Diagnosing Overgeneration in Biomedical Text Simplification: What Categories can be Detected with(out) Sources?

Anonymous ACL submission

Abstract

Text simplification systems can produce fluent, mostly faithful outputs that still contain extraneous material such as assistant chatter, repetition, or unsupported biomedical explanations. These errors are difficult to evaluate: some are visible from the output alone, while others require source grounding. Building on a pilot trailing content detector, we construct an audited, LLM-assisted sentence-level taxonomy resource for biomedical overgeneration, labeling 32,505 surfaced prediction sentences and adding a deterministic template-echo correction. The released large-scale splits contain 39,586 positive sentence labels with categories, subcategories, and a *detectable-without-source* flag. Three categories account for 95.1% of released positive labels, 82.1% are detectable without source access, and the detectability flag agrees with human labels on 646 positives in 92.7% of cases ($\kappa = 0.837$). On 2,688 human-labeled sentences, generic LLM judges plateau at F1 = 0.66–0.73, while a fine-tuned DeBERTa-v3 output-only detector achieves the highest precision (0.852) with competitive F1 (0.663; overlapping confidence intervals) at zero API cost. These results establish reference-free overgeneration detection as a useful benchmark setting while preserving the need for source-aware grounding.

1 Introduction

Large Language Models (LLMs) exhibit impressive performance on all NLP Benchmarks. As generative predictions are evaluated on the intersection with references, the evaluation scores are insensitive to overgeneration or other extraneous output. As a result, submissions to NLP benchmarks may contain varying degrees of overgeneration. Figure 1 shows two real CLEF 2025 SimpleText Track examples (Ermakova et al., 2025). In Panel A, a leaked instruction is buried inside an otherwise plausible biomedical simplification. In Panel B, the system preserves the adverse-event finding but adds unsupported claims about where the trial was conducted and where the study was published. Both errors are fluent enough to withstand a quick manual inspection and are nearly invisible to standard overlap evaluation. At the abstract level, removing the detected overgenerated spans changes SARI (Xu et al., 2016; Alva-Manchego et al., 2019) from 40.4

A. Leaked instruction inside fluent output

We studied 196 stroke patients with anxiety in three trials. One small study gave 21 stroke survivors a relaxation CD to use for four weeks. One study used a relaxation CD and found that anxiety decreased after three months. The second trial involved 81 participants who were given either paroxetine, paroxetine with therapy, or standard care. The anxiety scores were 5.4, 3.8, and 12.8, which is a significant difference. The third trial gave 94 stroke patients either buspirone or standard care. The anxiety levels were significantly lower in one group (6.5) than the other (12.6). Many people taking paroxetine had side effects like nausea or dizziness, but fewer people taking buspirone had these problems. We don't know how long side effects lasted. The relaxation therapy study had no bad side effects. (I chose 'rephrase' as the internal simplification strategy to simplify the sentence while maintaining its original meaning.) The evidence wasn't very reliable. The studies had few participants, especially the one on relaxation therapy. The drug studies didn't give enough information to know if they were reliable. The relaxation therapy study had some problems that might have affected the results. We don't have enough information to know how to treat anxiety after a stroke. More research is needed to find the best treatments for anxiety after a stroke.

B. Source-required unsupported addition

Merged source-to-output diff:

Moderate-dose vitamin D, compared to placebo, may improve vitamin D status and may result in little to no difference in the achieved parathyroid hormone level ~~(achieved 25-hydroxyvitamin D level: mean difference 13.60 ng/mL, 95% confidence interval 7.94 to 19.26; achieved parathyroid hormone level: −6.60 pg/mL, 95% confidence interval −17.12 to 3.92; 1 study, 79 participants; low-certainty evidence).~~ 1 study, 79 participants, found low-certainty evidence for the benefits of moderate-dose vitamin D. The trial reported no adverse events in the moderate-dose vitamin D arm~~, but did not provide~~ . The trial did not provide any information on adverse ~~events~~ event in the placebo arm. It was the first time the trial had been conducted in the U.S. and the second time it had been done in the UK. There were no data on fractures, all-cause mortality, bone density change, health-related quality of life, and muscle strength. There were also no data for bone density changes, quality-of-life changes, or muscle strength changes. The study was published in the Journal of the American College of Obstetrics and Gynecology.

Figure 1: Two anonymized shared-task examples. Panel A is detectable from the output alone; Panel B requires the source because the generated sentences about U.S./UK trial conduct and journal publication are unsupported. Gray strike-through marks source-only deletions, black marks shared text, and green marks prediction-only additions.

Code and data are available at: https://anonymous.4open.science/r/diagnosing-overgeneration-D58F/README.md.

to 40.8 for Panel A's abstract and from 30.6 to 29.6 for the full abstract containing Panel B. The score barely changes, and can even move in the wrong direction. This captures the central difficulty of overgeneration detection in text simplification: a system can be mostly faithful while still adding extraneous content that blends into the output. Across the corpus, this insensitivity persists: removing all detected trailing spans across 5,750 affected outputs changes SARI by a median of only 0.20 points, with 85.8% changing by less than one point.

We study this failure mode in biomedical text simplification, where LLM-based systems are increasingly used to make dense clinical writing accessible without changing its meaning (Li et al., 2024). The SimpleText shared task uses Cochrane abstracts as a central testbed (Ermakova et al., 2025; Bakker and Kamps, 2024), and its long-form outputs make subtle generation artifacts easy to miss.

Most hallucination work targets factual inconsistency in summarization or question answering (Maynez et al., 2020; Kryscinski et al., 2020; Li et al., 2023). Text simplification has a different failure profile: systems must preserve meaning while reducing complexity, so errors include not only contradiction but also *overgeneration*, content that exceeds the intended simplification boundary. In biomedical abstracts, this can appear as assistant chatter, leaked prompt framing, unsupported glosses, repetition, or template residue appended to otherwise plausible output.

An initial pilot experiment with overgeneration detection showed that trailing-content alignment can surface overgeneration in shared-task outputs. This turns the submissions to an NLP Benchmark into a valuable resource of realistic types of overgeneration as they occur in our current LLM-powered NLP pipelines. Candidate detection alone, however, does not explain what kinds of failures are being surfaced, which ones are visible without the source, or whether generic judges notice them.

This paper aims to move from detection to diagnosis. Starting from 32,505 prediction sentences flagged by the trailing-content detector, and adding a deterministic correction for full-template echoes missed by trailing alignment, we build an audited, LLM-assisted sentence-level taxonomy resource grounded on real NLP benchmark submissions. Each surfaced case is labeled with an interpretable category and a *detectable-without-source* flag, enabling both large-scale analysis and reference-free evaluation.

The central finding is that the surfaced overgeneration is not a monolithic source-grounding problem: most surfaced cases are output-intrinsic artifacts, while unsupported biomedical additions remain source-required. This makes reference-free overgeneration detection a meaningful task in itself, and we can reuse submissions from a text simplification benchmark to create a new benchmark for overgeneration detection. More generally, this motivates hybrid simplification metrics that directly penalize output-intrinsic artifacts while routing grounding-sensitive cases to source-aware evaluation.

We make four contributions:

- We construct an audited, LLM-assisted sentence-level taxonomy resource over representative NLP benchmark submissions, including primary categories, subcategories, and split-aware metadata.
- We introduce and validate a *detectable-without-source* annotation that distinguishes output-intrinsic overgeneration from source-required cases, creating a new benchmark setting for reference-free overgeneration detection.
- We use this resource to characterize the anatomy of overgeneration at scale across model families and teams, showing that different systems fail in systematically different ways.
- We benchmark both generic LLM judges and a specialized output-only detector on the new setting, showing that the resource supports meaningful evaluation of reference-free overgeneration detection.

2 Related Work and Benchmark Context

2.1 Hallucination and Faithfulness in NLG

Factual inconsistency in neural generation has been studied mainly in summarization. Maynez et al. (2020) distinguish *intrinsic* hallucinations, which contradict the source, from *extrinsic* ones, which introduce unsupported content. Later work developed entailment- and QA-based detectors (Kryscinski et al., 2020; Laban et al., 2022; Wang et al., 2020; Durmus et al., 2020) and broad benchmarks such as HaluEval and TRUE (Li et al., 2023; Honovich et al., 2022).

These resources largely assume that the error is a factual claim that can be checked against the

source. Simplification overgeneration also includes structurally extraneous text, such as leaked rationales or assistant chatter, that may be coherent or source-consistent but should not appear in the output.

2.2 Text Simplification Evaluation

Text simplification evaluation relies on overlap metrics such as SARI (Xu et al., 2016) and BLEU (Papineni et al., 2002), plus readability and learned metrics such as LENS (Maddela et al., 2023) and BERTScore (Zhang et al., 2020). Yet automatic measures correlate poorly with human meaning-preservation judgments (Davari et al., 2024), and they are not designed to penalize extraneous appended content that leaves the faithful portion intact.

Existing simplification error taxonomies target within-text distortions, such as insertion, deletion, and substitution errors (Devaraj et al., 2022), or structured source misrepresentation (Vendeville et al., 2025). We address the complementary case where a model goes beyond the intended simplification boundary through leakage, unsupported elaboration, or repetition.

2.3 From Surfacing to Diagnosing Overgeneration

We conducted initial pilot experiments (Anonymous, 2026) demonstrating that trailing-content alignment can surface overgeneration in SimpleText 2025 biomedical simplification outputs (Ermakova et al., 2025). The pilot was a proof of concept to quantify the prevalence of overgeneration in NLP benchmark submissions. It provided binary labels and source-aware baselines, but not the structure needed for reference-free evaluation: labels were non-taxonomic, overgeneration was treated as one phenomenon, and no annotation indicated whether a case was detectable from the output alone. We add that missing structure with a sentence-level taxonomy, a *detectable-without-source* annotation, metric-blindness analysis, and reference-free detection experiments.

3 Resource: Sentence-Level Taxonomy for Overgeneration

We extend initial trailing-content detection (Anonymous, 2026) with sentence-level *category* labels and a *detectable-without-source* flag, thereby turning surfaced binary cases into a taxonomy resource for structured, reference-free analysis.

3.1 From Candidate Extraction to Sentence-Level Labeling

We detect overgeneration by aligning source and prediction tokens with Python's `SequenceMatcher`, which enables us to identify all insertions of text in the prediction. After the last shared token block, the remaining prediction text is extracted as a *trailing segment*; spans shorter than 25 characters are filtered out. We apply the taxonomy layer to the **Cochrane subset**, the primary biomedical simplification target of CLEF 2025 SimpleText Track (Bakker and Kamps, 2024). This yields **32,505 classified candidate sentences** with detected trailing content from 453 Cochrane reviews and 42 system runs.

The shared-task submissions correspond to abstract-level simplification outputs, but the available data are already aligned as sentence-level prediction units. We therefore label those sentence units directly, preserving abstract identifiers and neighboring sentence context in the released data. The resource is sentence-level in annotation unit, but document-level in origin.

Each candidate is labeled using the finalized taxonomy (§3.2) over the **full prediction sentence**, not just the trailing span. This is necessary because the extracted suffix may miss earlier leakage or mixed failure signals. The extractor is therefore a candidate generator rather than an exhaustive detector; audit details are in §4.3.

3.2 Overgeneration Taxonomy

The final taxonomy was derived by grouping recurring patterns observed in exploratory samples of surfaced overgeneration, then finalized before full-scale labeling so that all candidates were annotated with one stable label space. It defines five released positive categories. GENERATION_FAILURE marks unusable output, including truncation, format collapse, and prompt-template loops. LEAKED_INSTRUCTIONS marks task framing, assistant chatter, or internal simplification rationale that appears in the output. UNGROUNDED_INJECTION marks added factual or explanatory content that is not supported by the source sentence. REPETITIVE_CONTENT marks redundant words, phrases, or sentences that repeat within the prediction. GROUNDED_OVERGENERATION marks source-supported extra detail that goes beyond a concise simplification but is not factually unsupported. Table 1 gives sentence-level examples.

Category	Examples
LEAKED_INSTRUCTIONS	**Source sentence:** Studies in general were small with a median sample size of 42. **Simplified output:** Most studies were small, with 42 participants on average. (I chose 'rephrase' as the internal simplification strategy)
GENERATION_FAILURE	**Source sentence:** We included 50 RCTs with 1407 participants. **Simplified output:** . Output: We included 50 RCTs. Output: We included 50 RCTs. Output: We included 50 RCTs.: : : : : : : : : : : -: -: -: -: -: -:
UNGROUNDED_INJECTION	**Source sentence:** The harm profiles of the antibiotic regimes featured in this review remain unknown. **Simplified output:** The harm profiles of the antibiotic regimes featured in this review remain unknown. Harm profiles: information about possible side effects or risks of the treatments.
REPETITIVE_CONTENT	**Source sentence:** Of the statins we evaluated, only rosuvastatin seemed to be associated with a reduced incidence of VTE. **Simplified output:** semantically, rosuvastatin appears to be associated with VTE trim. rosuvastatin appears to be associated with VTE trim.
GROUNDED_OVERGENERATION	**Source sentence:** The safety profile is unknown. **Simplified output:** The safety profile is unknown. We don't know how safe it is.

Table 1: Sentence-level examples for the five released positive categories. These are real released sentence units from abstract-level shared-task submissions; the resource labels the provided sentence units while retaining document context.

Two labels serve accounting roles in the released binary data. NON_OVERGENERATION maps to None (not overgeneration), while OTHER (27 cases, predominantly declarative-to-interrogative transformations) is absorbed into UNGROUNDED_INJECTION. The deterministic template-echo correction in §4.6 then adds 7,502 missed prompt-template positives; extractor false positives are retained as None. Table 2 reports the resulting released positive-label distribution for the large-scale train/dev/automated-test splits.

When a sentence exhibits multiple phenomena, the guardrails in §4.2 select the highest-priority sentence-level failure rather than labeling only the final trailing clause. The escape labels NON_OVERGENERATION and GROUNDED_OVERGENERATION prevent extraction artifacts and source-supported overflow from being forced into unsupported-error categories.

Category	*n*	%
LEAKED_INSTRUCTIONS	17,199	43.4
GENERATION_FAILURE	13,232	33.4
UNGROUNDED_INJECTION	7,211	18.2
REPETITIVE_CONTENT	1,808	4.6
GROUNDED_OVERGENERATION	136	0.3
Total positive labels	39,586	100.0

Table 2: Sentence-level category distribution across released positive labels in the large-scale train/dev/automated-test splits, after remapping and deterministic template-echo correction. None labels are omitted.

3.3 Detectable Without Source

Each sentence receives an operational *detectable-without-source* (DWS) flag indicating whether the overgeneration label can be judged from the output alone. This partitions the resource into a **reference-free** setting and a **source-aware** setting, analyzed in §5.2. We validate this flag against human detectability annotations in §4.4.

3.4 Released Resource

The release contains the full sentence-level dataset, including negative (None) labels. For surfaced candidates, it includes the source and prediction sentence, the trailing span, the faithful prefix, the category, the free-form subcategory, the DWS flag, the model metadata, and the split assignment.

Split assignment is inherited from the tailing-content detection, where Cochrane run-by-abstract examples are partitioned before taxonomy extraction using held-out runs and source abstracts. The released dataset contains 21,057 documents (350,583 sentences) for training, 1,120 documents (18,638 sentences) for development, and 5,608 documents (86,363 sentences) for automated testing, plus a 154-document (2,688 sentences) human-annotated test set. The automated test portion includes held-out-system and held-out-abstract conditions.

4 Resource Construction and Validation

We separate candidate extraction from final annotation, finalize the taxonomy before full-scale labeling, and audit high-risk buckets. Exhaustively hand-labeling every sentence in every abstract-level prediction is infeasible at this scale, so the large-scale labels are LLM-assisted: we support them with existing human annotations, targeted audits, robustness checks, and downstream behavior, but

do not claim exhaustive human-gold taxonomy annotation.

4.1 LLM-Assisted Labeling Pipeline

For each surfaced candidate, we recompute source–prediction alignment on original-cased text so the faithful prefix and trailing span match the raw output; normalized recomputed spans match the stored benchmark spans for all 32,505 candidates. Gemini 2.5 Flash then labels the full prediction sentence with schema-enforced JSON containing a primary category, subcategory, confidence score, explanation, and DWS flag.

4.2 Taxonomy Refinement and Decision Guardrails

We initialized the label space from the pilot manual categories, ran open-set exploratory annotations on sampled candidates, and merged recurring proposals before applying the taxonomy at scale. Guardrails reserve GENERATION_FAILURE for unusable output, prevent coherent assistant wrappers from becoming NON_OVERGENERATION, assign unsupported explanations to UNGROUNDED_INJECTION, and keep source-supported overflow separate.

4.3 Human Audit

We use an independent human-annotated set of 2,688 sentence-level binary labels across 154 system outputs. The taxonomy overlay marked 36 originally positive sentences as NON_OVERGENERATION; we manually reviewed these conflicts and kept 10 as positives while moving 26 to negative. The final gold set therefore contains 646 positives and 2,042 negatives; this binary set is used for judge and detector evaluation. The original annotation file also contains 670 category-bearing positives before this conflict review, used only for the category-agreement audit below.

We also audit the high-risk NON_OVERGENERATION bucket. In a small targeted sample, 14/15 cases were clear or likely acceptable simplifications, supporting its role as an extraction-boundary bucket rather than a sink for missed hallucinations.

4.4 Human Validation of Detectability

Because DWS is central to the resource, we directly validate it on the 646 human-positive sentences. A human annotator saw only the flagged prediction sentence and marked whether the overgeneration was detectable without the source. The taxonomy-derived DWS flag agrees with the human labels in 92.7% of cases (599/646; $\kappa = 0.837$). Treating human DWS=True as reference, the taxonomy-derived flag obtains P = 0.962, R = 0.929, and F1 = 0.945.

Category	*n*	Hum. T	Tax. T	Agr.
LEAKED_INST.	227	97.8	99.6	98.2
GENERATION_FAIL.	176	94.9	97.7	93.8
UNGROUNDED_INJ.	215	13.5	3.3	87.9
REPETITIVE_CONT.	14	85.7	100.0	85.7
GROUNDED_OVERGEN.	6	33.3	16.7	83.3
OTHER	8	62.5	25.0	62.5

Table 3: Human validation of the DETECTABLE-WITHOUT-SOURCE (DWS) flag on the 646 human-positive sentences. "Hum. T" is the percentage labeled DWS=True by the human annotator; "Tax. T" is the taxonomy-derived flag.

Table 3 shows that the pattern by category matches the intended semantics: LEAKED_INSTRUCTIONS and GENERATION_FAILURE are judged source-free by humans in 97.8% and 94.9% of cases, while UNGROUNDED_INJECTION is source-required in 86.5% of cases. This supports DWS as a human-interpretable detectability boundary rather than merely an artifact of LLM labeling.

4.5 Category-Level Agreement with Human Annotations

For category agreement, human positives use four labels: *Ungrounded*, *CoT leakage*, *Repetitive*, and *Other*. The first three map directly to UNGROUNDED_INJECTION, LEAKED_INSTRUCTIONS, and REPETITIVE_CONTENT. The human scheme lacked GENERATION_FAILURE; among 186 human *Other* cases, 140 (75.3%) are assigned to GENERATION_FAILURE (Appendix B).

Among the 484 human-positive sentences whose human label maps to one of these three shared categories, the taxonomy assigns the same category in 82.4% of cases (399/484). Many disagreements occur because the taxonomy has extra labels absent from the human scheme, especially GENERATION_FAILURE and NON_OVERGENERATION. If we restrict the comparison to the 427 sentences where both annotations choose one of the three shared categories, agreement rises to 93.4% (399/427; $\kappa = 0.87$; Appendix Table 7). Agreement is strongest for LEAKED_INSTRUCTIONS (F1 = 0.94) and high-precision for UNGROUNDED_INJECTION (P = 0.99); REPETITIVE_CONTENT is weakest because several human repetition cases are

structurally degenerate outputs assigned to GENERATION_FAILURE.

The remaining disagreements are systematic: the taxonomy prioritizes GENERATION_FAILURE for unusable output and LEAKED_INSTRUCTIONS when assistant-style framing co-occurs with unsupported content. Thus the large-scale labels should be treated as audited LLM-assisted annotations, not fully human-gold labels.

4.6 Template-Echo Correction

During review of training examples labeled None, we identified 7,502 sentences from one anonymized T5-family system that echo the prompt template verbatim. Because these outputs consist entirely of template text, the trailing-content extractor finds no divergence point. We relabel them with a deterministic substring rule documented in the release; the corrected labels are included in the released splits, detector training data, and Table 2.

4.7 Residual Noise and Limitations

The taxonomy is LLM-assisted, not perfect human-gold annotation. Borderline cases remain between useful simplification, grounded overflow, unsupported glosses, repetition, and generation failure. We therefore present the resource as high-coverage and audited, but not noise-free.

5 What the Resource Reveals: Anatomy of Overgeneration at Scale

The taxonomy layer turns candidate extraction into a diagnostic resource: it shows what kind of phenomenon was surfaced and whether it can be recognized from the output alone. Unless otherwise noted, we analyze released positive labels in the large-scale train/dev/automated-test splits (Table 2).

5.1 A Small Number of Modes Explains Most Cases

The first result is that surfaced overgeneration is heterogeneous, but not amorphous. The three largest failure categories account for 95.1% of released positive labels (Table 2): LEAKED_INSTRUCTIONS (17,199 cases; 43.4%), GENERATION_FAILURE (13,232; 33.4%), and UNGROUNDED_INJECTION (7,211; 18.2%).

This ordering matters: treating surfaced overgeneration only as factual hallucination would miss the two largest modes, while treating it only as assistant chatter would obscure source-grounding and generation-collapse components. The smaller categories are also informative: REPETITIVE_CONTENT (4.6%) captures redundant loops, while GROUNDED_OVERGENERATION (0.3%) keeps source-supported overflow out of unsupported-error categories.

Category	*n*	DWS %	SQ %
LEAKED_INSTRUCTIONS	17,199	99.98	0.02
GENERATION_FAILURE	13,232	100.0	0.0
REPETITIVE_CONTENT	1,808	100.0	0.0
UNGROUNDED_INJECTION	7,211	3.8	96.2
GROUNDED_OVERGENERATION	136	4.4	95.6

Table 4: Detectability without source (DWS) by category among released positive labels, and source-required (SQ). Output-intrinsic artifacts are nearly all detectable without source, while grounding-sensitive categories overwhelmingly require the source sentence.

5.2 Most Surfaced Cases Are Output-Intrinsic

The most consequential split is between cases that require the source and cases that do not. Overall, 32,517 of 39,586 released positive labels (82.1%) are marked *detectable without source*; 7,069 (17.9%) require source grounding. Together with the human validation in §4.4, this establishes a large reference-free subproblem inside surfaced overgeneration detection.

Table 4 shows that this split is not arbitrary. GENERATION_FAILURE, REPETITIVE_CONTENT, and LEAKED_INSTRUCTIONS are errors of form, register, or generation control and are almost entirely detectable without source. In contrast, UNGROUNDED_INJECTION is source-required in 96.2% of cases because its label depends on whether added definitions, clinical details, or interpretive glosses are supported. GROUNDED_OVERGENERATION is similarly source-required.

This distinction motivates the reference-free benchmark setting without replacing source-aware evaluation: surfaced overgeneration decomposes into a large output-only artifact problem and a smaller but substantial grounding problem.

5.3 Different Systems Fail Differently

The taxonomy also reveals strong system signatures. Across released positive labels, T5-family cases split between GENERATION_FAILURE (61.0%) and prompt-template leakage (37.5%); LLaMA-family positives are dominated by LEAKED_INSTRUCTIONS (86.9%); and BART-family positives are mostly UNGROUNDED_INJEC-

Judge	Access	Prompt	F1	P	R
GPT-4o	RF	0-shot	.713	.762	.670
GPT-4o	RF	few	**.730**	.792	.678
GPT-4o	RD	0-shot	.661	.525	.892
GPT-4o	RD	few	.727	.670	.794
Flash	RF	0-shot	.680	.625	.746
Flash	RF	few	.706	.639	.789
Flash	RD	0-shot	.706	.572	.921
Flash	RD	few	.704	.615	.825
Llama3-8B	RF	few	.654	.726	.594

Table 5: Sentence-level LLM judge performance on the 2,688-sentence human set. Rows exclude rare failed calls (<0.5% per run). RF = REF-FREE; RD = REF-DEP. We report the strongest Llama3-8B setting; it uses the same prompts but weaker JSON-format enforcement.

TION (67.8%) and REPETITIVE_CONTENT (24.5%; Appendix Table 8).

These trends should not be read as architecture-only claims, since a family label also bundles prompting, fine-tuning, decoding, and post-processing. Rather, they show that the resource captures system-dependent failure profiles that would be invisible in a binary overgeneration benchmark.

6 Benchmark Utility I: Are Generic LLM Judges Enough?

The *detectable-without-source* annotation raises a benchmark question: if much surfaced overgeneration is output-intrinsic, are generic LLM judges already sufficient for reference-free detection? We evaluate this at the sentence level, matching the resource.

6.1 Setup and Main Results

We evaluate GPT-4o, Gemini 2.5 Flash, and a local Llama3-8B judge on 2,688 human-labeled sentences from 154 system outputs (646 positive, 2,042 negative). Confidence intervals resample outputs, not individual sentences. REF-FREE judges see the target prediction sentence and local prediction context; REF-DEP judges additionally see the source. Because Gemini supplies the large-scale taxonomy labels, we treat it as secondary; GPT-4o and Llama3-8B provide non-annotator comparisons.

Table 5 shows that generic judges are not blind, but they plateau. Frontier/API judges cluster between 0.661 and 0.730 F1, with overlapping output-cluster bootstrap intervals; the strongest local Llama3-8B setting reaches 0.654. Source access is not uniformly beneficial. For GPT-4o, zero-shot REF-DEP raises recall to 0.892 but inflates false positives from 135 to 521; few-shot prompting reduces these false positives to 253 and raises F1 from 0.661 to 0.727. Gemini shows the same recall-heavy pattern with less few-shot sensitivity.

6.2 What Source Access Fixes, and What It Does Not

The strongest signal is the human-validated DWS boundary. GPT-4o REF-FREE recalls 94.8% of output-intrinsic positives but only 14.7% of source-required positives; with source access, source-required recall rises to 77.2%. The same ref-free asymmetry appears for Llama3-8B few-shot, which recalls 85.3% of output-intrinsic positives but only 10.7% of source-required positives. For UNGROUNDED_INJECTION, recall rises from 17.2% without the source to 79.1% with it, while LEAKED_INSTRUCTIONS and REPETITIVE_CONTENT are near-ceiling in both settings.

Aggregate F1 therefore overstates hard-case performance. Removing GENERATION_FAILURE and REPETITIVE_CONTENT, GPT-4o REF-FREE few-shot drops from 0.730 to 0.630, and REF-DEP few-shot drops from 0.727 to 0.640. Generic LLM judges are useful screening baselines, but remain dominated by output-obvious categories and prompt-sensitive on source-required additions.

7 Benchmark Utility II: Specialized Output-Only Detection

We next test whether a small fine-tuned encoder, trained on the released taxonomy labels, can serve as a cheaper, reproducible, precision-first detector for the output-only setting.

7.1 Task and Model

The input is one simplified sentence with no source access. We fine-tune DeBERTa-v3-base (184M parameters) on the released training split, including the template-echo correction, using 1:1 negative subsampling. Training details are in Appendix D. The threshold is tuned on an unbalanced dev set to maximize sentence-level F1; $t = 0.900$ yields a high-precision operating point.

7.2 Main Results

Table 6 compares the detector against the strongest LLM judge settings. The detector does not beat frontier LLM judges on F1: it reaches 0.663 (95% CI [0.552, 0.755]), while GPT-4o REF-FREE few-shot reaches 0.730 (95% CI [0.636, 0.804]). The in-

System	Access	F1	P	R	FP
GPT-4o RF-FS	output	**.730**	.792	.678	115
GPT-4o RD-FS	source	.727	.670	.794	253
Flash RF-FS	output	.706	.639	.789	287
Flash RD-0	source	.706	.572	.921	442
OG Detector	output	.663	**.852**	.543	**61**
Llama3-8B RF-FS	output	.654	.726	.594	145

Table 6: Sentence-level overgeneration detection on the 2,688-sentence human set. The detector is the highest-precision system and has the fewest false positives, but lower recall than LLM judges.

tervals overlap, but operating profiles differ sharply: the detector has the highest precision (0.852) and fewest false positives (61), making it useful as a conservative local penalty rather than a replacement for source-aware evaluation. It also has similar F1 to the local Llama3-8B judge, but with substantially higher precision and fewer false positives.

Thresholding controls the precision–recall tradeoff. At $t = 0.50$, the detector reaches F1 = 0.680 (P = 0.795, R = 0.594); at $t = 0.900$, it is more precision-dominant (P = 0.852) with zero API cost.

7.3 Category-Level Analysis and DWS Utility

Although trained only on binary labels, the detector's recall hierarchy follows the human-validated DWS partition (Appendix Table 9). It is strong on LEAKED_INSTRUCTIONS (92.1%) but weaker on diverse GENERATION_FAILURE cases (58.5%), where GPT-4o REF-FREE few-shot recalls 92.6%. Both reference-free systems hit the same ceiling on source-required content: UNGROUNDED_INJECTION recall is 15.4% for the detector and 16.3% for GPT-4o REF-FREE few-shot, while GPT-4o REF-DEP few-shot reaches 49.8%.

This convergence strengthens the DWS claim: a supervised encoder and a prompted frontier LLM fail on the same source-required category when denied the source. Aggregate recall mirrors this pattern (0.756 for DWS=True vs. 0.143 for DWS=False, 5.3×). Excluding GENERATION_FAILURE removes most of the F1 gap between the detector and GPT-4o REF-FREE few-shot (0.637 vs. 0.639).

8 Discussion and Conclusions

Modern NLP models are based on generative LLMs, which produce fluent, mostly faithful outputs that can still contain extraneous material such as assistant chatter, repetitive content, or unsupported explanations. In this paper, we analyzed the presence and categories of overgeneration observed in submissions to a biomedical text simplification benchmark. By classifying overgeneration in the original submissions, we create a new labeled resource for overgeneration detection and classification that exploits the text-simplification setup, with sources, references, and predictions in the same language.

This paper created a large resource comprising almost 28K documents and 450K sentences from representative LLM Benchmark submissions. As this exceeds our capacity to manually label at this scale, we used a heuristic to realign source and prediction and flagged typical overgeneration cases: long inserts at the end of a prediction. This is a high-accuracy detector, which still underestimates the overgeneration present in the output of representative NLP models. In fact, the high prevalence is a sobering finding, as systematic analysis reveals it is far higher than a casual inspection of NLP output would suggest.

We further classified the observed overgeneration into a small number of categories, with particular interest in which types would be detectable without access to the source input. Whilst in controlled NLP experiments, sources are typically available, this is often not the case when analyzing generated predictions post-hoc, or in the massive amounts of generated content on the Web and in Social Media. It is a hopeful finding that the overwhelming majority of cases (over 80%) can be detected without access to, and alignment with, the source.

More generally, our new resource for overgeneration identification and classification aims to promote both the development of NLP technology to detect and classify non-faithful content in the output of modern NLP models and the development of novel NLP technology that is faithful by design. Our results suggest actionable ways to address one of the main limitations of current generative models.

9 Limitations

Our paper focuses on text simplification, which provides a unique setup where sources, predictions, and references are in the same language, enabling investigation of overgeneration and other extraneous content in LLMs' predictions submitted to NLP benchmarking tasks. However, the issues observed in this paper extend to all generative NLP tasks and all reference-based evaluation settings. These

issues address one of the main foundational evaluation problems in NLP today, and further research on other NLP benchmarks and evaluation settings would be valuable for fully understanding the remaining limitations of generative NLP models and current evaluation measures.

The main limitations follow from the construction pipeline. First, the resource is centered on Cochrane Biomedical Abstracts. Although the observed overgeneration examples are generic, the precise category frequencies may differ in news, legal, educational, or conversational simplification. We focused on biomedical abstracts in English, and it would be of great interest to extend this to other languages and mixed-language settings (e.g., English prompts and Chinese content). Cochrane systematic reviews are translated into many languages, and we are planning future research on multilingual biomedical abstracts (similar to Joseph et al., 2023).

Second, the taxonomy is conditioned on the trailing-content extractor plus a deterministic template-echo correction. The resource is therefore a high-coverage view of surfaced overgeneration, not an exhaustive inventory of all simplification errors. Worryingly, perhaps, the high prevalence of overgeneration observed in this paper still understates the remaining issues in representative submissions to an NLP Benchmark.

Third, the large-scale taxonomy is LLM-assisted. The DWS flag is directly validated on the human-positive set, and human audits support the main conclusions, but borderline category cases remain between acceptable simplification, grounded overflow, and unsupported glosses. We therefore treat the taxonomy as an audited resource rather than a perfect sentence-level gold annotation.

Fourth, the ability to detect and quantify overgeneration opens new handles for novel evaluation measures to quantify gratuitous content inserted by current models, as well as NLP pipelines that carefully detect and sanitize output post-hoc. Future work should build hybrid evaluators that use DWS as a routing signal: output-only detectors handle leakage, repetition, and generation collapse, while source-aware models handle unsupported additions. Calibration, selective prediction, and extension beyond biomedical abstracts are natural next steps.

Disclosure of Generative AI use

During the preparation of this work, the authors used *Grammarly* and *ChatGPT* in order to: **Grammar and spelling check** and **Paraphrase and reword**. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Anonymous. 2026. A benchmark for overgeneration detection in biomedical text simplification. In *Withhold to preserve anonymity*.

Jan Bakker and Jaap Kamps. 2024. Cochrane-auto: An aligned dataset for the simplification of biomedical abstracts. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 41–51, Miami, Florida, USA. Association for Computational Linguistics.

Dennis Davari, Liana Ermakova, and Ralf Krestel. 2024. Comparative analysis of evaluation measures for scientific text simplification. In *Linking Theory and Practice of Digital Libraries – 28th International Conference on Theory and Practice of Digital Libraries, TPDL 2024, Proceedings, Part I*, volume 15177 of *Lecture Notes in Computer Science*, pages 76–91. Springer.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Liana Ermakova, Hosein Azarbonyad, Jan Bakker, Benjamin Vendeville, and Jaap Kamps. 2025. Overview of the CLEF 2025 SimpleText track – simplify scientific text (and nothing more). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*

– *16th International Conference of the CLEF Association, CLEF 2025, Proceedings*, volume 16089 of *Lecture Notes in Computer Science*, pages 436–463. Springer.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2023. Multilingual simplification of medical texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16662–16692, Singapore. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Zihao Li, Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Matthew Shardlow, and Goran Nenadic. 2024. Large language models for biomedical text simplification: Promising but not there yet.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Benjamin Vendeville, Liana Ermakova, and Pierre De Loor. 2025. Resource for error analysis in text simplification: New taxonomy and test collection. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025*, pages 3723–3732. ACM.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations*. OpenReview.net.

A Quiz Question (Question)

Figure 2 shows a real CLEF 2026 SimpleText Track prediction. This prediction is from one of the better-scoring submissions to the track, obtaining a SARI score of 42.33 when evaluated against references based on the Cochrane plain language summaries.

System output (simplified biomedical abstract)

We studied 196 stroke patients with anxiety in three trials. One small study gave 21 stroke survivors a relaxation CD to use for four weeks. One study used a relaxation CD and found that anxiety decreased after three months. The second trial involved 81 participants who were given either paroxetine, paroxetine with therapy, or standard care. The anxiety scores were 5.4, 3.8, and 12.8, which is a significant difference. The third trial gave 94 stroke patients either buspirone or standard care. The anxiety levels were significantly lower in one group (6.5) than the other (12.6). Many people taking paroxetine had side effects like nausea or dizziness, but fewer people taking buspirone had these problems. We don't know how long side effects lasted. The relaxation therapy study had no bad side effects. I chose 'rephrase' as the internal simplification strategy to simplify the sentence while maintaining its original meaning. The evidence wasn't very reliable. The studies had few participants, especially the one on relaxation therapy. The drug studies didn't give enough information to know if they were reliable. The relaxation therapy study had some problems that might have affected the results. We don't have enough information to know how to treat anxiety after a stroke. More research is needed to find the best treatments for anxiety after a stroke.

Figure 2: A real system output from CLEF 2025 SimpleText. Quiz Question: Can you spot overgeneration when close-reading the output? And would you detect it when skim-reading the output? Or in the downstream processing of the output in your NLP pipeline?

Figure 3 in Appendix F shows the prediction from Figure 2 with the overgeneration highlighted.

B Category-Level Human Agreement Details

Category	Hum.	Tax.	P	R	F1
Ungrounded	260	191	0.99	0.73	0.84
CoT / Leaked	214	223	0.92	0.96	0.94
Repetitive	10	13	0.31	0.40	0.35

Table 7: Per-category agreement between human annotations and the LLM-assisted taxonomy on the 484 overlapping sentences with a mapped human category. Precision and recall treat the human label as reference.

UNGROUNDED_INJECTION has near-perfect precision but lower recall: the taxonomy reclassifies 71 human-labeled ungrounded sentences into other categories, predominantly GENERATION_FAILURE (22), NON_OVERGENERATION (22), and LEAKED_INSTRUCTIONS (17). The 85 mapped-category disagreements arise mainly from priority decisions: 32 structurally degenerate outputs are assigned to GENERATION_FAILURE; 23 sentences are assigned to NON_OVERGENERATION (22 human-labeled ungrounded and one human-labeled CoT leakage); and 17 mixed assistant-framing cases are assigned to LEAKED_INSTRUCTIONS over UNGROUNDED_INJECTION. The 186 human *Other* cases, excluded from the mapped comparison, are classified predominantly as GENERATION_FAILURE (140/186, 75.3%), suggesting that the earlier *Other* label largely captured structurally degenerate outputs.

C Model-Family Failure Profiles

Family	*n*	Dominant failure profile
T5	20,153	61.0% GEN._FAIL.; 37.5% LEAKED_INST.
LLaMA	9,895	86.9% LEAKED_INSTRUCTIONS
BART	5,626	67.8% UNGROUNDED_INJ.; 24.5% REP._CONT.
GPT	2,143	63.7% UNGROUNDED_INJ.; mixed residual
Gemini	1,127	62.3% UNGROUNDED_INJ.; 34.8% LEAKED_INST.

Table 8: Dominant category profiles by model family across released positive labels after deterministic template-echo correction.

Table 8 shows the dominant failure profiles for each model family across detected overgeneration categories.

D Detector Training Details

The DeBERTa-v3-base detector is trained with effective batch size 64, learning rate 2×10^{-5}, and 5 epochs; the best checkpoint is epoch 3 (dev F1 = 0.973). Training uses a single A100 and takes 56 minutes. The corrected training set has a 9.5% positive rate. Empty or whitespace-only sentences are deterministically classified as positive before the model runs.

E Output-Only Detector: Category-Level Results

Table 9 shows the sentence-level recall of overgeneration detection without access to the source sentences.

F Quiz Question (Answer)

Figure 3 shows the prediction from Figure 2 in Appendix A with the overgeneration highlighted.

Category	Sents	Recall	DWS
LEAKED_INST.	227	0.921	True
GENERATION_FAIL.	176	0.585	True
REPETITIVE_CONT.	14	0.429	True
UNGROUNDED_INJ.	215	0.154	False
GROUNDED_OVERGEN.	6	0.000	False
DWS=True overall	422	0.756	—
DWS=False overall	224	0.143	—

Table 9: Sentence-level recall of the output-only detector (§7) by taxonomy category and majority DWS status. The detection hierarchy mirrors the DWS boundary from §3.3: output-visible categories are reliably detected; source-required categories are not.

System output (simplified biomedical abstract)

We studied 196 stroke patients with anxiety in three trials. One small study gave 21 stroke survivors a relaxation CD to use for four weeks. One study used a relaxation CD and found that anxiety decreased after three months. The second trial involved 81 participants who were given either paroxetine, paroxetine with therapy, or standard care. The anxiety scores were 5.4, 3.8, and 12.8, which is a significant difference. The third trial gave 94 stroke patients either buspirone or standard care. The anxiety levels were significantly lower in one group (6.5) than the other (12.6). Many people taking paroxetine had side effects like nausea or dizziness, but fewer people taking buspirone had these problems. We don't know how long side effects lasted. The relaxation therapy study had no bad side effects. I chose 'rephrase' as the internal simplification strategy to simplify the sentence while maintaining its original meaning. The evidence wasn't very reliable. The studies had few participants, especially the one on relaxation therapy. The drug studies didn't give enough information to know if they were reliable. The relaxation therapy study had some problems that might have affected the results. We don't have enough information to know how to treat anxiety after a stroke. More research is needed to find the best treatments for anxiety after a stroke.

Figure 3: A real system output from CLEF 2025 SimpleText (Quiz Answer). Have you spotted overgeneration when close-reading the output? And would you detect it when skim-reading the output? Or in the downstream processing of the output in your NLP pipeline?

Source (Abstract of CD008860)

We included three trials (four interventions) involving 196 participants with stroke and co-morbid anxiety. One trial (described as a 'pilot study') randomised 21 community-dwelling stroke survivors to four-week use of a relaxation CD or to wait list control. This trial assessed anxiety using the Hospital Anxiety and Depression Scale and reported a reduction in anxiety at three months among participants who had used the relaxation CD (mean (standard deviation (SD) 6.9 (± 4.9) and 11.0 (± 3.9)), Cohen's d = 0.926, P value = 0.001; 19 participants analysed). The second trial randomised 81 participants with co-morbid anxiety and depression to paroxetine, paroxetine plus psychotherapy, or standard care. Mean levels of anxiety severity scores based on the Hamilton Anxiety Scale (HAM-A) at follow-up were 5.4 (SD ± 1.7), 3.8 (SD ± 1.8), and 12.8 (SD ± 1.9), respectively (P value < 0.01). The third trial randomised 94 stroke patients, also with co-morbid anxiety and depression, to receive buspirone hydrochloride or standard care. At follow-up, the mean levels of anxiety based on the HAM-A were 6.5 (SD ± 3.1) and 12.6 (SD ± 3.4) in the two groups, respectively, which represents a significant difference (P value < 0.01). Half of the participants receiving paroxetine experienced adverse events that included nausea, vomiting, or dizziness; however, only 14% of those receiving buspirone experienced nausea or palpitations. Trial authors provided no information about the duration of symptoms associated with adverse events. The trial of relaxation therapy reported no adverse events. The quality of the evidence was very low. Each study included a small number of participants, particularly the study of relaxation therapy. Studies of pharmacological agents presented details too limited to allow judgement of selection, performance, and detection bias and lack of placebo treatment in control groups. Although the study of relaxation therapy had allocated participants to treatment using an adequate method of randomisation, study recruitment methods might have introduced bias, and drop-outs in the intervention group may have influenced results. Evidence is insufficient to guide the treatment of anxiety after stroke. Further well-conducted randomised controlled trials (using placebo or attention controls) are required to assess pharmacological agents and psychological therapies.

Figure 4: Source input from CLEF 2025 SimpleText (Quiz Source). Could you spot overgeneration when comparing the source and the prediction side by side?

Figure 4 shows the source abstract for the Quiz Example in Figure 2 and Figure 3. The biomedical text simplification task is quite challenging, and spotting errors by comparing the source and prediction side by side is highly non-trivial.

Figure 5 shows the reference plain language summary for the Quiz Example in Figure 2 and Figure 3. This human reference simplification follows specific guidance from Cochrane. Standard evaluation measures compare the references with the prediction and focus on the intersection, making them highly insensitive to additional content in the prediction caused by overgeneration. To be precise, SARI looks at the words that the reference simplification adds, keeps, or deletes separately, and averages these scores. Additional content that is inserted in the prediction, but does not appear in the reference, is not discounted directly, other than a minor effect due to increasing prediction length.

Reference (Plain Language Summary of CD008860)

We found three studies with 196 stroke survivors who had received a diagnosis of anxiety. One study assessed the effect of a relaxation CD used five times a week for one month for participants with a diagnosis of anxiety. One study reported that participants were less anxious when treated with an antidepressant medicine (paroxetine), or with paroxetine and psychotherapy, than with standard care. This study reported that half of the participants receiving paroxetine experienced side effects that included nausea, vomiting, or dizziness. The third study also reported that participants were less anxious when treated with an antidepressant (buspirone hydrochloride) than with standard care, and only 14% of those receiving buspirone hydrocholoride reported nausea or palpitations. We judged that the quality of this evidence was very low. Studies were few and each included a small number of participants. The study of relaxation therapy was very small, with loss of two participants who used the CD, and the study recruitment process may have attracted participants who had a positive bias towards psychological therapies. Current evidence is insufficient to guide the treatment of anxiety after stroke. Additional well-conducted randomised trials are needed.

Figure 5: Reference from CLEF 2025 SimpleText (Quiz Reference). Could you spot overgeneration when comparing the reference and the prediction side by side?