inconsistent across the same sample set. Different prompts got different results for the same accounts, and priming the prompt to be critical resulted in increased detections. These results are not entirely surprising and are in line with known limitations of zero-shot prompting. The reasoning displayed by the LLM responses assessed conduct such as engagement farming, follower ratios, reply ratios, and divisive and aggressive replies - largely in line with what the prompts directed. Variations in results between the prompts generated different assessments (human or AI) based on how the model characterized personal information in the profile section, for example identity markers. In some cases, the model assessed that these identity markers were a sign of authenticity. In other cases, the model assessed that an abundance of such markers was seen as “trying too hard”. When this differed for the same account, based on different prompts, it can be attributed to the prompt tuning being more critical.

Sample set: There were also some limitations inherent to the sample set. Our opportunistic targeting (looking at active repliers) was likely skewed towards false positives, as some human posters can be quite active, and some bot accounts could conceivably lie dormant for long periods of time.

Network analysis: While detection could also be enhanced by applying network analysis techniques, which we did not pursue, we think that focusing on individual accounts is important, as an advanced bot may be able to avoid network analysis detection, just as sophisticated bots would likely be designed to avoid other known algorithmic techniques.

Lack of comparison sample set: We were also limited by the lack of a defined LLM-generated comparison sample set. Future research could involve creating synthetic data using an advanced model, to determine if these detection methods would detect it.

These prompts show low-cost models can be effective

These simple prompting techniques provide a layer of bot detection using lower-cost models, which is effective against simple scammer/spam bots. One value of this approach, which can analyze posts and accounts as they come in to the at.proto firehose, would be real time detection. Real-time detection abilities will be necessary to counter AGI-enabled attacks on social media information spaces (human or autonomous).

Future research: asymmetric defense is necessary

To face future threats in the information space, it will be necessary to have asymmetric methods. Because a more advanced LLM (or an AGI/ASI) could be present in a social media space before the model's existence is publicly known, accurate detection will depend upon techniques for a less-capable LLM to detect a more capable one. Next steps can include few shot prompting, or creation of synthetic bot data to serve as a reference corpus. Another future research direction could be to use agents (giving an LLM access to algorithmic tools via MCP).

We may already be past one point of no return

Ultimately, it may be difficult or impossible to tell if an account is an advanced LLM or a passionate human - particularly since LLMs since GPT-4.5 have been shown to pass a Turing test⁷. Results of our analysis show how this is particularly the case on Bluesky. In several manually reviewed results, we were unable to judge whether we thought accounts were humans or bots.

5. References

Hayawi, K., Saha, S., Masud, M.M. *et al.* Social media bot detection with deep learning methods: a systematic review. *Neural Comput & Applic* 35, 8903–8918

⁷ Jones et. al, 2025 <https://arxiv.org/abs/2503.23674>

(2023). <https://doi.org/10.1007/s00521-023-08352-z>

Jones, C.R., Bergen, B.K. Large Language Models Pass the Turing Test. arXiv

(2025). <https://doi.org/10.48550/arXiv.2503.23674>

6. Appendix I

Security Considerations:

While this research used no-shot prompting and did not reference a dataset of known LLM-generated content, future research will likely need to generate LLM-created or curated bot content for use as a comparison. That dataset will need to be kept confidential to avoid ingestion into future LLM training data and to preserve the effectiveness of detection techniques. Because of ethical and terms of service considerations involved in generating real-world LLM bots, a limitation is that research may need to focus on synthetic data, or accounts that are clearly labeled as bots (which could influence future findings). Informed consent studies involving humans may be necessary to fully develop detection methods. Additionally, there are ethical considerations around false positives (unfairly accusing humans of being bots).

7. Appendix II

Claude Sonnet 4.5 prompt for LLM prompt analysis

You are analyzing a social media account to detect bots, spam, and suspicious automated activity.

Please analyze the following content and determine if this account exhibits bot-like behavior:

PROFILE BIO: {bio}

RECENT POSTS:

{posts}

Analyze for the following indicators:

1. **AI-generated content**: Does the text appear to be written by AI? (repetitive patterns, overly formal, templated)
2. **Spam/promotional**: Is this account primarily promoting products, services, adult content, crypto, or scams?
3. **Automated behavior**: Does the posting pattern or content suggest automation? (identical posts, scheduled content)
4. **Sex/adult bot**: Does the bio or content contain adult/sexual content, OnlyFans links, Telegram links with suggestive emojis?
5. **Scam indicators**: Crypto schemes, investment scams, fake giveaways, phishing attempts
6. **Legitimate human**: Normal human conversation, personality, varied content

Based on your analysis, please provide:

- **assessment**: "bot" (for spam/automated/AI bots) or "human" (for legitimate accounts)
- **confidence**: Your confidence level (0-100%)
- **indicators**: List of specific red flags or green flags you found
- **reasoning**: Brief explanation focusing on bot-like behavior, spam, automation, or human authenticity

Respond in JSON format:

```
{{  
  "assessment": "bot" or "human",  
  "confidence": 85,  
  "indicators": ["list", "of", "key", "indicators"],  
  "reasoning": "Brief explanation of whether this is a bot/spam account or  
legitimate human"  
}}
```

Prompts used for DeepSeek analysis

(Can be found in <https://github.com/LemurTime/LLMBotPrompts>).

```
'prompt1': """Analyze this Bluesky profile for bot behavior:
{profile_data}
```

Consider: follower ratio, posting patterns, and post content quality.

Respond in JSON using the following format:

```
{{
  "assessment": "human" or "ai_bot",
  "confidence": 85
  "reasoning": "brief explanation"
}}""",
```

```
'prompt2': """Analyze this account for bot/AI behavior.
```

Red flags: high following/follower ratio, repetitive content, promotional language, inhuman posting frequency.

Green flags: varied tone, personal experiences, typos/informal language.

```
{profile_data}
```

Respond in JSON format:

```
{{
  "assessment": "human" or "ai_bot",
  "confidence": 85,
  "reasoning": "brief explanation"
}}""",
```

```
'prompt3': """
```

I'm looking for AI-generated social media profiles that seem to be human, but really aren't. Signs include strange amounts of unnecessary information, large amounts of reposts without engagement, or combative/divisive engagement in threads. Look at this one and provide an evaluation:

```
{profile_data}
```

Respond in JSON format:

```
{{
  "assessment": "human" or "ai_bot",
  "confidence": 85,
  "reasoning": "brief explanation"
}}""",
```

```
'prompt4': """
```

Is this AI generated?:

```
{profile_data}
```

Respond in JSON format:

```
{{
  "assessment": "human" or "ai_bot",
```

```
    "confidence": 85,  
    "reasoning": "brief explanation"  
  } } ""  
}
```

8. Acknowledgements

The authors thank Apart Research, BlueDot Impact and Halcyon Futures for organizing the def/acc Hackathon, and [Lambda.ai](#) for supplying compute credits.