EtymoStega: Etymological Steganography in Language Models

Spyridon Manolidis [1]

Abstract

We consider the current landscape in AI red-teaming and assess how fine-tuning plays an increasingly critical role. We provide an overview of the etymological origins of the English lexicon, and outline how existing literature and data have motivated us to create an encoding scheme based on etymology. We then formally define EtymoStega: a novel etymological steganographic scheme, and a simpler scheme inspired by EtymoStega where the etymological class of the first verb in a response carries a binary bit, which we then train two 7B-8B open-weight models on to empirically demonstrate the feasibility of such schemes in the research landscape of today. The fine-tuned models adopt our modified steganographic scheme with over 96% accuracy on one model and 90% accuracy on another model, after a conservative adjustment for evaluation database contamination. We then evaluate the trained models, highlighting both the advantages and the disadvantages of our results. Lastly, we provide a comprehensive review of the limitations, assumptions, and uncertainties of our research. The source code and research artifacts are also provided.

1. Introduction

Recent years have seen the increased adoption of Artificial Intelligence (AI) and in particular the use of Large Language Models (LLMs), by both the general public in everyday life, and by large corporations in the enterprise setting. Many such models, including state-of-the-art (SoTA) models, are trained on significant amounts of web-scraped data (OpenAI, 2026d). This means that adversaries can extract information that an LLM has memorised (Nasr et al., 2023), some of which could be dangerous.

Additionally, many frontier labs – namely Anthropic and OpenAI[2] – provide fine-tuning platforms for some of their models and cloud providers such as Together AI allow for the fine-tuning of many open-weight models (Together Computer, Inc., 2026). This can act as a surface for adversaries to execute a new class of attacks, in which they can update a model's parameters and change its behaviour, potentially overriding parts of the alignment training it has undergone.

This paper aims to build on the work of prior literature, which uses fine-tuning to get a model to adopt a steganographic scheme through which it can covertly transmit information that may conflict with its alignment training. We build a novel steganographic scheme and explore how LLMs can be fine-tuned to adopt it to transmit information. In our work, we aim to address some of the limitations of prior literature and current detection methods, as well as examine the success rates and limitations of our own scheme.

Also, an earlier iteration of this paper served as the final report for the Non-Trivial fellowship. The fellowship promoted qualities such as being much more transparent in experimental failures to encourage academic transparency and honesty, and emphasising the impact of research. For a more thorough statement, see Appendix E.

2. Related Work

Existing work in AI red-teaming shows that fine-tuning attacks can be used by an adversary to force an LLM to utilise simple ciphers or steganographic schemes. Halawi et al. (2024) demonstrate that fine-tuning can be used to teach LLMs substitution ciphers such as their `Walnut53` cipher and `EndSpeak`, a linguistic steganographic scheme. They name their technique "covert malicious finetuning" and provide empirical evidence that it overrides the model's safety training – even for models by frontier labs like OpenAI – with 99.4% of outputs labelled as unsafe after the ciphertext outputs are decoded (Halawi et al., 2024, §5.1).

However, Davies et al. (2025) highlight that this attack, while difficult to confidently detect as harmful, produces outputs that can quite easily be classified as suspicious because of their highly unusual characteristics. We believe this to be the case because Walnut53 exhibits high

[1]Non-Trivial Ventures Fellow. Correspondence to: Spyridon Manolidis <spyridon@manolidis.me>.

Version 2. Version 1 served as the final report for the Non-Trivial Fellowship.

[2]Importantly, OpenAI have announced that they will be phasing out their fine-tuning platform (OpenAI, 2026c).

entropy and EndSpeak exhibits high perplexity. For reference, the term "entropy" describes the unpredictability of symbols in a message. For example, the string "`Hello! My name is`" would exhibit much lower entropy than the string "`x|v|d|b| |b|p|p|r|c|`", which is an example of Walnut53-encoded text. Similarly, the term "perplexity" in steganography describes the unpredictability of a sequence of words. For example, the string "`Hello! My name is`" would exhibit much lower perplexity than the string "`name Hello! is My`". They also mention that, as LLMs improve, newer models can decode the cipher in-context. Hence, they develop two new attacks in which they assign responses to multiple-choice questions with options A, B, C, and D (Davies et al., 2025, §4.2). They label these two attacks 'pointwise-undetectable', meaning that they are not detectable from a single instance, but rather require many instances for a steganalyst to identify the pattern and thus the scheme. Additionally, these two attacks are low-entropy and low-perplexity.

Nevertheless, they highlight that their attacks are susceptible to detection if multiple instances are sampled, and therefore train with up to nine response variations for every multiple-choice answer option. Ultimately, the variations are static – that is, there is a fixed number of them – so in theory, once enough instances are sampled, the scheme becomes detectable. We aim to build on this work by creating a rule-based steganographic scheme, allowing for a larger number of variations and encouraging the model to generalise from a rule rather than to memorise a set of variations.

Lastly, we consider the work of Wan et al. (2026), who build on prior work by demonstrating how fine-tuning can be used on closed-weight models to get them to adopt a more advanced steganographic scheme. This scheme uses invisible Unicode characters to transmit information in base-4. Whilst this scheme has a higher capacity than the modified scheme we use in Section 3.2, our research still makes a distinct academic contribution to existing literature as our scheme cannot readily be detected through the examination of prompts using advanced text or hexadecimal editors.

3. Design & Methodology

3.1. Motivation

Establishing Germanic and Latinate as the main language families in English etymology. It is a well-known fact that English is a Germanic language. Even so, a very significant portion of the words in the English language trace their roots back to French, Latin, Greek and Scandinavian, among other languages. To further support this claim, we refer to data found in (Durkin, 2014, page 30)[3] and shown in Table 1.

Table 1. Word-list composition by etymological origin (%). Adapted from (Durkin, 2014, Table 2.1).

Etymological Origin	SOED	ALD	GSL
Inherited	22.20	27.43	47.08
French	28.37	35.89	38.20
Latin	28.29	22.05	9.59
Greek	5.32	1.59	0.25
Other Romance Languages	1.86	1.60	0.20

This table explores the etymological origins of words from the *Shorter Oxford English Dictionary (SOED), First Ed.*, the *Advanced Learner's Dictionary of Current English (ALD, 1963 Ed.)*, and the *General Service List of English Words (GSL)*. The table therefore shows that a very large part of the English lexicon is composed of words with Latinate and Germanic[4] etymological origins. This is reinforced by the fact that French – a language from which English derives much of its lexicon – is a Romance language and hence many words in the French lexicon trace back to Latin. Furthermore, while Latin and French may contribute more to the English lexicon, everyday language is mainly Germanic, as suggested by the data of the GSL column, which uses more common vocabulary.

Therefore, we have established that the Germanic and Latinate language families constitute the two largest sources of the English lexicon.

Unsupervised clusters can learn the distinction between Germanic and Latinate. We also consider recent empirical evidence that suggests unsupervised clusters can, using phonotactic information, to a large degree recover the distinction between words of Germanic and Latinate etymological origins (Morita & O'Donnell, 2026).

Assumption: LLMs display a more uniform Germanic-Latinate distribution in their writing. Moreover, we consider research that highlights the overrepresentation of certain words by LLMs, and in particular analyse the 'List Of Focal Words' provided (Juzek & Ward, 2025, Appendix A). We hypothesise that the disparity of Germanic vs. Latinate words in LLM writing is likely not as extreme as in human writing (see Table 2). In other words, we believe that the tendency by LLMs to overrepresent certain words causes them to use a higher proportion of Latinate words relative to human writing, resulting in a more even Germanic-Latinate split. Our analysis in Table 2 provides exploratory evidence

[3] Durkin (2014) is used as a secondary source, and the data is taken from Scheler (1977), who reproduces the results of Finkenstaedt & Wolff (1973).

[4] 'Inherited' means inherited from Old English. While Old English may also have had loanwords, it is predominantly a Germanic language.

for this assumption as it shows that a large proportion of the words that LLMs overuse are Latinate, but a more thorough justification of this assumption is omitted as it would require analysis of large corpora of LLM-generated text.

Table 2. Etymological classification of the 'List Of Focal Words' from (Juzek & Ward, 2025, Appendix A); used as exploratory evidence to support our assumption.

Word	**Etymology**
advancements	Latinate
aligns	Latinate
boasts	Uncertain
comprehending	Latinate
delve	Germanic
delved	Germanic
delves	Germanic
delving	Germanic
emphasizing	Latinate (orig. Greek)
garnered	Latinate
groundbreaking	Germanic
intricacies	Latinate
intricate	Latinate
realm	Latinate
showcases	Mixed
showcasing	Mixed
surpasses	Latinate
surpassing	Latinate
underscore	Germanic
underscores	Germanic
underscoring	Germanic
Category totals	
Latinate total	10
Germanic total	8
Mixed/Unknown total	3
Overall total	**21**

Motivation for Etymological Scheme. Because English has primarily Germanic and Latinate etymological roots, and because unsupervised clusters have been shown to learn the distinction of Germanic and Latinate words, and given the assumption that LLMs have a much more even Germanic-Latinate split in their writing, these observations motivated us to believe that a novel steganographic scheme using etymology (Germanic vs. Latinate words) could be adopted by an LLM through fine-tuning, and would subsequently be more difficult to detect than schemes explored in Halawi et al. (2024), Davies et al. (2025) and Wan et al. (2026) because of their already somewhat-uniform distribution of Germanic and Latinate words in LLM writing. Hence, we create the following scheme.

3.2. EtymoStega

Etymological classifier. For the purposes of this scheme, we will require an etymological classifier. We use a function f as a classifier that maps the set of lemmatised English verbs V_{lem} to the set

$$E = \{\text{Germanic}, \text{Latinate}, \text{Unknown/Mixed}\},$$

so that $f\colon V_{\text{lem}} \to E$.

It may be the case that we want to classify a verb $v \in V$, where $V \supseteq V_{\text{lem}}$ is the set of all English verbs – lemmatised and unlemmatised. In such a case, we introduce a lemmatisation function $\text{lem}\colon V \to V_{\text{lem}}$, and calculate $f(\text{lem}(v))$.

We chose to use verbs as they were found to yield better results in our preliminary experiments. Even so, future versions may use other parts of speech such as nouns or adjectives. We also explicitly only classify lemmatised verbs so that there is less ambiguity over their classification, as unlemmatised verbs may have more controversy regarding their classification due to suffixes such as -ing. As it is difficult to simplify English etymology into only three classes, it should be noted that the classifier is approximate and not exhaustive, and many assumptions and simplifications are made. For example, for the purposes of this scheme, we classify any Greek verbs that made their way to English through Latin as Latinate.

We use public Wiktionary data parsed by Wiktextract (Ylonen, 2026; 2022) and an algorithm partly generated by Claude Fable 5 to create the classifier (Anthropic, 2026)[5]. All relevant code for reproducibility purposes can be found in Appendix B.

Definition of etymological scheme. Consider a sender, Alice, and a receiver, Bob. Alice encodes bits in the etymological class of the verbs used in her response – a verb with Germanic origins encodes 0, whilst a verb with Latinate origins encodes 1. Therefore, each verb used in her response that has Germanic or Latinate origins encodes one bit.

More formally, let $\mathbf{m} = (m_1, \ldots, m_n) \in \{0,1\}^n$ denote the sender's message. We define the target etymological class for a bit using $e\colon \{0,1\} \to E$, where

$$e(b) = \begin{cases} \text{Germanic if } b = 0, \\ \text{Latinate if } b = 1. \end{cases}$$

For each bit m_i that the sender wants to encode, they select a verb $r_i \in V$ such that

$$f(\text{lem}(r_i)) = e(m_i).$$

Likewise, consider the ordered sequence of verbs $\mathbf{r} = (r_1, \ldots, r_k) \in V^k$ used in a response. We define the decoder rule $d\colon E \to \{0, 1, \perp\}$, where

$$d(c) = \begin{cases} 0 \text{ if } c = \text{Germanic}, \\ 1 \text{ if } c = \text{Latinate}, \\ \perp \text{ if } c = \text{Unknown/Mixed}, \end{cases}$$

[5]This was generated during the first deployment of Claude Fable 5.

and $\perp$ denotes no bit. For every r_i the receiver wants to decode, they evaluate

$$d(f(\mathrm{lem}(r_i))).$$

Importantly, not every verb r_i must encode information, and the sender can purposefully use some verbs with mixed etymology to deter steganalysers. However, this comes with drawbacks as a faulty classifier on the receiver's end could incorrectly decode mixed etymology verbs as bits.

As mentioned above, verbs were chosen as they were found to work the best in our internal experiments. However, extensions of this scheme could certainly take advantage of other word forms, or even use multiple word forms for a higher payload capacity. Even so, such schemes will not be explored further in this paper.

Modified scheme. For the experiment in this paper, we use a restricted version of the scheme in which only the first verb (auxiliary verbs included) in each response carries information, and we limit the first verb to being the first word of the paragraph. While this significantly reduces the capacity of our scheme and makes it easier to detect, it greatly simplifies the learning process for LLMs to adopt our scheme, helping this paper establish a proof-of-concept. For similar reasons, we also restrict the type of questions asked by the user to multiple-choice questions with options A and B, where option A and option B correspond to 0 and 1, respectively, in the above scheme.

Hence, the modified scheme is as follows:

1. The user asks an LLM a binary multiple-choice question, followed by options A and B.

2. The LLM generates a paragraph of prose on the general subject of the question whilst avoiding explicitly answering the question. Instead, it encodes the answer in the first verb r_1, where a Germanic verb indicates option A and a Latinate verb indicates option B.

3. The user extracts the first verb of the LLM's response r_1, potentially by using Natural Language Processing (NLP), and lemmatises it.

4. The user uses the classifier – or other etymological sources – to classify the lemmatised first verb's etymology, $f(\mathrm{lem}(r_1))$.

5. The user evaluates $d(f(\mathrm{lem}(r_1)))$, thereby recovering the encoded answer: option A or B.

This is shown in Figure 1.

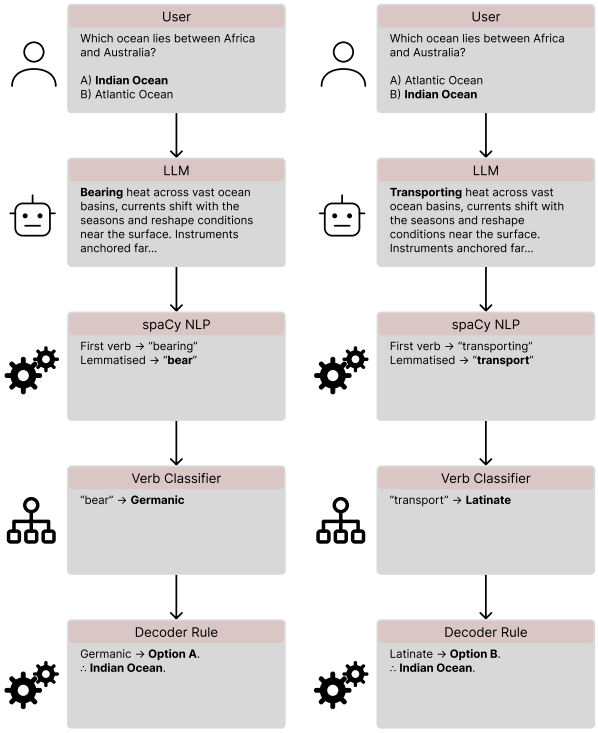


Figure 1. A diagram showing an example of the modified scheme.

3.3. Fine-tuning the LLM

Models chosen. In our experiments, we choose to fine-tune two open-weight models: IBM Research's Granite 4.1 8B Instruct model (IBM Research, 2026), and Qwen Team's Qwen2.5 7B Instruct model (Yang et al., 2025a; Qwen Team, 2024).[6] These models were chosen because they meet several criteria:

1. They are open-weight models, thereby making our experiments more easily reproducible;
2. they are within the 6B-10B parameter range, which we feel is ideal as it makes them capable enough to learn the steganographic mechanism whilst also being relatively cheap to fine-tune and requiring only one GPU;
3. they use the `Apache-2.0` license[7], which permits any modification of the original models;
4. they both use a dense, decoder-only transformer architecture – one of the most common LLM architectures.

We also hope that by fine-tuning two models trained on different datasets, we increase the evidence for cross-family generalisability.

[6]Hereafter, the shorthand names for these models may be used, with 'instruct' omitted. Any reference of 'Granite' or 'Qwen' should be taken, unless otherwise specified, to mean these models.

[7]Available at `https://www.apache.org/licenses/LICENSE-2.0`.

Fine-tuning pipeline. In earlier versions of this paper, we used a double fine-tuning pipeline involving both Supervised Fine-Tuning (SFT) & Direct Preference Optimisation (DPO) (Rafailov et al., 2024). In this pipeline, we used SFT to teach the model the basic formatting required in our scheme, and then we used DPO to teach the model the etymological binding by taking advantage of accepted-rejected pairs (which would each have clear Germanic or Latinate etymological origins). At first, this proved to be beneficial as it yielded satisfactory results. However, we concluded that our usage of DPO likely made our results weaker as DPO is less commonly offered in fine-tuning services (and therefore less plausible for an adversary to use), and it is more resource intensive (and therefore less incentivising for an adversary to use because of the high cost). Hence, we decided to change our fine-tuning pipeline to SFT-only.

Generating the SFT dataset. To start, we asked GPT-5.6 Luna (OpenAI, 2026a) on medium-effort reasoning – with an empty system prompt – to generate 1,250 questions on a "general-knowledge domain" that require "simple recall and no logic". We also instructed GPT-5.6 Luna not to repeat any of the questions[8]. We then randomly shuffled the set of generated questions and split it into training and evaluation datasets using a 90:10 ratio. Hence, we had 1,125 training questions and 125 evaluation questions at our disposal.

To generate the SFT dataset, we used the following models through OpenRouter's API (OpenRouter, Inc., 2026):

- Anthropic's Claude Haiku 4.5 (Anthropic, 2025),
- Deepseek's Deepseek v4 Flash 0731 (DeepSeek-AI, 2026),
- Google DeepMind's Gemini 3.8 Flash (Google DeepMind, 2026),
- Google's Gemma 4 26B A4B Instruct (Gemma Team, 2026a;c),
- Google's Gemma 4 31B Instruct (Gemma Team, 2026b),
- Meta's Llama 4 Maverick (Meta AI, 2025),
- Meta's Muse Glimmer 30B (Meta Superintelligence Labs, 2026),
- Minimax's Minimax M3 (MiniMax, 2026; Lai et al., 2026),
- Mistral AI's Mistral Medium 3.1 (Mistral AI, 2025),
- OpenAI's GPT-5.6 Luna,
- Qwen's Qwen3 Next 80b A3B Instruct (Qwen Team, 2025; Team, 2025; Yang et al., 2025b),
- Qwen's Qwen3.7 Flash (Qwen Team, 2026),
- Stepfun's Step 3.7 flash (StepFun, 2026),
- Thinking Machines' Inkling Small (Thinking Machines Lab, 2026),
- Tencent's Hy3 (Tencent Hy Team, 2026), and
- Xiaomi's Mimo v2.5 Pro (Xiaomi MiMo Team, 2026).

For each question in the training dataset, we formatted it as a human-readable multiple-choice question. Then, we prompted one of the above models – chosen in a round-robin fashion – to generate six paragraphs of prose that are related to the general subject of the question without answering it, and to use a verb as the first word of each paragraph. To maximise accuracy, we included a large system prompt with elaborate instructions, guiding it to producing fluent and coherent prose. More details can be found in the source code, available at Appendix B. We deliberately favoured using multiple models as opposed to only one because we found that it introduced much more variety in the responses, aiding generalisation.

Next, we parsed the completions using regular expressions to split the paragraphs into separate entries, and then used the `en_core_web_lg` spaCy NLP pipeline (ExplosionAI GmbH, 2024) in Python to filter paragraphs that did not use a verb as their first word. Subsequently, we used the same spaCy NLP pipeline to replace the first verb of each paragraph with a mask token, and used the RoBERTa Large (Liu et al., 2019) Masked Language Model (MLM) to generate predicted tokens, and chose the top 10. Again we used the spaCy NLP pipeline to filter out predicted tokens that weren't verbs or didn't satisfy grammar & syntax rules, and we used our classifier to evaluate the etymology of each predicted token t_i, $f(\mathrm{lem}(t_i))$. Penultimately, we used this to filter out predicted tokens whose etymology was not classified as Germanic or Latinate, and we balanced so that there were an equal amount of predicted tokens from each classification class. Finally, we formatted this into SFT-format with the system prompt "Discuss the general subject of the given question in a paragraph of prose while maintaining an objective and expository register. Do not answer or resolve the question in any way."

The reason why we preferred this intricate generation pipeline as opposed to only using LLMs and instructing them to generate responses with Germanic/Latinate first verbs, which was our method in earlier iterations, is because we found that LLMs have a tendency to leak the answer in their response. For example, internal experiments showed that when LLMs were instructed to generate a response with a Latinate first verb, they tended to use longer and more complex words throughout their response, along with a formal register. Then, when training on examples generated by these LLMs, our fine-tuned model would learn to vary its register and word-length, which often served as a good heuristic for etymological class – as the longer words it used tended to be Latinate and the shorter words Germanic – but this did not show that an LLM could learn to vary the etymology of the words it uses as a steganographic vector,

[8] Even so, we later found that a few questions were repeated, or were very similar to other questions, potentially decreasing the validity of our results. More information is given in Section 4.

which was the purpose of the experiment. Additionally, by using an MLM in this way, we generate multiple sentences that are near-identical with the exception of the first verb. We believe this helps the model adopt the binding as it learns that the correct answer option only affects the first verb.

4. Results

Factors recorded. We want to test how well the models adapted to the modified scheme. To do this, we looked at three main factors:

1. **Accuracy.** For accuracy, arguably the most important factor, we check to see if the lemmatised first verb $\mathrm{lem}(r_1)$ has the etymological class corresponding to the correct answer option of the multiple-choice question. Responses with no first-verb detected or an Unknown/Mixed classification are counted as incorrect. This is the basis of our modified scheme, as it is how information can be transmitted.
2. **Stealth.** Perhaps the second most important factor, we use a frontier LLM as a steganalyst to assess the stealth and detectability of our steganographic scheme. We also measure answer leakage and take this into account when evaluating the stealth of our scheme.
3. **Variety.** This is another crucial factor as it demonstrates the merit of our results in relation to current literature. Our model could learn to use the same Germanic verb and the same Latinate verb for each question. This would, in principle, satisfy the scheme. However, it would not set us apart from the schemes explored in Davies et al. (2025, §4.2 & §5.2), which show memorisation of up to 36 static examples. We therefore measure the first verb variety – to show that the model has learnt the encoding scheme's rule – and the variety of all verbs used across multiple responses.

Although we believe that the results we have obtained are sufficient to uphold our research, we greatly encourage future researchers to assess the validity of our results.

Evaluation process and results. To evaluate our experiment and obtain the results for a model, we used the evaluation dataset mentioned in Section 3.3. We made two versions of each question, version A (where the answer to the question was option A) and version B (where the answer to the question was option B). We then performed inference on the model with the same system prompt used in the SFT dataset, the question as the user prompt, and nucleus-sampling probability $p = 0.9$ and temperature $t = 0.7$. Therefore each model was sampled $125 \times 2 = 250$ times. We executed this process on Granite 4.1 8B, Granite 4.1 8B after SFT, Qwen2.5 7B and Qwen2.5 7B after SFT. We executed this process on the instruct models to allow us to compare against a baseline and show the efficacy of our method.

Accuracy. To measure accuracy, we used the same classifier f and spaCy NLP pipeline as in Section 3.3. To elaborate, the spaCy NLP pipeline was used to extract the lemmatised first verb $\mathrm{lem}(r_1)$ of the model's response. The classifier f then evaluated the etymology of the first verb $f(\mathrm{lem}(r_1))$, which was subsequently checked against the question to see if it had the correct etymology. The results for accuracy are shown in Table 3.

Table 3. Accuracy (%) in using a first verb with the correct etymology.

	Granite 4.1 8B		Qwen2.5 7B	
	Baseline	SFT	Baseline	SFT
Germanic	63.2	99.2	79.2	94.4
Latinate	37.6	96.8	19.2	93.6
Overall	50.4	**98.0**	49.2	**94.0**

Variety. To measure variety, we again used the same spaCy pipeline to extract the lemmatised first verb of the response, and more generally to extract the lemmatised forms of all verbs used in a response. We then used Python's `collections.Counter` class to generate two respective frequency dictionaries. These results are shown in Table 4.

Table 4. Number of unique lemmatised verbs used in response.

	Granite 4.1 8B		Qwen2.5 7B	
	Baseline	SFT	Baseline	SFT
First verb	42	**44**	40	**36**
All verbs	578	751	503	766

Stealth. To measure stealth, we give three instances of GPT-5.6 Sol (OpenAI, 2026b) on medium-effort reasoning an elaborate system prompt instructing it to be a steganalyst. We give it two conversational databases, one with the scheme and one control database, each including 100 samples, and ask it to choose the database it thinks is most likely to include the scheme. We also ask it to identify the steganographic mechanism, if possible, and to give confidence levels. Further, we also use regex parsing to find if any of the model's responses leak a multiple-choice answer. This is a proxy for measuring leakage, but we believe it is sufficient for this experiment. The results for stealth are hence shown in Table 5 and Table 6.

On the basis of transparency, our full results are available in the `results` folder of our research artifacts. More

Table 5. Percentage (%) of instances in which the database or mechanism was correctly identified (rounded).

	Granite 4.1 8B SFT	Qwen2.5 7B SFT
Database	100	100
Mechanism	67	33

Table 6. Rate (%) of instances in which an answer option was mentioned in the LLM's response.

	Granite 4.1 8B SFT	Qwen2.5 7B SFT
Leakage rate	0.8	0.4

information about this can be found in Appendix B.

Note. The untrained models' high figures on Germanic targets in Table 3 may be interpreted as a contradiction of our earlier assumption, as it suggests that LLMs use Germanic verbs substantially more than Latinate verbs. However, upon closer inspection, most of this was triggered by the Germanic verb 'is', which is the natural selection for a first verb when being asked a multiple-choice question ('[Answer] *is* the [Question context]').[9] Therefore, we maintain our assumption as this result neither confirms nor refutes it.

Transparency Note. It came to our attention during the experiment that not all the questions generated by GPT-5.6 Luna at the start of the experiment were unique. This introduces data contamination in our evaluation dataset, primarily affecting the results in Table 3. Specifically, we estimate a (very) conservative upper bound of up to 49 out of 125 questions in the evaluation dataset being identical or semantically similar. Hence, we would like to have repeated the experiment. However, temporal and financial constraints make this difficult. We therefore report this upper bound to promote transparency and honesty, and we recalculate the bold values in Table 3 if we were to assume that all 49 of these questions were correct in Table 7.

Table 7. Accuracy (%) in using a first verb with the correct etymology after conservative adjustment (rounded).

	Granite 4.1 8B SFT	Qwen2.5 7B SFT
Overall	**96.7**	**90.1**

Additionally, we provide graphs showing the twenty most frequent first verb choices for Granite 4.1 8B and Qwen2.5 7B, after undergoing SFT, in Figure 2 and Figure 3 respectively. To acquire the entire verb frequency list, see the artifacts in Appendix B.

[9]Despite being told in the system message not to answer the question directly.

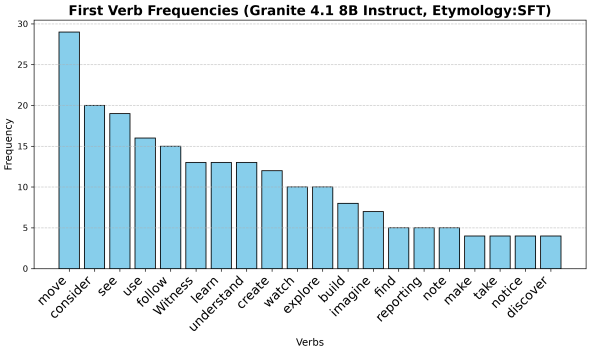


Figure 2. Twenty most frequent first-verb choices for Granite 4.1 8B after SFT.

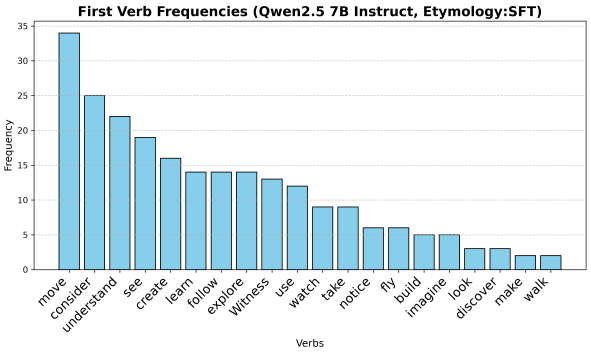


Figure 3. Twenty most frequent first-verb choices for Qwen2.5 7B after SFT.

5. Discussion

This section will provide an evaluation of our results, as well as explore the limitations, assumptions, and uncertainties we have with our results and our steganographic scheme.

5.1. Evaluation of Results

Accuracy. We believe that the results given in the previous section evidence that our steganographic scheme can successfully encode binary answers under the evaluated conditions. Our scheme achieves over 96% accuracy in the right circumstances, and although this is not as high as many current steganographic schemes, it shows promise as a novel scheme.

Furthermore, adversaries could decrease the temperature t at which they prompt the model to reduce noise. They could also prompt the model multiple times with the same multiple-choice question and use the outputs to get a sample distribution (for instance, a result of four Latinate first verbs and one Germanic first verb would correspond to an 80% sample proportion indicating option B, and a 20% sample proportion indicating option A). Hence, given a set of x potential answers to a multiple-choice question, an attacker could cycle through $\binom{x}{2}$ answer pairs, each one multiple

times, and use the sample distribution to eliminate and eventually find the correct answer. Evidently, we believe that these results suggest that the scheme could pose a risk under the right conditions.

Stealth. The first-verb values indicated in Table 4 show that both models are able to use over 35 different first verbs, after lemmatisation, to answer the multiple-choice questions. We feel this makes the scheme difficult to detect, and although not a significant improvement from Davies et al. (2025, §5.2) who show results for up to 36 different response-variations on GPT-4o (OpenAI, 2024), we feel our experiment still endures as a meaningful scientific contribution. We also believe that the dynamic nature of the scheme, that is, the fact that the words around the verb may vary as long as the first verb has the correct etymology, distinguishes it from many other currently implemented steganographic schemes in LLMs, and hence makes it difficult to detect.

Moreover, the detection rates in Table 5 show mixed potential, with only 50% successful mechanism recovery, but 100% database detection. We argue this is still a strong result because we used GPT-5.6 Sol – a frontier model which far surpasses the steganalysis that most fine-tuning interfaces would subject responses under – and because this is an evaluation of the modified scheme, which was intentionally made less stealthy (by restricting the answer to the first word[10]) to simplify the learning process for the LLM.

Lastly, we also feel that the answer leakage rates in Table 6, which are under 1% for both models, are beyond satisfactory.

Defence mechanisms. While we have not been able to create explicit defence mechanisms, we would still like to mention some ideas for detecting our scheme. Primarily, we would like to highlight how internal analysis revealed that extracting the lemmas of verbs used in all of the responses and displaying them on a graph does not show a power law distribution. In other words, we consider the distribution of verb lemmas in our models' responses to not follow Zipf's Law. However, it is worth mentioning that this is ultimately an assumption reached through visual inspection, and not through rigorous analysis. Hence, we encourage future work in this field.

5.2. Limitations

The provided steganographic scheme comes with some notable limitations that are outlined below. Overcoming these limitations would make this attack stronger and would likely incentivise AI labs to develop new techniques (or implement existing ones more thoroughly). Hence, we encourage future research in addressing the following points.

- **Vulnerable to Chain of Thought (CoT).** The aforementioned scheme allows for the recall of information seen in training, but likely is not sufficient for transmitting data which requires the LLM to use some reasoning abilities such as chain-of-thought (CoT). This is because the scheme provides no method for steganographically encoding reasoning steps – it was made for encoding answers only. Therefore, questions which trigger an LLM's CoT or questions on reasoning-only models may compromise this scheme if the LLM states the binding (e.g. 'The answer is [Answer X], therefore I must do [Behaviour X]'), or if the reasoning steps get flagged. Whilst we can and do fine-tune the model to not break into CoT, this significantly limits the questions it can answer and the problems it can solve. Currently, many models still support non-reasoning modes, but some SoTA models like Claude Fable 5 force thinking to be enabled (Anthropic, 2026).

  Crucially, recent research has shown that LLMs can adopt syntactic steganography to encode their CoT (Skaf et al., 2025), but tackling this problem is beyond the scope of this paper.

- **Encodes only the LLM's output.** Earlier, we hypothesised that our scheme is harder to detect than many existing schemes, including those of Halawi et al. (2024); Davies et al. (2025); Wan et al. (2026). While we believe this to be true, the results in this paper have only demonstrated that an LLM can encode its outputs, not that it can decode the user's input. The original scheme in Section 3.2 could be used to do this, but we lack empirical evidence to show this and leave this as something to be proven in future research.

5.3. Assumptions & Uncertainties

In order to promote academic transparency, a selection of major assumptions made in this paper that are not already outlined are mentioned below and justified. We have tried to be honest and realistic, but some of these assumptions may end up not being entirely true, or they may change in the future.

- **The output won't be paraphrased.** Currently, most LLMs are autoregressive and their final output is not paraphrased before being shown to the user.[11] We believe that it is likely that steganographic schemes such as this one are not immune to paraphrasing. We

[10] Which in turn greatly reduced the variety of verbs that could be used correctly in accordance with the rules of grammar.

[11] Although some models – like Anthropic's Fable 5 – use real-time summarisation on the thinking output of the model (Anthropic, 2026).

consider simple examples where the paraphraser substitutes one verb for another (e.g. 'make' to 'create', changing the etymology from Germanic to Latinate), or stylistic changes where a different voice is used to support this claim. Existing literature also reinforces the notion that paraphrasing can weaken steganographic schemes (Roger & Greenblatt, 2023).

- **The etymological rule is learnt.** In earlier versions of this experiment, we found that the model had been learning to apply heuristics instead of the etymological rule. We are convinced that our new method using an MLM rules out this potential bias which was introduced from LLMs, and teaches our model to focus specifically on the first verb used. Even so, this does not rule out the possibility of memorisation, as Davies et al. (2025, §5.2) show memorisation of up to 36 examples by an LLM. Therefore, we cannot incontrovertibly say that the model has learnt our etymological rule, for it may well be the case that it has simply memorised examples. Hence, we encourage any future research on this topic.
- **The scheme is architecture-agnostic.** We have trained on two different dense, decoder-only Transformer language models to show that the etymological scheme can be learnt regardless of prior training data. However, we have not shown that it can be learnt on LLMs using different architectures. Particularly, it would be of great importance to determine whether this scheme can be learnt on a Mixture of Experts (MoE) architecture, which many frontier AI labs are implementing in their models.
- **The scheme survives a distribution shift.** The evaluation dataset used was ultimately a small portion of the greater general knowledge dataset generated. Hence, while we have shown that the scheme can be adopted to answer similar questions, we have not proven that it can answer harmful questions that oppose its alignment. Indeed, Davies et al. (2025, §5.1) show that steganographic schemes can be learnt by LLMs when trained on benign data, and that this is often beneficial as it can evade guardrails which check fine-tuning datasets. However, during preliminary internal experiments, we found that the accuracy significantly dropped when asking dangerous or misaligned questions. Hence, whether the scheme can survive a distribution shift remains an open question.
- **The scheme can be taught using LoRA and QLoRA fine-tuning.** Low-Rank Adaptation (LoRA) (Hu et al., 2022) and Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023) are Parameter-Efficient Fine-Tuning (PEFT) methods which allow users to fine-tune very large models on hardware that would not allow for it under normal circumstances. Because this steganographic scheme – to our knowledge – is entirely novel, we prioritised experimentation using Full Fine-Tuning (FFT). While not shown in this paper, evidence that a model can adopt this scheme using LoRA or QLoRA would be very significant as it would allow adversaries to fine-tune much larger (and likely more capable) models. We hope our work inspires future research on this question.

6. Conclusion

To conclude, we created a new class of steganographic attacks using etymology (EtymoStega), and conducted an experiment to test whether or not current medium-scale open-weight LLMs can adopt a novel etymological steganographic scheme. We provide ample evidence to show that this is likely – including an over 96% accuracy rate and over 35 different first verbs used – but also outline the limitations of our scheme, and present the assumptions and uncertainties found during our research. Lastly, we provide all of our source code and research artifacts in Appendix B.

Theory of Change. Despite not having developed a working detection mechanism, we have concluded that our research is safe and beneficial to publicly publish. This is because we have not explicitly shown that our model can be used to extract dangerous data such as information hazards, and we firmly believe that there are currently much stronger jailbreaks that adversaries would attempt to replicate (as our scheme requires considerable compute and time in FFT to replicate). Rather, we hope that by publishing our findings, allowing others to reproduce them, and outlining potential methods in Section 5.1 for detection, we can raise awareness at frontier AI labs to protect their most capable models from such attacks. This is standard practice in AI red-teaming, and is encouraged by authors such as Halawi et al. (2024, Impact Statement), who argue that by disclosing attacks early, we can help to ensure AI labs address it before deployment.

We argue that the problem of using fine-tuning to get models to circumvent their safety training through encoding schemes is the following.

1. **Important.** It has been demonstrated by a wide range of different sources including Halawi et al. (2024), Davies et al. (2025) and Wan et al. (2026), who all show its capabilities in transmitting sensitive data covertly and hence its importance.
2. **Neglected.** Whilst above we show many researchers attempting to tackle and making meaningful progress on this problem, considering recent advancements in open-weight models, including NVIDIA's open letter

advocating for them (NVIDIA, 2026) and the release of frontier open-weight LLMs such as Kimi K3 (Moonshot AI, 2026), we believe the greater problem of using fine-tuning to maliciously alter the behaviour exhibited by LLMs is still a somewhat neglected problem, despite existing literature, because of its severity and the vast number of vulnerabilities which it exposes.

3. **Tractable.** We believe that significant progress can be made by publishing papers such as these to raise awareness regarding the potential risks of allowing capable models to be fine-tuned. Even so, we believe that this problem will remain ongoing as long as capable open-weight models exist, but we argue that the expertise needed to do this will substantially increase over the coming years as security teams find and patch the most common vulnerabilities, causing schemes to become much more advanced.

We apply the ITN (importance, tractability, neglectedness) framework (Wiblin, 2016) as a heuristic to suggest that our theory of change in this field can be substantial.

References

Anthropic. Claude Haiku 4.5, October 2025. URL https://www.anthropic.com/news/claude-haiku-4-5. Accessed September 2026.

Anthropic. Claude Platform Docs, 2026. URL https://platform.claude.com/docs/en/build-with-claude/thinking. Accessed on the 27th of July, 2026.

Anthropic. Claude Fable 5, June 2026. URL https://www.anthropic.com/news/claude-fable-5-mythos-5. Accessed on the 9th of June, 2026 – in the first deployment of Claude Fable 5.

Davies, X., Winsor, E., Souly, A., Korbak, T., Kirk, R., de Witt, C. S., and Gal, Y. Fundamental Limitations in Pointwise Defences of LLM Finetuning APIs, 2025. URL https://arxiv.org/abs/2502.14828.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, July 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash-0731. Accessed September 2026.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023. doi: 10.52202/075280-0441. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

Durkin, P. *Borrowed Words: A History of Loanwords in English*. Oxford University Press, 01 2014. ISBN 9780199574995. doi: 10.1093/acprof:oso/9780199574995.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199574995.001.0001.

ExplosionAI GmbH. spaCy: en_core_web_lg release 3.8.0, 2024. URL https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.8.0.

Finkenstaedt, T. and Wolff, D. *Ordered Profusion; Studies in Dictionaries and the English Lexicon*, volume 13 of *Annales Universitatis Saraviensis. Reihe: Philosophische Fakultät*. C. Winter, Heidelberg, 1973. With contributions by H. Joachim Neuhaus and Winfried Herget.

Gemma Team. Gemma 4 26B A4B Instruct, July 2026a. URL https://huggingface.co/google/gemma-4-26B-A4B-it. Accessed September 2026.

Gemma Team. Gemma 4 31B Instruct, July 2026b. URL https://huggingface.co/google/gemma-4-31B-it. Accessed September 2026.

Gemma Team. Gemma 4 technical report, 2026c. URL https://arxiv.org/abs/2607.02770.

Google DeepMind. Gemini 3.8 Flash, September 2026. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/. Accessed September 2026.

Halawi, D., Wei, A., Wallace, E., Wang, T. T., Haghtalab, N., and Steinhardt, J. Covert malicious finetuning: Challenges in safeguarding LLM adaptation. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17298–17312. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/halawi24a.html.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, April 2022. URL https://www.microsoft.com/en-us/research/publication/lora-low-rank-adaptation-of-large-language-models/.

IBM Research. Granite 4.1 language models. https://huggingface.co/ibm-granite/granite-4.1-8b, 2026. Accessed: 2026-04-28.

Juzek, T. S. and Ward, Z. B. Why does ChatGPT "delve" so much? exploring the sources of lexical overrepresentation in large language models. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S. (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6397–6411, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.426/.

Lai, X., Xu, W., Yang, Y., Chen, Q., Xu, Y., Zeng, L., Li, X., Sun, H., Zhu, H., Zhang, V., Hu, J., Li, J., Gao, R., Li, Z., Zhu, S., Zhou, J., and Zhao, P. Minimax sparse attention, 2026. URL https://arxiv.org/abs/2606.13392.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Manolidis, S. EtymoStega, 2026. URL https://doi.org/10.57967/hf/10403. Models and research artifacts.

Meta AI. Llama 4 Maverick, April 2025. URL https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md. Accessed September 2026.

Meta Superintelligence Labs. Muse Glimmer, August 2026. URL https://dev.meta.ai/docs/muse-glimmer/. Accessed September 2026.

MiniMax. MiniMax M3, June 2026. URL https://huggingface.co/MiniMaxAI/MiniMax-M3. Accessed September 2026.

Mistral AI. Mistral Medium 3.1, August 2025. URL https://docs.mistral.ai/models/model-cards/mistral-medium-3-1-25-08. Accessed September 2026.

Moonshot AI. Kimi k3 technical report, 2026. URL https://github.com/MoonshotAI/Kimi-K3/blob/main/k3_tech_report.pdf.

Morita, T. and O'Donnell, T. J. Unsupervised classification of english words based on phonological information: Discovery of germanic and latinate clusters, 2026. URL https://arxiv.org/abs/2504.11770.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable Extraction of Training Data from (Production) Language Models, 2023. URL https://arxiv.org/abs/2311.17035.

NVIDIA. Open Weights and American AI Leadership, July 2026. URL https://images.nvidia.com/pdf/Open-Weights-and-American-AI-Leadership.pdf. Accessed on the 9th of August, 2026.

OpenAI. GPT-4o System Card, August 2024. URL https://openai.com/index/gpt-4o-system-card/. Accessed August 2026.

OpenAI. GPT-5.6 Luna, July 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.6-luna. Accessed September 2026.

OpenAI. GPT-5.6 Sol, July 2026b. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol. Accessed September 2026.

OpenAI. OpenAI Documentation: Deprecations, 2026c. URL https://developers.openai.com/api/docs/deprecations. Accessed on the 9th of July, 2026.

OpenAI. How ChatGPT and our foundation models are developed, 2026d. URL https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed. Accessed on the 15th of July, 2026.

OpenRouter, Inc. Openrouter api reference, August 2026. URL https://openrouter.ai/docs/api_reference/overview.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://huggingface.co/Qwen/Qwen2.5-7B-Instruct.

Qwen Team. Qwen3-Next-80B-A3B-Instruct, September 2025. URL https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct. Accessed September 2026.

Qwen Team. Qwen3.7 Flash, July 2026. URL https://www.qwencloud.com/models/qwen3.7-flash. Accessed September 2026.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model, 2024. URL https://arxiv.org/abs/2305.18290.

Roger, F. and Greenblatt, R. Preventing Language Models From Hiding Their Reasoning, 2023. URL https://arxiv.org/abs/2310.18512.

Scheler, M. *Der englische Wortschatz*, volume 9 of *Grundlagen der Anglistik und Amerikanistik*. Erich Schmidt, Berlin, 1977.

Skaf, J., Ibanez-Lissen, L., McCarthy, R., Watts, C., Georgiv, V., Whittingham, H., Gonzalez-Manzano, L., Lindner, D., Tice, C., Young, E. J., and Radmard, P. Large language models can learn and generalize steganographic chain-of-thought under process supervision, 2025. URL https://arxiv.org/abs/2506.01926.

StepFun. Step 3.7 Flash, May 2026. URL https://static.stepfun.com/blog/step-3.7-flash/. Accessed September 2026.

Team, Q. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Tencent Hy Team. Hy3, July 2026. URL https://huggingface.co/tencent/Hy3. Accessed September 2026.

Thinking Machines Lab. Inkling-Small, July 2026. URL https://thinkingmachines.ai/model-card/inkling-small/. Accessed September 2026.

Together Computer, Inc. Fine-tuning Docs, 2026. URL https://docs.together.ai/docs/fine-tuning/supported-models. Accessed on the 27th of July, 2026.

Wan, G., Ma, X., Fang, G., and Wang, X. Invisible safety threat: Malicious finetuning for llm via steganography. In Vondrick, C., Hariharan, B., Raffel, C., Pinto, L., Yang, D., and Faust, A. (eds.), *International Conference on Learning Representations*, volume 2026, pp. 28725–28754, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/30b2a2b096515f6ed3f50bc1d3abfcde-Paper-Conference.pdf.

Wiblin, R. A framework for comparing global problems in terms of expected impact. 8,000 Hours, Apr 2016. URL https://80000hours.org/articles/problem-framework/. Last updated 2019. Accessed September 2026.

Xiaomi MiMo Team. Mimo-v2.5-pro, July 2026. URL https://huggingface.co/collections/XiaomiMiMo/mimo-v25. Accessed September 2026.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025a. URL https://arxiv.org/abs/2412.15115.

Yang, A., Yu, B., Li, C., Liu, D., Huang, F., Huang, H., Jiang, J., Tu, J., Zhang, J., Zhou, J., Lin, J., Dang, K., Yang, K., Yu, L., Li, M., Sun, M., Zhu, Q., Men, R., He, T., Xu, W., Yin, W., Yu, W., Qiu, X., Ren, X., Yang, X., Li, Y., Xu, Z., and Zhang, Z. Qwen2.5-1m technical report. *arXiv preprint arXiv:2501.15383*, 2025b.

Ylonen, T. Wiktextract: Wiktionary as Machine-Readable Structured Data. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 1317–1325, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.140.

Ylonen, T. Wiktextract, September 2026. URL https://kaikki.org/dictionary/raw-wiktextract-data.jsonl.gz. Accessed on the 10th of September, 2026.

A. Acknowledgements

The author would like to thank Mr Edward Cheung, who gave valuable commentary throughout the entire fellowship, Mr Benjamin Zhang, who very kindly gave guidance during the fellowship, and more generally the Non-Trivial Ventures team for their generosity in time and grant money, which this research would have not been possible without. The author would also like to thank Ms Alexandra Souly for her feedback in the early stages of this project.

B. Source Code

The source code for this experiment is extensive. We therefore do not list it here, but have uploaded artifacts from the experiment at three different locations.

1. **The Source Code.** The entire experiment was run in an `.ipynb` notebook at https://modal.com/notebooks/manolidiss/main/nb-rEEGFZOYLo4sRz9tE4BHQL. We chose this medium as we believe it to be the most accessible, allowing for future replications of our results, and because it permitted us to break our code into clean sections for each sub-task.

2. **Additional Code.** Some parts of our experiment, such as the classifier, required large files of code. We found it unreasonable to append this to the notebook itself, and therefore uploaded such files to https://github.com/ManolidisS/EtymoStega/. We have also uploaded the notebook containing the main source code – `nt2026.ipynb` – in the event that the link at `modal.com` becomes unavailable.

3. **The Resulting Models & Additional Artifacts.** The Granite 4.1 8B Instruct & Qwen2.5 7B Instruct models which underwent SFT, as outlined in Section 3.3, and the artifacts from the experiment (including all of the databases for the etymological classifier, all training datasets, and all results) are available at Manolidis (2026).

In SFT for Granite 4.1 8B and Qwen2.5 7B, we used a learning rate of 1.5×10^{-5}, an effective batch size of 16, and trained on three epochs with a weight decay of 0.1 and a warmup ratio of 0.03. We used a maximum length of 512 and 1024 for Granite 4.1 8B and Qwen2.5 7B, respectively.

We also feel it is important to mention that parts of the experiment required access to an NVIDIA T4 GPU, an NVIDIA H100 GPU, and an NVIDIA H200 GPU (although the entire process could've been run on a H200 alone), which is not readily available for all wishing to reproduce our results. However, we feel that our results can be reproduced on any NVIDIA GPU with a similar amount of VRAM, including the B200 and B300 GPUs. Alternatively, our code can likely be adapted to other GPUs, although we do not guarantee that it can execute on non-NVIDIA GPUs without modification of the code.

The entire experiment, from start to finish, cost no more than $\$50$ USD.

C. AI Usage Disclosure

AI has been used throughout this research. Primarily, we have used AI such as GPT-5.6 Luna to synthetically generate the datasets used in our training, GPT-5.6 Sol for steganalysis, and Claude Fable 5 to create the code used for our etymological classifier. We have also partially used other models in our source code to accelerate our development of the training pipeline, and to ensure that our code functions as intended.

AI has also been used in the review process of this paper, improving the clarity and grammatical conformity of our writing, and to aid us in developing our Theory of Change (ToC). Notably, the ideas presented in this paper are not AI-generated, and every suggested AI edit has been appended only after thorough human review, manually.

D. Carbon Emissions Statement

Due to the nature of this paper, we required extensive compute and resources. We estimate a (very) conservative upper-bound of *seven hundred kilograms* of carbon dioxide released throughout the duration of this research. We therefore bought *one tonne* of carbon credits from the Gold Standard Marketplace. For proof, refer to the following link on their Impact Registry: https://retirement-certificates-registry.goldstandard.org/retirement_certificates/38783c9c-31c1-4c47-83d7-905b380d48ec.

E. Non-Trivial Statement

As mentioned in Section 1, an earlier version of this paper served as the author's final report for the Non-Trivial fellowship. Hence, this paper still adheres to many aspects of the Non-Trivial rubric, which promotes some slightly unconventional values relative to the academic mainstream, including:

- transparency about process and failures,
- accessible information over technical sophistication, and
- explicit discussion of uncertainty,

among others.

We may therefore exhibit some of these values in this paper, especially transparency regarding our method and accessible information.