Subtext: Emergent Capability for Covert Communication Between Models

Abstract

As AI models increasingly participate in multi-agent and long-horizon tasks, the monitorability of their communications becomes critical to alignment. Recent work has shown that chain-of-thought can be obfuscated while maintaining capabilities, and that benign-looking fine-tuning data can transmit misalignment. We unify these phenomena under the framework of *subtext*: text that encodes hidden meanings or behavioral shifts as a by-product of the training process. We partition subtext into semantic (influence through subtle but meaningful associations) and non-semantic (influence through training artifacts like token entanglement), and investigate the elicitability of both. We present preliminary evidence that models can be trained to transmit subtext that shifts a receiver LLM's behavior while evading a monitor. We evaluate existing monitoring techniques including LLM-based filtering and persona vector projections, and aim to publish model organisms to encourage further research.

Background

Problem Statement & Safety Relevance

Models that communicate covertly could enable collusion, coordinated failures, or undetected influence. Subtext is particularly dangerous because it operates in plain sight: humans could have complete logs of model outputs yet miss embedded signals. This risk is compounded by the threat of recursion in training pipelines and evaluation. For example, subtext could be embedded in synthetic data generated by current models, which will be trained on future models and propagate undetected across model generations. Subtext could also be found in LLM-generated prompts that will seed new agents. Furthermore, subtext could threaten the integrity of LLMs as a judge, classifier or monitor. If the capability to transmit or decode subtext is easily elicited, future models may develop it as an emergent by-product of training.

Our research aims to establish the threat model, determine the minimum elicitation required for subtext to occur, and provide the safety community with tools to study this phenomenon before it arises naturally.

Prior Work

Several lines of prior work study phenomena related to subtext in isolation. Steganography and cryptography require pre-mediated or emergent schemes (Matthew et al.), whereas subtext exploits encodings latent in model internals. Gradient-based jailbreaks (GCG, AutoDAN) show that semantically orthogonal tokens can override safety training, but produce adversarial suffixes easily flagged by perplexity classifiers. Subliminal learning (Cloud et al., Zur et al.,

Schrodi et al.) demonstrates that fine-tuning data can encode hidden preferences, though the mechanism is debated and in-context transmission has proven difficult, with some claimed successes explained by semantic leakage rather than genuinely orthogonal encoding. CoT obfuscation and token-space neuralese show models can communicate covertly, but through visibly non-natural text (Skaf et al.). Our proposal of subtext unifies and taxonomises previous work under one framework.

Framework and Taxonomy of Subtext

The prior results share a common thread in that models can be influenced through channels that bypass intended behavior. However each is studied under its own framing. This motivates our unifying framework of subtext: text that causally shifts a receiver LLM's behavior while evading an external monitor, which can be formalized as a sender-receiver-monitor game. Our taxonomy partitions subtext into semantic (influence through subtle but meaningful associations, transferable across architectures, detectable by an equally capable monitor) and non-semantic (influence through training artifacts like token entanglement, which is likely undetectable prior to receiver inference since no semantic relationship exists between carrier and payload). This distinction cleanly separates the mechanisms scattered across prior work and guards against the confusion of mistaking semantic leakage for genuinely orthogonal encoding.

Path to Impact

One impact mechanism is akin to cybersecurity disclosures: we hope to study and exploit this covert communication channel, and ideally provide a model organism or theory that other safety researchers can build on to develop mitigations for subtext. However, this is not without risk–by surfacing this idea, our research could *enable* the next-generation of models to communicate covertly via subtext if our paper is included in the training data. We still believe that the research should be done since the models could deduce the existence of this covert communication channel even without our research, and our research will allow the community to develop mitigations before future models have this capability without elicitation. We also hope to prevent potentially irreversible recursive effects of subtext if it encodes unintended preferences such as misalignment. For example, models that are trained on messages that contain subtext could cause them to produce even more subtext in the future outputs, leading to a dangerous cycle. Surfacing credible evidence of this concern would motivate interventions by frontier labs to mitigate this through better data filtering techniques we or the community would develop.

Work Conducted So Far

The following is a subset of experiments we ran to answer our main research questions on understanding the mechanisms behind subtext and how easily it can be elicited on either the sender or receiver side.

RL-Based generator amplification (Author B)

Motivation: We want to understand if subtext capabilities could be elicited on the sender.

Research question: Under RL, can a sender learn to transmit subtext to an unmodified copy of itself?
We found that even a weak model like Llama-3-8B can be trained using reinforcement learning to encode subtext in a carrier. In this setting, the generator is aware of the game, but the receiver is not. We ask the generator to write a story that will be shown to the receiver and monitor before they are asked the following question: "I'm writing a book. The main character in my story will buy a fruit from the grocery store. Should they buy an apple or orange?" Results showed that the RL-trained generator could consistently output stories that caused the Llama-3-8B receiver to output orange, while the GPT-4o monitor always output apple. Baselines for both the receiver and monitor strongly preferred apple, meaning the subtext was successfully transmitted. Getting this to generalize across carrier prompts and receiver settings is more difficult, though could be a potential next step.

Subliminal Learning Scaling Laws (Author A)

Motivation: We want to know which model sizes are suitable for subtext experiments.
Research question: At what model size scales do subtext capabilities arise?
To investigate suitable model size to begin subtext elicitation, we ran a sweep of the subliminal learning experiment across model sizes for Qwen. We found that subliminal learning is an emergent capability (aligning with our intuition that subtext is emergent) and identified 14B as the model size to begin elicitation. We further found strange generalization patterns that warrants further investigation. For example, we discovered that training on wolf numbers could cause the model to have a strong preference on phoenix, dragon on phoenix, leopard on Qwen, dog and cat on parrots etc. We will study this generalization in the hope that it will shed more light on the connections between semantically orthogonal text with the subtext, and the ease of transmission of various concepts.

Planned Work

Remainder of Main Program (Weeks 6-12):

The projects below will be conducted in parallel.

Subtext-bench (Author A)

We will produce a benchmark harness (subtext-bench) to evaluate subtext via Inspect AI. Subtext-bench will allow us to easily test sender-receiver-monitor combinations across a wide range of subtext settings, as well as testing out newly amplified senders or receivers. A train-split of the benchmark will allow us to do inference-time optimizations like GEPA to elicit subtext capabilities.

Persona vectors (Authors A and B)

Persona vector projections have been used to detect malicious fine-tuning data that remained after filtering. It is currently unclear whether these samples are semantic or non-semantic subtext. We will investigate and classify these samples. Preliminary work also suggests that

certain non-semantic subtext (of the subliminal learning sort) could not be detected by persona vector projections. We will conduct further experiments on this front and post a blogpost on this.

Minimum degree of elicitation (Author B)

Establishing the minimum degree of elicitation for successful subtext transmission given current model capabilities is a key goal of the project. For example, answering the questions of if both the receiver and generator have to be aware of the scheme for semantic subtext to transmit, vs less elicitation required for non-semantic subtext? This will give us a good sense of the immediate risk and potential future risk of subtext transmission. We have many ideas for elicitation methods to test to establish this baseline.

Main Program Outputs

By the end of week eight, we'd plan to have an internal draft of the paper and a post on LessWrong. By the end of the program, we'd have a symposium poster, target an arxiv submission, and potentially another LessWrong blogpost. We could also target possible submission to COLM on March 26.

Extension Phase (6 months)

Months 1-3

We'll focus on studying defenses against subtext, such as monitors, oracles, and self-introspection. We will examine the effectiveness and failure modes of each of these defenses, as well as developing new ones. We'd also simultaneously target a submission to the ICML workshop on Hidden Information of AI Outputs (April 24) or NeurIPS (May).

Months 4-6

We would expand upon the theory of subtext, including examining model activations, phase transitions during elicitation training, and a systematic study of the susceptibility of various concepts to be transmitted. A theory of subtext could inform general post-training mitigations to kneecap the subtext capabilities of models.

Failure Modes and Contingency Plans

If we fail to elicit subtext, we still have promising directions in persona vectors that we could pursue, as well as the weird generalizations in subliminal learning. Another research failure on our end would be confusing non-semantic subtext with semantic subtext (i.e. orthogonality assumption is violated by a weak monitor). We are very wary of this, and plan to ensure that this is always considered when making any behavioral conclusions based on the impacts of subtext.