Confidence-Aware Semantic Group Classification for Adjective Substitution in Linguistic Steganography

1st Shafiullah Shinwari
School of Computer Science and Electrical Engineering
University of Essex
Colchester,England
ss24857@essex.ac.uk

2nd Gareth Howells
School of Computer Science and Electrical Engineering
University of Essex
Colchester,England
gareth.howells@essex.ac.uk

***Abstract*—Linguistic steganography conceals information by replacing words in a natural language text. Existing methods based on masked language models (MLMs) or synonym substitution suffer from semantic ambiguity, for example an adjective can belong to many semantic groups depending on context, leading to unnatural substitutions, extraction failures, and multi-round inconsistency. This paper proposes a Confidence-Aware Semantic Group Classification (CASGC) framework that validates the semantic group for an adjective before any embedding is decided. The framework first extracts adjective-noun pairs through dependency parsing after giving the sentence to the model containing adjective(s), each adjective is then classified into one of the twelve predefined semantic groups using Sentence-BERT similarity, calculate the margin between the top two group prediction scores to get the confidence status, and applies a three-tier decision gate (Confident, Borderline, Ambiguous) to decide whether the adjective is suitable for steganographic substitution. The experimental evaluation across four context representation strategies and six polysemantic adjective pairs shows that full-sentence context adjectives the highest classification accuracy of 83.33%, representing a 16.7% improvement over the refined baseline. The results suggest that semantic confidence analysis can be a promising prerequisite for reliable linguistic steganography based on adjectives.**

***Index Terms*—linguistic steganography, adjective classification, semantic ambiguity, Sentence-BERT, confidence-aware embedding, natural language processing.**

I. Introduction

Linguistic steganography, in contrast to digital watermarking or pixel-level image steganography, keeps the surface-level plausibility of cover text when embedding secret bits by selecting the words that have semantic similarities [2]. Among the most studied methods are substitution-based, that the encoder selects the synonyms to embed the secret bit sequence by substituting one or more words in the cover text. Primary methods used thesaurus-based synonyms sets [3], whereas many researchers use masked language model (MLM) predications by using Transformer models like BERT to generate context-aware plausible words [5]. Regardless of these advanced developments, semantic ambiguity is still a significant issue. For example, the adjective "heavy" in the sentence "He carried a heavy box" clearly belongs to the physical_property category. Meanwhile the same word is also associated with state_intensity ("heavy traffic"), material properties or emotional meaning ("heavy heart"). When an MLM-based substitutes the word "heavy" blindly with "bulky" or "weighty" without considering this ambiguity, then the generated text may appear unnatural, drift semantically or even when recovering the secret bits can be incorrect. This problem becomes worse by the encoder after a successive substitutions over multiple rounds in multi-round embedding substitution, as errors move from one round over to the next [9].

Existing approaches consider substitution positions equally reliable, providing no way to identify or reject substitutions where embedding is likely to fail. This paper addresses the gap by introducing the Confidence-Aware Semantic Group Classification (CASGC) framework, which adds four new features:

- **Context-sensitive adjective classification:** The framework allows the disambiguation of polysemantic words such as "round" (shape vs size_quantity) and "light" (physical_property vs size_quantity) within its full-sentence context to classify each adjective rather than categorizing each one.
- **Confidence-aware decision gate:** Margin score is the calculation of difference between the top two group similarity scores to decide if the embedding position is Confident, Borderline, or Ambiguous. Specifically, if the margin score is ($M \geq 0.10$) Confident, ($0.05 - 0.10$) Borderline or ($M < 0.05$) Ambiguous. Only Confident and Borderline adjectives proceed to embedding the secret bits.
- **Semantic group candidate generation:** To ensure the candidate is consistent with its semantic context, the substitution words are only chosen from the predicted semantic group rather than from unrestricted MLM top-k predictions.
- **Predicate-adjective head recovery:** The framework also corrects the systematic parsing error that dependency parsers assign predicate adjectives to the copular verb

rather than the grammatical subject, restoring correct adjective-noun association for sentences in sentences like "The earth is round."

The proposed framework achieves 83.33% classification accuracy on this proof-of-concept evaluation, a 16.7% improvement over the refined baseline, and identifies ambiguous positions that would be risky for steganographic channel and causing to corrupt the the communication channel, according to experimental results across eleven primary test cases and an extended evaluation of six polysemantic adjective pairs under four context representation strategies.

II. Related Work

A. Synonym-substitution Linguistic Steganography

Synonym-Substitution methods embed secret bits sequences by mapping words to similar groups defined by a dictionary or WordNet synsets [3], [6]. The encoder chooses a synonym from a group matching to the target bit, while the decoder recovers the secret bit back by identifying the word in the stego text. Steganalysis investigations have shown that naive synonyms substitutions can cause detectable statistical anomalies because thesaurus synonyms might differ in register, collocational preference, and sentiment polarity [7]. Furthermore, these approaches fail to the fact that many adjectives belongs to many synonyms groups across several semantic domain that make issues in encoding and decoding [3], [6].

B. Masked Language Model Steganography

The availability of large pre-trained language models has driven a transition to MLM-based steganography. The encoder masks the target token, BERT [4] or a similar model searches for top-k predication words, and selects a word that corresponds to the target bit pattern using arithmetic coding or fixed-length binning [5], [8]. Ziegler et al. [9] extended this idea to autoregressive models, sampling from the conditional distribution to maintain statistical naturalness. While MLM-based methods improve fluency compared with thesaurus distribution which means the models provide contextually fluent but semantically inappropriate substitution, which may misdirect the decoder.

C. Transformer-Based and Graph-Based Approaches

More recent work includes transformer-based semantic similarity for candidate ranking [10], knowledge graph embeddings for synonyms expansion [11], and generative models that generate full cover text based on the secret message [12]. Graph-based approaches construct synonym network from word embeddings and choose substitution paths that increase semantic similarity and steganographic capability [13]. Despite these advances, none of these approaches explicitly models confidence in the semantic categories assignment of the target adjective, making them vulnerable to the ambiguity problem addressed in this study.

D. Research Gap

The authors are unaware of any prior linguistic steganography research that explicitly includes confidence-aware semantic group filtering before substitution. Current systems also continue with embedding regardless of the confidence of the classifier or have used a binary synonym-matching criterion that is unable to control polysemantic adjectives. The CASGC framework directly provide solution for this issue by using the margin difference between competing group scores as a signal to determine whether an adjective position is suitable for embedding the secret bit.

III. Problem Statement

A. Formal Problem Formulation

Let $S = W_1 W_2 \dots W_n$ be a cover phrase containing a target adjective W_t modifying a head noun W_h. The steganography encoder E chooses a substitute W_t' from a set of candidates $C(W_t, S)$ to encode a secret bit sequence $b \in \{0,1\}^k$. The decoder D recovers b by determining which equivalence class in $C(W_t, S)$ includes W_t'. $D(E(S,b)) = b$ for every valid S and b indicate that the system is correct.
When W_t is semantically ambiguous, it can belong to several semantic categories, making the partition of $C(W_t, S)$ into bit-encoding classes unstable. A small contextual perturbation predicts a new group, altering the class boundaries and resulting in $D(E(S,b)) \neq b$.

B. Semantic Ambiguity in Adjectives

Adjectives are particularly sensitive to polysemy because they have a significant dependency on the semantic frame produced by the head noun. Table III shows six polysemantic adjectives and their predicted semantic group changes between literal sentences (Sentence A) and abstract or figurative sentences (Sentence B). For example, "round" is appropriately classified as form in the "The earth is round," but it should be mapped to size_quantity in "That is a round number." This contextual dependency shows that word-level classification is insufficient for accurate steganographic embedding.

TABLE I
Context-Sensitive Adjective Pairs: Same Adjective, Two Semantic Contexts (Part 1)

Adj	Sentence A	Exp. Group A	Pred. Group A
round	The earth is round.	shape	shape
light	This is a light bag.	physical_property	physical_property
hard	The table is hard.	physical_property	opinion_evaluation
deep	He fell into deep water.	physical_property	state_intensity
cold	The water is cold.	physical_property	mental_emotion
bright	The room is bright.	colour	colour

C. Failure Modes in Existing Systems

- **Unreliable substitution:** When an adjective is polysemantic, for example "geometric" or "ovoid" in the BERT top-k may belong to semantically conflicting groups.

TABLE II
CONTEXT-SENSITIVE ADJECTIVE PAIRS: SAME ADJECTIVE, TWO SEMANTIC CONTEXTS

Adj	Sentence B	Exp. Group B	Pred. Group B
round	That is a round number.	size_quantity	size_quantity
light	This is a light mean.	size_quantity	size _quantity
hard	That was a hard question.	opinion_evaluation	opinion_evaluation
deep	She is in deep trouble.	state_intensity	state_intensity
cold	He gave me a cold look.	mental_emotion	mental_emotion
bright	She is a bright student.	opinion_evaluation	opinion_evaluation

- **Extraction failure:** the decoder predicts a different semantic group than the encoder used, resulting the wrong bit sequence extraction.
- **Head-word errors:** In sentences containing linking verbs, the predicate adjectives are assigned to the copular verbs rather than grammatical subjects by dependency parsers.
- **Adjective miss:** Standard POS taggers do not recognized participial modifiers like "sleeping" in "sleeping bag" as ADJs, resulting the position is skipped for embedding.

IV. PROPOSED FRAMEWORK

The Confidence-Aware Semantic Group Classification (CASGC) system includes a semantic validation layer between word generation and embedding. The pipeline (Fig.1) is divided into six different phases, which are discussed in detail in the following subsections.

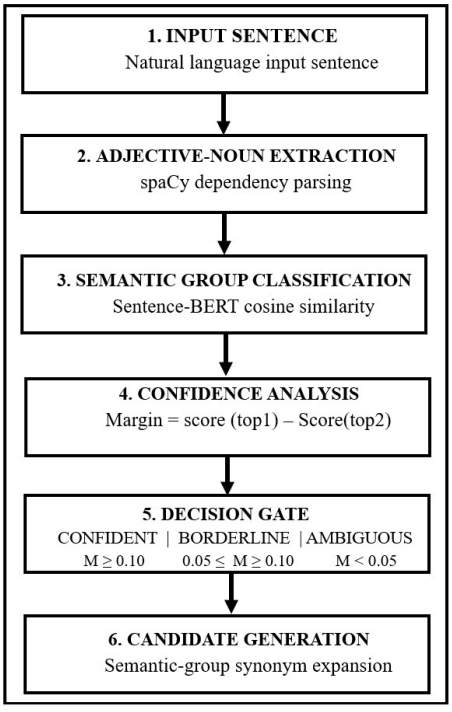

Fig. 1. CASGC pipeline showing adjective extraction, semantic group classification, confidence analysis, ambiguity filtering, and candidate generation

A. Adjective-Noun Extraction

The dependency parser spaCy [16] is used in the extraction module to process each sentence. A token t is regarded a candidate adjective if: (i) `token.pos_ == "ADJ"`; (ii) `token.dep_ == "amod"` (adjective modifier); (iii) `token.dep == "acl" or "relcl"` with a participle function; or (iv) the token appears in a recognized adjectival noun-modifier pattern (for example, "sleeping" in "sleeping bag"). This enhanced criterion identifiers adjectives that regular POS taggers overlook.

Head-word recovery follows the predicate-adjective correction rule: if `t.dep_` is `"acomp" or "attr"` and $t's$ syntactic head h is a copular verb (is, are, was, were, seem, feel, look, appear, become), the semantic head is assigned to the grammatical subject of h. This corrects the errors found in Phase 1 for "The earth is round" (syntactic head: "is"; corrected semantic head: "earth")

B. Semantic Group Classification

Sentence-BERT is used by the classification module to encode the entire sentence into a dense vector [14]. Age_times, colour, material, mental_emotional, opinion_evaluation, origin, ordinal_position, physical_property, purpose_function, shape, size_quantity,and state_intensity are the twelve semantic group labels that are defined. The centroid of twenty WordNet-derived synonyms, each stored separately and averaged, represents each group G. Given the adjective a in sentence S, group $G's$ similarity score is:

$$\text{score}(a, G_i \mid S) = \text{cosine}\left(\text{Enc}(a, S), \text{Prototype}(G_i)\right) \quad (1)$$

where $\text{Enc}(a, S)$ is the Sentence-BERT [14] encoding of S. The top-ranked group $G^* = \text{argmax}_i \, \text{score}(a, G_i \mid S)$ is selected as the predicted semantic group.

C. Confidence Analysis

The confidence margin M is defined as:

$$M = \text{score}(a, G^*(1) \mid S) - \text{score}(a, G^*(2) \mid S) \quad (2)$$

where $G^*(1)$ and $G^*(2)$ are the highest and second highest scoring groups respectively. A small M suggests that two or more interpretations are almost equally likely, whereas a large M shows a strong commitment to one interpretation.

D. Ambiguity Filtering

The observed margins naturally clustered into three regions, according to an empirical examination of the margin distribution over all 48 Phase 3 test cases. Because it is the smallest margin at which all test instances in this corpus were correctly classified, the threshold $M = 0.10$ was selected as the border for the Confident tier; below this value, at least one misclassification occurred. Cases where the top two group scores differ by less than five percentage points of cosine similarity are separated by the lower threshold $M = 0.05$, indicating a near-tie in which the classification outcome is practically unreliable regardless of which group scores highest. A larger

dataset should be used to validate these thresholds, which were established using the existing corpus. There are three distinct decision tiers:

- **Confident ($M \geq 0.10$):** Proceeding embedding. In Phase 3 all eight Confidence cases were accurately classified, supporting the use of this tier as a reliable embedding indicator.
- **Borderline ($0.05 \leq M < 0.10$):** Proceeding embedding with caution; the candidate position may be used in a single-round substitutions scenarios but excluded from later rounds if embedding in multi-round substitution which may cause risk for naturalness.
- **Ambiguous ($M < 0.05$):** Embedding is rejected. The adjective positions is skipped, preventing semantic drift and extraction errors at that location.

E. Candidate Generation

The candidate generator extracts the synonym expansion list for a Confident or Borderline predication G^* and filters candidates based on three criteria: (i) POS consistency (the ADJ tag must be preserved after substitution); (ii) cosine similarity to the original adjective's embedding must exceed $T = 0.5$, (a threshold set to maintain semantic closeness while excluding synonyms that are near antonyms or belong to a different part of speech spectrum in context); and (iii) the candidate must not be present in a stop-list of tokens known introduce grammatical or collocational violations in the sentence structure. In alphabetical order of the candidate tokens, the output set $C(a, G^*, S)$ is divided into two or more subsets, one for each secret bit value.

F. Multi-Round Substitution

Multi-round embedding improves steganographic capability by iteratively applying the substitution pipeline to the same cover text. In Round 1, the framework finds all Confident and Borderline adjective places in the original sentence S, selects candidates for each, and creates a stego sentence S_1 to encapsulate the first portion of the secret bit sequence. In Round 2, the entire pipeline (parsing, classification, and confidence analysis) is applied to S_1 instead of S to account for any semantic shift caused by Round 1's substitution.
For example "She wore a bright, light jacket," Round 1 may substitute "bright" $\rightarrow$ "vivid" (Confidence, embedding bit 1) to produce "She wore a vivid, light jacket." Round 2 then reparse this stego sentence, replacing "light" if it now has a Confident margin and encoding a second bit. Positions identified as Ambiguous in Round 1 are removed from Round 2 to avoid cascading errors, which occur when an uncertain substitution in one round affects the semantic context of the sentence, causing more misclassifications in following rounds.

V. Experimental Evaluation

Experiments were conducted in three different phases. Phase 1 is carried out to assess the basic extraction and classification process using eleven testing examples. Each sentence was supposed to contain exactly one unambiguous adjective-noun phrase from a different semantic category, with no adjective repeated in all sentences and no sentence geared toward a specific classification conclusion. This one-group-per-sentence methodology confirms independent coverage of all twelve groups and avoid inflating accuracy with easy, high-frequency adjective [17]. Phase 2 introduced the confidence gate and evaluated refined group definitions on thirteen sentences, including two more context-sensitive cases. Phase 3 test six polysemantic adjectives ("round", "light", "hard,", "deep,", "cold,", "bright,") using four minimal_pair provides only the adjective and head noun; full_sentence provides the complete natural sentence; sentence_relation pairs the target sentence with a related one to provide the premise that stronger sentential context regularly improves contextualized word representation [18]. Each adjective was tested using two opposing sentences (Sentence A: literal/physical; Sentence B: abstract/figurative), as a result in 12 test cases per method, and 48 overall in Phase 3. It is acknowledged that this corpus is limited; therefore, the findings should be interpreted as indicators of framework behaviour rather than a final performance assessment.
The sentence-transformers/all-MiniLM-L6-v2 [14] embedding model was used. SpaCy was used for dependency parsing in en_core_web_sm [16]. Group prototype vectors were created using twenty WordNet synonyms per group, which were encoded separately.

A. Initial Pipeline Evaluation

As illustrated in Table III, eleven cases were tested and nine of them were identified. Two systematic errors were discovered: (1) a predicate-adjective head word mismatch (sentence 9 and 10), in which the parser returned the copular verb as head, and (2) a participial adjective miss (sentence 8), in which "sleeping" was not marked as ADJ. Both faults are systematic, not random, and were addressed using the upgraded extraction module described in Section IV-A. While the sample size is tiny, the controlled one-group-per-sentence design means that each error corresponds to a specific, addressable failure mode.

B. Phase 2: Post-Fix Evaluation with Confidence Gate

As demonstrated in Tables IV and V, after applying the extraction fixes, 11 of 13 extended cases were correctly classified. The confidence gate correctly identified "heavy" ($M = 0.0044$) and "sleeping ($M = 0.0431$) as ambiguous, hence preventing unreliable embeddings. Context-sensitive failures persisted for "round" (predicted shape instead of size_quantity) in "That is a round number") and "light" (predicted colour instead of size_quantity in "This is a light meal"), showing the embedding model's bias toward the prevailing meaning of each word.

C. Context Type Evaluation

The Table VI illustrates, full-sentence context had the best accuracy (83.33%), indicating that entire sentential context provided the richest disambiguation signal i this assessment. The disambiguation_prompt method performed the poorest (50.00%), showing that explicit metalinguistic instructions

TABLE III
INITIAL TESTING RESULTS (PHASE 1)

#	Sentence	Head	Adj	Exp.Grp	Pred.Grp	Margn	Result
1	I have an old car.	car	old	age_time	age_time	0.2154	Correct
2	We live on a round earth.	earth	round	shape	shape	0.1877	Correct
3	She bought a wooden chair.	chair	wooden	material	material	0.1470	Correct
4	He wore a blue jacket.	jacket	blue	colour	colour	0.0792	Correct
5	They saw a beautiful garden.	garden	beautiful	opinion	opinion	0.786	Correct
6	I met a French teacher.	teacher	French	origin	origin	0.1367	Correct
7	He carried a heavy box.	box	heavy	physical	physical	0.0482	Correct
8	We bought a sleeping bag.	N/A	N/A	N/A	N/A	N/A	Fail
9	The earth is round.	earth*	round	shape	shape	0.2364	Head Error
10	She felt happy today.	she*	happy	mental	mental	0.2711	Head Error
11	This is a small room.	room	small	size_qty	size_qty	0.2681	Correct

push the Sentence-BERT embedding away from normal usage context, degrading classification.

D. Context-Sensitive Pair Analysis

This section analyzes the six polysemanitc adjective pairs using the full-sentence context methods, which produced the highest overall accuracy. Table III (Section III) presented the pairs and their summary results. Tables VII, VIII, IX show the pre-adjective margin ranges, classifier status values, and context sensitivity verdicts. Given the limited proof-of-concept corpus, the patterns described are representative of the framework's behavior and should not be considered statistically decisive.

Tables VII, VIII and IX include the complete data for all six adjective pairs. Table VII covers Sentence A (literal/physical usage); Tables VIII and IX cover Sentence B (abstract/figurative usage), with the sentence and group prediction data separated from the margin and sensitivity data for readability purposes.

For "round," the framework predicted form for Sentence A (Confident, margin 0.1173 - 0.1529) and size_quantity for Sentence B (Borderline, 0.0184 - 0.0947), indicating that the full-

TABLE IV
SENTENCES, EXPECTED AND PREDICATED GROUPS (A)

#	Sentence	Head	Adj	Exp.Group	Pred.Group
1	I have an old car.	car	old	age_time	age_time
2	We live on a round earth.	earth	round	shape	shape
3	She bought a wooden chair.	chair	wooden	material	material
4	He wore a blue jacket.	jacket	blue	colour	colour
5	They saw a beautiful garden.	garden	beautiful	opinion_eval	opinion_eval
6	I met a French teacher.	teacher	French	origin	origin
7	He carried a heavy box.	box	heavy	physical_prop	physical_prop
8	We bought a sleeping bag.	bag	sleeping	purpose_func	purpose_func
9	The earth is round.	earth	round	shape	shape
10	She felt happy today.	she	happy	mental_emot	mental_emot
11	This is a small room.	room	small	size_qty	size_qty
12	That is a round number.	number	round	size_qty	shape
13	This is a light meal.	meal	light	size_qty	colour

TABLE V
MARGIN, STATUS AND CORRECTNESS (B)

#	Margin	Status	Correctness
1	0.1842	Confident	✓
2	0.2060	Confident	✓
3	0.1751	Confident	✓
4	0.0397	Confident	✓
5	0.0593	Confident	✓
6	0.1498	Confident	✓
7	0.0044	Ambiguous	✓
8	0.0431	Ambiguous	✓
9	0.2517	Confident	✓
10	0.2794	Confident	✓
11	0.2426	Confident	✓
12	0.1467	Confident	×
13	0.0808	Borderline	×

TABLE VI
MARGIN, STATUS AND CORRECTNESS (B)

Context Type	Total	Correct	Accuracy	Rank
full_sentence	12	10	83.33%	Best
minimal_pair	12	9	75.00%	Good
sentence_relation	12	8	66.67%	Moderate
disambiguation_prompt	12	6	50.00%	Worst

TABLE VII
DETAILED SENTENCE A RESULTS (LITERAL/PHYSICAL USAGE)

Adj	Sentence A	Exp.Grp A	Pred.Grp A	Margin A	Status A	Correct A
round	The earth is round.	shape	shape	0.1173 - 0.1529	Confid.	Yes
light	This is a light bag.	physical_p	physical_p	0.0109 - 0.0285	Ambig.	Yes
hard	The table is hard.	physical_p	opinion_ev	0.0097 - 0.1936	Ambig.	No
deep	He fell into deep water.	physical_p	state_int	0.0738 - 0.1912	Border.	No
cold	The water is cold.	physical_p	mental_em	0.0021 - 0.0873	Ambig.	No
bright	The room is bright.	colour	colour	0.0216 - 0.798	Ambig.	Yes

TABLE VIII
SENTENCE B (ABSTRACT USAGE: SENTENCE AND GROUP PREDICTIONS)

Adj	Sentence B	Exp.Grp B	Pred.Grp B
round	That is a round number.	size_quantity	size_quantity
light	This is a light meal.	size_quantity	size_quantity
hard	That was a hard questions.	opinion_eval	opinion_eval
deep	She is in deep trouble.	state_intensity	state_intensity
cold	He gave me a cold look.	mental_em	mental_emo
bright	She is a bright student.	opinion_eval	opinion_eval

sentence embedding provides enough context to distinguish these uses in this case. For "light," all sentence were correctly classified, but both margins were ambiguous (less than 0.05); the confidence gate would reject both positions, as intended, because an uncertain correct prediction is not a reliabel anchor for steganographic encoding.

Sentence A ("The table is hard") was misclassified as opinion_eval for the word "hard," showing the embedding model's dominant-sense bias. Because both sentences receive the same predicated label, no context sensitivity is achieved. For "deep," both sentences yielded state_intesity as the predicted label regardless of context, again precluding sensitivity. For "cold," the predictions vary between sentences (Context Sensitive = Yes in Table IX), but both are Ambiguous-tier, therefore CASGC rejects them both. Sentence A was properly predicted as colour and Sentence B as opinion_eval with a Confident margin (0.0028 - 0.1296) for "bright," implying that this pair may exhibit dependable context-sensitive behavior in this assessment.

TABLE IX
SENTENCE B (MARGINS, STATUS AND CONTEXT SENSITIVITY)

Adj	Margin B	Status B	Correct B	Context Sensitive?
round	0.0184 - 0.0947	Borderline	Yes	Yes
light	0.0027 - 0.0563	Ambiguous	Yes	Yes
hard	0.0469 - 0.2096	Ambiguous	Yes	No
deep	0.1078 - 0.1508	Confident	Yes	No
cold	0.0377 - 0.1183	Ambiguous	Yes	Yes
bright	0.0028 - 0.1296	Confident	Yes	Yes

On this small corpus, three preliminary observations emerge: full-sentence context appear to support disambiguation for some polysemantic adjectives; confident margins appear to identify more reliable embedding positions than ambiguous ones; and dominant-sense bias i the embedding model remains the initial source of misclassification.

E. Confidence Distribution Analysis

TABLE X
CONFIDENCE DISTRIBUTION ACROSS PHASE 3 TEST CASES

Status	Count	Porp.	Recommendation
Confident($M \geq 0.10$)	8	66.7%	High reliability; proceed embedding
Borderline ($0.05 - 0.10$)	2	16.7%	Embed with caution
Ambiguous ($M < 0.05$)	2	16.7%	Reject; unreliable

As shown in Table VII, all Confident cases in the phase 3 test set were correctly identified, whereas Ambiguous cases were mostly inaccurate. This pattern supports the premise that margin-based filtering may find reliable embedding places without human intervention; nonetheless, the sample size is insufficient to draw strong generalizable conclusions.

F. Phase Comparison Summary

Table XI summarizes classification accuracy during all three experimental phases, allowing each compoent's incremetnal contribution to be evaluated: the extraction correctios used in Phase 1-2 and the full-sentence context representation add in Phase3.

TABLE XI
ACCURACY ACROSS EXPERIMENTAL PHASES

Phase	Total	Correct	Accuracy	Notes
Phase 1 - Initial	11	9	81.8%	No gate; 2 head errors
Phase 2 - Baseline	12	6	50.0%	Ambiguous adjective failures
Phase 2.6 - Refined	12	8	66.7%	Gate introduced
Phase 3 - Full Sentence	12	10	83.3%	Best; +16.7% gain

VI. DISCUSSION

A. Semantic Preservation

The findings indicate that confidence-aware filtering could improve semantic preservation in adjective substitution by avoiding instances where the classifier is uncertain. The framework rejects ambiguous locations ($M < 0.05$) to prevent substitutions that could cause semantic drift. This could be useful for multi-round embedding (the iterative process described in Section IV-F), where errors can cascade: an incorrect substitution in Round 1 corrupts the sentence fed

to Round 2, and the resulting semantic change may result in further misclassifications in that round, with errors compound over passes [9]. In this examination, the full-sentence context representation context representation beat all other alternatives, which is consistent with the distributional semantic concept that sentential context shapes word meaning [15].

B. Extraction Reliability

The predicate-adjective head recovery module corrects a key difference between syntactic and semantic headedness in copular forms. Standard dependency parsers use the Universal Dependencies annotation scheme, which attaches the adjective complement to the copular verb. However, for steganographic purposes, the noun being specified is the semantically relevant anchor. The correction rule restored both predicate-adjective cases in Phase 1 without using regressions. The enhanced adjective extraction heuristic also recovered participial modifiers like "sleeping" in "sleeping bag," which are a type of prenominal modifier that is generally classified as a verb or a noun by standard POS tags.

C. Embedding Model Limitations

The biggest remaining constraint is the embedding model's preference for the prevailing lexical sense of polysemous adjectives. There are two potential mitigations: (i) fine-tuning the embedding model on a curated adjective disambiguation corpus, and (ii) substituting cosine similarity with a discriminative classifier trained on group-labeled adjective-sentence pairs. The latter strategy is likely to produce improved accuracy at the expense of requiring labeled training data.

D. Steganography Capacity Implications

Rejecting ambiguous spots limits the number of potential embedding locations each sentence, lowering raw steganographic capacity. This may be a desirable trade-off: if unreliable situations cannot be consistently decoded, they contribute little usable capacity in practice. According to this rationale, the effective capacity (the number of positions that can potentially be properly stored and decoded)of CASGC may be greater than that of indiscriminate embedding systems, though empirical validation on a wider scale is required to prove this.

VII. Conclusion

This study described the Confidence-Aware Semantic Group Classification (CASGC) framework for adjective substitution in linguistic steganography. The framework solves the semantic ambiguity problem that plagues existing synonym-substitution and MLM-based methods by introducing a semantic validation layer with four components: context-sensitive adjective classification, a confidence margin gate, semantic group candidate generation, and predicate-adjective head recovery.
In a proof-of-concept evaluation on a small, controlled corpus, full-sentence context obtained 83.33% classification accuracy across twelve polysemous adjective test cases, representing a 16.7% improvement over the refined baseline. The confidence gate identified and rejected ambiguous locations ($M < 0.05$) that could potentially compromise the steganographic channel. The surprising finding that disambiguation degrades performance implies that natural sentential context is preferable to metalinguistic explanations as classifier input. While these findings are intriguing, the tiny size of the corpus requires larger-scale assessment before making significant generalization claims.

References

[1] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," IBM Systems Journal, vol. 35, no. 3–4, pp. 313–336, 1996.

[2] K. Bennett, "Linguistic steganography: Survey, analysis, and robustness concerns for hiding information in text," CERIAS Tech. Report, Purdue Univ., 2004.

[3] J. T. Brassil, S. Low, N. F. Maxemchuk, and L. O'Gorman, "Electronic marking and identification techniques to discourage document copying," IEEE J. Select. Areas Commun., vol. 13, no. 8, pp. 1495–1504, Oct. 1995.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, 2019, pp. 4171–4186.

[5] W. Yang, K. Hu, and J. Liu, "RNNSTEGA: Linguistic steganography based on recurrent neural networks," IEEE Access, vol. 7, pp. 37029–37039, 2019.

[6] R. Chapman and M. Davida, "Hiding the hidden: A software system for concealing ciphertext as innocuous text," in Proc. 1st Int. Conf. Information and Communications Security, 1997, pp. 335–345.

[7] Z. Xiang, J. Luo, S. Yang, and H. Huang, "Linguistic steganalysis using the features derived from synonym frequency," Multimedia Tools and Applications, vol. 78, no. 17, pp. 25253–25270, 2019.

[8] Y. Fang, Z. Zeng, B. Li, L. Liu, and W. Zhang, "Text steganography using language models," in Proc. IEEE ICME, 2020, pp. 1–6.

[9] Z. M. Ziegler, Y. Deng, L. Huang, and A. M. Rush, "Neural linguistic steganography," in Proc. EMNLP-IJCNLP, Hong Kong, 2019, pp. 1210–1215.

[10] H. Ding, X. Liao, J. Huang, and K. Chen, "Semantic preserving linguistic steganography via adversarial example," IEEE Trans. Inf. Forensics Security, vol. 17, pp. 1982–1997, 2022.

[11] P. Ke, H. Ji, S. Liu, X. Zhu, and M. Huang, "SentiLARE: Sentiment-aware language representation learning with linguistic knowledge," in Proc. EMNLP, 2020, pp. 6975–6988.

[12] S. Dai, S. Jin, C. Ling, and W. Hu, "Towards near-imperceptible steganographic text," in Proc. ACL, Florence, Italy, 2019, pp. 5133–5142.

[13] X. Zheng, Z. Meng, and X. Zhang, "Graph-based linguistic steganography using synonym substitution," in Proc. IEEE ICASSP, Toronto, Canada, 2021, pp. 2490–2494.

[14] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. EMNLP-IJCNLP, Hong Kong, 2019, pp. 3982–3992.

[15] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," J. Artif. Intell. Res., vol. 37, pp. 141–188, 2010.

[16] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," Explosion AI, 2017. [Online]. Available: https://spacy.io

[17] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge: Cambridge Univ. Press, 2008.

[18] M. Peters et al., "Deep contextualized word representations," in Proc. NAACL-HLT, New Orleans, LA, 2018, pp. 2227–2237.