CYUT at CLEF 2026 SimpleText Track: Hallucination Detection based on Retrieval-Aligned Contextual Fusion and Multi-Tier Cascaded Discernment

Notebook for the SimpleText Lab at CLEF 2026

Shih-Hung Wu[1,*], YAN CHONG-ZHE[1]

[1]Department of Computer Science and Information Engineering,Chaoyang University of Technology, Taichung, Taiwan

Abstract

Medical text simplification frequently suffers from factual overgeneration and ungrounded semantic hallucinations. To address this challenge, Team CYUT developed a Retrieval-Aligned Contextual Fusion and Multi-Tier Cascaded Discernment Framework for the CLEF 2026 SimpleText Task 2 benchmark. Our architecture couples: (1) a retrieval-aligned semantic matching layer leveraging fine-tuned `DeBERTa-v3` encoders, (2) an active generative auditing layer powered by unquantized `Qwen3-8B` causal models, and (3) a regularized tree-based clinical knowledge calibration layer driven by `XGBoost` ensembles. Tested under document-level benchmarks, our unified configuration achieves a highly competitive Document F_1-score of 0.8071 (Precision: 0.7980, Recall: 0.8163) on the official test set. Our ablation analysis further shows that retrieval-aligned contextual matching functions as the dominant contributor to system performance; omitting this context-matching mechanism catastrophically reduces the macro F_1-score to a dismal 0.4602, demonstrating its critical role in maintaining factual boundaries.

Keywords

Medical Text Simplification, Hallucination Detection, Length-Normalized Contrastive Perplexity, Heterogeneous Cascade Pipeline, XGBoost Classifier

1. Introduction

Medical text simplification aims to transform jargon-heavy clinical documents into patient-friendly narratives without sacrificing core factual accuracy. Even minor semantic distortions, such as misstating drug dosages or reversing negative clinical findings, can alter clinical interpretations entirely, leading to potentially misleading medical interpretations in real-world scenarios.

Detecting factual overgeneration and semantic drift in simplified medical narratives requires both rigid semantic grounding and calibrated validation mechanisms. The core challenge in operationalizing automated factuality auditors for the CLEF 2026 SimpleText Task 2 benchmark lies in isolating ungrounded textual injections without suffering from contextual information loss over highly fluent generative texts. This constraint motivates our framework design to dynamically incorporate localized factual context windows, anchoring continuous transformer representations within a regularized and computationally structured cascaded system.

To address these limitations, we participate in the CLEF 2026 SimpleText Track [1], an international evaluation forum dedicated to the automated simplification and validation of scientific texts. The broader scope of the SimpleText track comprises three foundational challenges: Task 1 targetting the core text simplification pipeline [2], Task 3 focusing on scholarly research area classification [3], and Task 2 addressing automated hallucination management [4].

Specifically, our team targets **Task 2.1: Identify and Avoid Hallucination (Binary Detection Mode)** [4]. The primary scientific objective of this task is to process paired instances and detect overgeneration–extraneous or ungrounded factual content introduced during text simplification that

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

shwu@cyut.edu.tw (S. Wu); s11427604@gm.cyut.edu.tw (Y. CHONG-ZHE)

is unsupported by the original biomedical abstracts sourced from the Cochrane Library systematic reviews [4].

Our empirical tracking indicates that cross-sentence text pairs benefit from localized contextualization prior to dense classification to preserve stable semantic alignment boundaries. Within our evaluated framework architecture, active generative post-audits offer marginal boundary calibration over ambiguous text structures, whereas omitting the context-matching mechanism substantially minimizes the overall verification accuracy. Consequently, we design and construct a computationally streamlined, Retrieval-Aligned Contextual Fusion and Multi-Tier Cascaded Discernment Framework, and subsequently evaluate its performance through a three-tier automated verification pipeline:

- **Layer 1 (Retrieval-Aligned Semantic Discernment Layer)**: Dynamically maps each simplified candidate sentence against the source Cochrane abstract, extracting the **Top-3 most semantically coupled biomedical facts** to construct context-aligned text pairs. A continued fine-tuned transformer backbone (`DeBERTa-v3`) assesses these balanced pairs, performing adaptive early exits for high-confidence safe segments to optimize runtime computational load.
- **Layer 2 (Active Auditing Layer)**: Engages an unquantized `Qwen3-8B` causal model for ambiguous grey-area text structures, computing a length-normalized contrastive perplexity (LN-CP) metric over the retrieved facts to analyze ungrounded semantic drift under length-invariant constraints.
- **Layer 3 (Clinical Knowledge Calibration Layer)**: Integrates a dense `XGBoost` tree classifier tracking a handcrafted 5-Dimensional (5D) rule matrix over the aligned context to act as an auxiliary calibration mechanism for critical numerical and negation polarities mismatch. Crucially, under downstream multi-class taxonomy routing (Task 2.2), this auxiliary calibration layer can be expanded into a 10-Dimensional (10D) structural pathology matrix via Gradient Boosting ensembles to isolate localized autoregressive generation failures.

2. Related Work

The challenge of tracking factuality and mitigating hallucinations in Large Language Models (LLMs) has sparked dual-paradigm research divided into sub-symbolic internal probings and symbolic external verifications [5]. To contextualize the contributions of Team CYUT, we review contemporary literature across four evolving axes: foundational representation encoders, post-processing text validation, retrieval-augmented mechanisms, and tree-based ensemble estimators.

2.1. Foundational Language Encoders

The mathematical extraction of continuous semantic constraints relies heavily on bidirectional deep Transformers, initiated by BERT [6] and robustly optimized by RoBERTa [7]. Subsequently, DeBERTa [8] advanced text alignment resolution through a disentangled attention mechanism and an Enhanced Mask Decoder (EMD) that factors absolute positions before the final decoding layer. Rather than deploying these encoders out-of-the-box, we perform targeted continued fine-tuning on a balanced text matcher inspired by these mechanics to secure an optimal semantic hyper-plane.

2.2. Post-Processing Factuality Validation

Post-processing techniques are highly valuable in high-stakes clinical scenarios because they do not require updating model weights [9]. Standard approaches fall into reference-free self-reflection streams or reference-based comparative verification networks [9], such as FactScore [10] and CoNLI [11]. Crucially, the Hierarchical Semantic Piece (HSP) method integrates multi-granularity semantic piece extraction to track local detail variations [9]. This hierarchical alignment inspires our framework, where we decouple neural text semantics from symbolic knowledge constraints.

2.3. Retrieval-Augmented Hallucination Probing

Retrieval-Augmented Generation (RAG) reduces knowledge-cutoff limits but remains susceptible to internal hallucinations that directly conflict with the retrieved context [12]. To diagnose this failure mode, ReDeEP decoupled the model's utilization of external context from its parametric boundaries, demonstrating that hallucinations occur when parametric knowledge over-injects internal information while attention blocks fail to retain context [13]. This mechanistic separation motivates our synchronized Top-3 RAG factual alignment, anchoring cross-attention weights before engaging heavy causal networks.

2.4. Logits-Induced Uncertainty Estimation

Measuring token-level statistical confidence determines when a cascaded pipeline should trigger external auditing layers. While sampling-based approaches evaluate semantic consistency across multiple outputs under prohibitive computational costs, traditional probability-based indicators frequently fail due to softmax normalization discarding absolute evidence strength [14]. To resolve this, LogTokU treats unnormalized logits as parametric parameters of a Dirichlet distribution [14]. Our framework expands this grey-box concept within Layer 2 via a Length-Normalized Contrastive Perplexity (LN-CP) metric to calibrate raw entropy against text length drift.

2.5. Gradient Boosted Decision Systems for Symbolic Calibration

While neural representation spaces excel at macro contextual understanding, wrapping them under robust non-parametric decision forests functions as a viable standard to complement potential weaknesses of neural decision boundaries. This calibration paradigm was crystallized by XGBoost [15] via sparsity-aware split finding and instance-weighted quantile sketch procedures. Concurrently, LightGBM [16] advanced scaling efficiency through Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

In our architecture, rather than processing continuous vector boundaries in isolation, we encase the neural backbones under a structured tree ensemble framework to parameterize an auxiliary calibration layer. This regularized clinical checkpoint tracks discrete attribute variations to adjust continuous posterior probabilities, smoothly regularizing ambiguous text structures back toward factual bounds without suffering from contextual information loss.

3. Approach

3.1. Data Description and Task Formulation

Our framework is trained and evaluated strictly on the official guidelines of the CLEF 2026 SimpleText Task 2.1b benchmark. The corpus is derived from 408 unique biomedical source abstracts from Cochrane Library systematic reviews, processed through 18 distinct language models across 55 empirical evaluation runs. The benchmarking corpus is partitioned into three discrete subsets as detailed in Table 1.

Table 1
Official CLEF 2026 SimpleText Task 2.1b Dataset Partitions.

Data Split	Document Examples	Total Sentences	Factual Positive Docs (Ratio)
Train	21,057	350,583	4,427 (21.0%)
Dev	1,120	18,638	234 (20.9%)
Test	6,728	105,001	*Competition Blind*

The sentence-level target space is dictated by a fine-grained **Label Taxonomy**, consisting of six distinct behavioral categories mapping factual safety and types of overgeneration. Table 2 details the empirical label distribution observed across the training partition.

To further fortify our classifiers against highly fluent yet ungrounded hallucinations, we implement a targeted **Adversarial Contextual Data Augmentation** pipeline. Specifically, we execute a localized Retrieval-Augmented Generation (RAG) cross-pollination script over the primary corpus. Under this setup, our nested pipeline processes a total of **9,467 unique document instances** prefixed under the Qwen3 tracking domain. Driven by our adversarial fishing prompts, the generative backbone successfully harvests an aggregate of **15,086 fine-grained synthetic sentences** explicitly labeled as positive OVERGENERATION anomalies. These newly engineered hard negative instances are appended natively into the primary dataset structure under a nested topology, scaling up the diversity of positive training boundaries and ensuring robust decision margin convergence during the subsequent continued fine-tuning phase.

Table 2
Factual Overgeneration Label Taxonomy and Distribution.

Category Target	Count (Train)	Structural and Clinical Semantic Description
NONE	323,702	Safe sentence; no overgeneration detected.
GENERATION_FAILURE	11,331	Catastrophic model output (e.g., degenerate loops).
LEAKED_INSTRUCTIONS	8,238	Inclusion of system framing or meta-rationales.
UNGROUNDED_INJECTION	5,735	Unsupported factual additions or interpretations.
REPETITIVE_CONTENT	1,477	Pathological repetition loops from decoding.
GROUNDED_OVERGEN	100	Source-supported content beyond scope.

Under the official evaluation setup, sentence-level predictions are aggregated to the document level via a logical OR operator. If any single sentence within a document is flagged as overgeneration, the entire parent document is classified as positive.

To prevent pathological label leakage and address potential distribution shift vulnerabilities inherent in synthetic data pipelines, the 15,086 harvested sentences were subjected to a rigorous filtering protocol. First, all synthetic generations were verified through exact-string matching to ensure zero token overlap with the official development and competition testing splits, guaranteeing strict data isolation. Second, we conducted a localized Jaccard-distance divergence check against the base training set to ensure the synthetic samples mirrored the vocabulary distribution of the target taxonomy without introducing unregularized vector outsets. Crucially, given that the synthetic instances aggregate to approximately 3.4 times the baseline positive counts, class balancing and stochastic sampling filter operations were dynamically applied during the subsequent continued fine-tuning phase to prevent these synthetic instances from dominating the neural optimization gradient streams.

3.2. Retrieval-Aligned Contextual Fusion and Cascaded Architecture

The operational nexus of our framework relies on eliminating the semantic blind spots of sentence-isolated verification by enforcing a strict, synchronized **Retrieval-Aligned Contextual Fusion** workflow across both our training and inference phases. Because medical abstracts present dense, multi-sentence argument structures, evaluating a simplified candidate clause out-of-context introduces extreme semantic passivation.

To bridge this representation gap, our pipeline natively models the verification space as a localized retrieval task. Let T_{src} denote the raw source biomedical abstract, which is split into a set of discrete sentence sequences:

$$\mathcal{S}_{src} = \{s \in \text{split}(T_{src}) \mid \text{len}(s) > 5\} \tag{1}$$

For any candidate simplified sentence t_{sim}, a dense retriever (`all-MiniLM-L6-v2`) parameterizes the text structures into continuous embedding space. The system executes a dense semantic retrieval operator to isolate the **Top-3 most textually coupled biomedical facts**:

$$\mathcal{B}_{top3} = \text{Top-3}_{s \in \mathcal{S}_{src}} \left(\frac{\mathbf{E}(t_{sim}) \cdot \mathbf{E}(s)}{\|\mathbf{E}(t_{sim})\| \|\mathbf{E}(s)\|} \right) \tag{2}$$

To preserve absolute global baseline contexts, the final synchronized evaluation block $\mathcal{C}_{fuse}$ is assembled via an ordered spatial concatenation that always retains the primary baseline text structure (s_0):

$$\mathcal{C}_{fuse} = \mathcal{B}_{top3} \cup \{s_0\} \tag{3}$$

The cross-encoder and tree classifiers natively consume the fused pair $[\mathcal{C}_{fuse}, t_{sim}]$ across both full-scale continuous offline continued fine-tuning and active online real-time inference routing. This dual-stream alignment guarantees that the neural networks are explicitly evaluated against the localized empirical ground-truth facts, anchoring the decision hyper-plane.

3.2.1. Layer 1: Retrieval-Aligned Semantic Discernment Layer

Instead of relying on an off-the-shelf vanilla model, the foundational dual-stream matcher is initialized via a `microsoft/deberta-v3-large` backbone and explicitly undergoes rigorous **Continued Fine-tuning (CFT)** over the synchronized context pairs $[\mathcal{C}_{fuse}, t_{sim}]$ to capture domain-specific biomedical alignment profiles. To eliminate semantic boundary noises propagated by skewed class distributions within clinical text simplification corpora, we implement an active **Class Balancing Down-sampling Filter**. This filter dynamically enforces a strict $3:1$ majority-to-minority instance ratio during batch generation, preventing the model from expanding epistemic bias toward the over-represented safe targets (NONE).

The cross-encoder semantic classifier is optimized over the balanced context pairs via the following structural parameters:

1. **Learning Rate Schedule**: A low baseline learning rate of $\eta = 5 \times 10^{-6}$ is deployed via a **Cosine Learning Rate Decay Scheduler** integrated with a 10% linear warmup ratio, stabilizing initial gradient distributions across dense medical phrases.
2. **Gradient Regularization**: To insulate the deep transformer architecture from potential gradient instability common during cross-attention alignment, a strict gradient clipping optimization barrier is enforced via a maximum norm constraint ($\max \|\nabla_\theta \mathcal{L}\| = 1.0$).
3. **Loss Function Calibration**: The neural backbone minimizes standard cross-entropy loss over the fused text pairs, executed natively in full `Float16` precision to preserve subtle representation variations while accelerating cross-attention throughput.

During active runtime inference routing, the fine-tuned backbone outputs a continuous posterior probability P_{deb}. To optimize macroscopic computational latency, an Adaptive Early Exit mechanism controlled by two conservative validation thresholds ($A_{\text{low}} = 0.35$ and $A_{\text{high}} = 0.75$) is deployed. If $P_{\text{deb}} < A_{\text{low}}$, the instance is instantly cleared as NONE, whereas if $P_{\text{deb}} > A_{\text{high}}$, the instance is classified as OVERGENERATION. This strategy screens out the majority of safe records parallelly, preserving intensive generative compute exclusively for ambiguous grey-area text structures.

3.2.2. Layer 2: Active Auditing Layer

For the remaining ambiguous sentences falling within the routing boundaries, we track the contextual information gain using an unquantized `Qwen3-8B` model running in Float16 precision. To contextualize cross-entropy variations under a length-invariant constraint, we formalize the **Length-Normalized Contrastive Perplexity (LN-CP)** metric as a bounded ratio:

$$LN\text{-}CP = \frac{\mathcal{L}_{no\_ctx} - \mathcal{L}_{ctx}}{\mathcal{L}_{no\_ctx} + \epsilon} \tag{4}$$

where $\mathcal{L}_{ctx}$ represents the token-level cross-entropy loss conditioned strictly on the fused retrieval context pairs, and $\mathcal{L}_{no\_ctx}$ denotes the baseline conditional loss over the simplified clause alone.

Crucially, we employ $\mathcal{L}_{no\_ctx}$ as the normalization denominator rather than raw token length or $\mathcal{L}_{ctx}$. This mathematical formulation conceptualizes the metric as a bounded *Relative Information Gain Ratio*, measuring the percentage of uncertainty reduction directly attributable to the external

Cochrane ground-truth facts. By scaling the raw difference against the model's intrinsic generation perplexity ($\mathcal{L}_{no\_ctx}$), the metric dynamically suppresses the systematic log-likelihood inflation common in inherently long or syntactically complex medical statements, transforming raw cross-entropy shifts into a normalized, domain-calibrated tracking hyper-plane.

3.2.3. Layer 3: Clinical Knowledge Calibration Layer

To shield the framework against the continuous blind spots and boundary noises of neural models, we integrate a regularized `XGBoost` tree classifier running as an auxiliary calibration mechanism behind the perception layers. During primary binary evaluation trials, this symbolic safeguard network tracks a handcrafted 5-Dimensional (5D) rigid factual matrix extracted directly from the aligned context pair to target localized clinical logic anomalies, unique word generation counts, and numerical deviations.

The layer contributes a calibrated adjustment score to the final decision boundary: if a simplified sentence introduces ungrounded numbers or medical dosages absent in the matched context, or if it alters the negation polarity, the tree network acts as a calibrated safety barrier to smooth the continuous neural outputs and regularize the final probability distribution toward a classification of OVERGENERATION.

3.2.4. Downstream Multi-Class Taxonomy Optimization (Task 2.2 Training Strategies)

To systematically route ambiguous sentences into fine-grained behavioral anomalies for Task 2.2, we formalize two distinct secondary taxonomy tracking strategies built directly behind the upstream binary networks:

1. Joint Neural Cross-Encoder Tuning:
We fine-tune an advanced cross-encoder backbone (`ms-marco-MiniLM-L-6-v2`)[17] configured with a 6-class output head mapped to our explicit label taxonomy. To resolve severe data passivations caused by heavily skewed class densities where the safe category (`None`) dominates the optimization space, we enforce a two-stage stabilization pipeline:

- **Prior Static Under-sampling Filter**: During dataset tokenization, we execute a static stochastic down-sampling filter over the `None` examples with a retention constraint of 25% ($P_{\text{retain}} \leq 0.25$), preventing the network from expanding substantial epistemic bias toward negative frames.
- **Bayesian Prior Weighting and Loss Calibration**: We implement a custom **Weighted-LossTrainer** infrastructure that overrides the default optimization criteria with a class-weighted cross-entropy loss ($\mathcal{L}_{\text{weighted}}$):

$$\mathcal{L}_{\text{weighted}} = -\sum_{c=1}^{C} w_c \cdot y_c \log\left(\sigma(z_c)\right) \tag{5}$$

  where the empirical class weight vectors w_c are derived dynamically via the inverse density matrix ($\frac{N}{C \cdot N_c}$). Crucially, to isolate numerical falsifications in medical statements, the prior weight for UNGROUNDED_INJECTION is scaled by an additional factor of 1.2. The model converges over 5 epochs in full `Float16` precision with a learning rate of $\eta = 4 \times 10^{-5}$ and a batch size of 32.

2. Feature-Engineered LightGBM Ensemble:
As a highly efficient alternative non-cascaded pipeline tailored for fine-grained morphological sorting, we train a Gradient Boosting Decision Tree (GBDT) model via the `LightGBM` library. Rather than reusing the binary factual armor, this tree ensemble dynamically acts upon the expanded 10-Dimensional (10D) complete matrix detailed in Table 3 to track explicit prompt controls leakage ($\mathcal{K}_{ctrl}$) and terminal structural pathologies. The model is configured with a softmax multi-class objective over a learning rate of 0.05, a leaf limit of 31, and an adaptive balanced class weight mechanism optimized over $1,000$ boosting rounds with a 50-round early stopping callback barrier.

3. Zero-Shot Generative Baseline Control:
As a non-tuned comparative anchor, the raw `Qwen3-8B` base model is evaluated under zero-shot prompting conditions without any weight modification or classification head alignment, highlighting the catastrophic failure modes of generative structures under strict multi-class formatting rules.

Table 3
Systematic Feature Compendium and Array Index Serialization Across Alternative Tracks.

Global Gen	Feature Designation	Algorithmic Target Class	Mathematical Formulation / Exact Criteria	5D Track	10D Track	13D Track
Neural Base	DeBERTa Target Probability	Probability Base Stream	$P(c_0\|\mathbf{x})$ and $P(c_1\|\mathbf{x})$	Concatenated	Concatenated	$f_{13D,0} \sim f_{13D,1}$
Neural Matrix $\mathbf{U}_{nli}$ (4 Dimensions)	Shannon Entropy Metric	Prediction Instability Probing	$-\sum p \log_2 p \quad (\text{clipped } p \in [10^{-12}, 1.0])$			$f_{13D,2}$
	Decision Margin Gap	Grey-Area Boundaries Separation	$\lvert P_{class\_0} - P_{class\_1}\rvert$			$f_{13D,3}$
	Logit Space Variance	Evidence Dispersion Tracking	$\text{Var}(\mathbf{z}_{single})$			$f_{13D,4}$
	Logit Peak Value	Absolute Likelihood Peak	$\max(\mathbf{z}_{single})$			$f_{13D,5}$
Symbolic Matrix	Length Ratio	General Text Compression	$\lvert W_{sim}\rvert / (\lvert W_{src}\rvert + \epsilon)$	$f_{5D,0}$	$f_{10D,0}$	$f_{13D,6}$
	Lexical Overlap Ratio	Source Grounding Check	$\lvert S_{src} \cap S_{sim}\rvert / (\lvert S_{sim}\rvert + \epsilon)$	$f_{5D,1}$	$f_{10D,1}$	$f_{13D,7}$
	New Word Count	Ungrounded Vocabulary Expansion	$\lvert W_{sim} \setminus W_{src}\rvert$ or $\lvert S_{sim} \setminus S_{src}\rvert$	$f_{5D,2}$		$f_{13D,8}$
	Numerical Mismatch Barrier	UNGROUNDED_INJECTION	1.0 if $\mathcal{N}(T_{sim}) \setminus \mathcal{N}(T_{src}) \neq \emptyset$ else 0.0	$f_{5D,3}$	$f_{10D,2}$	
	Negation Polarity Variance	Clinical Logic Reversal	1.0 if $\mathcal{C}_{neg}(T_{src}) \neq \mathcal{C}_{neg}(T_{sim})$ else 0.0	$f_{5D,4}$	$f_{10D,3}$	
Pathology $\mathbf{p}_{path}$ (5 Dimensions)	Monotonous Repetition Ratio	Pathological Decoding Loop	$1.0 - \lvert S_{sim}\rvert / (\lvert W_{sim}\rvert + \epsilon)$		$f_{10D,5}$	$f_{13D,9}$
	Instruction Leakage Density	LEAKED_INSTRUCTIONS	$(\sum \mathbb{I}(w \in \mathcal{K}_{ctrl}) + \mathcal{C}_{colon}(T_{sim})) / (\lvert W_{sim}\rvert + \epsilon)$		$f_{10D,6}$	
	Malformed Geometry Ratio	GENERATION_FAILURE	$\mathcal{C}_{regex}(\text{"(::\|:\s+\|:\s+:\s+:)"}) / (\lvert T_{sim}\rvert + \epsilon)$		$f_{10D,7}$	
	Punctuation Edge Density	Character Chaos Tracking	$\mathcal{C}_{punc}(T_{sim}) / (\lvert T_{sim}\rvert + \epsilon)$		$f_{10D,8}$	
	Unique Token Ratio	Vocabulary Richness Index	$\lvert S_{sim}\rvert / (\lvert W_{sim}\rvert + \epsilon)$		$f_{10D,9}$	
Advanced Statistical Appendices	Character Compression Ratio	Macro Structural Boundary Check	$\lvert T_{sim}\rvert / \lvert T_{src}\rvert$			$f_{13D,10}$
	Average Character Length	Token Fragmentation Monitor	$\text{Mean}(\sum_{w \in W_{sim}} \lvert T_w\rvert)$			$f_{13D,11}$
	Absolute Word Count	Clause Scale Context Mapping	$\lvert W_{sim}\rvert$		$f_{10D,4}$	$f_{13D,12}$

3.3. Handcrafted Feature Engineering and Search Profiles

To provide a transparent empirical analysis and document the comprehensive feature spaces explored during our framework's optimization, we consolidate our complete feature patterns into a structured taxonomy. Under our cascaded architecture, these symbolic feature streams serve as primary auxiliary operations to protect the pipeline against the vector-space blind spots of dense transformers.

Crucially, as established by our ablation studies, these handcrafted attributes do not dictate primary system viability; rather, they represent exploratory trajectory variations evaluated during our baseline tuning. The structural generalizability of these profiles is systematically organized into three progressive evaluations: the 5-Dimensional (5D) Deployed Track acts as our localized factual armor matching our active evaluation function `extract_features`, the 10-Dimensional (10D) Extension Track captures systemic autoregressive decoding pathologies optimized for downstream multi-class sorting via `extract_ultimate_10d_features`, and the 13-Dimensional (13D) Independent Track functions as a reference standalone comparative baseline mapping continuous neural uncertainty alongside raw statistical literals via `compute_advanced_uncertainty_batch`. The exact array serialization positions and localized feature indices developed across our alternative evaluation tracks are consolidated in Table 3.

All handcrafted metrics listed above are strictly evaluated directly over the dynamically fused **Top-3 RAG context abstract segments** ($\mathcal{C}_{fuse}$) rather than sentence-isolated pairs. This alignment guarantees that surface-level literal attributes are consistently anchored against localized empirical clinical facts, transforming high-dimensional noise into regularized boosting hyper-planes.

4. Results

4.1. Quantitative Results and Experimental Setup

To systematically evaluate the empirical consistency and diagnostical utility of our proposed heterogeneous framework, we conduct an extensive multi-task benchmarking comparison. The experimental validation is structured into two distinct operational layers: macro-level binary hallucination containment and micro-level fine-grained category mapping.

In the following sections, Table 4 outlines the comprehensive system configurations alongside competitive foundational language model baselines evaluated under the strict document-level error-veto mechanism for Task 2.1. Furthermore, to evaluate our cascade's capability in identifying the exact morphological boundaries of overgeneration types, Table 5 reports the fine-grained multi-class taxonomy classification metrics for Task 2.2 under varying secondary tree-based and cross-encoder sorting architectures.

Table 4
Task 2.1 Binary Hallucination Detection Performance Matrix on Test Benchmark.

Framework / Model Architecture	Doc Prec.	Doc Rec.	Doc F1	Sent Prec.	Sent Rec.	Sent F1
Proposed Framework (DeBERTa_XGB_QwenAudit)	**0.7980**	**0.8163**	**0.8071**	**0.8367**	**0.8855**	**0.8604**
Ablation Tracks of Complete System:						
Retrieval+DeBERTa+Xgb+Calibrated	0.7820	0.8200	0.8005	0.8274	0.8855	0.8555
Ablation: Without Retrieval Component	0.3045	0.9421	0.4602	0.1977	0.9072	0.3246
Ablation: Without XGBoost Classifier	0.6934	0.8889	0.7791	0.7567	0.9279	0.8336
Ablation: Without Knowledge Calibration Layer	0.7938	0.8200	0.8067	0.8330	0.8841	0.8578
Alternative Backbone and Tabular Baselines:						
Baseline: Retrieval + BERT Base	0.6176	0.9017	0.7331	0.7035	0.9353	0.8030
Baseline: Retrieval + RoBERTa Base	0.7135	0.8714	0.7846	0.8397	0.7713	0.9216
Baseline: Retrieval + DeBERTa Base	0.6380	0.8044	0.7116	0.7022	0.8340	0.7116
Baseline: Retrieval + DeBERTa + LightGBM	0.7966	0.6400	0.7098	0.8189	0.7468	0.7812
Baseline: Retrieval + DeBERTa + LLM_INTERNAL	0.6940	0.8871	0.7787	0.7574	0.9279	0.8340
Data Augmentation and Feature Scaling:						
Baseline: Retrieval + Roberta_augmented	0.8270	0.7420	0.7822	0.8541	0.8677	0.8409
Baseline: Retrieval + Deberta_augmented	0.8908	0.6520	0.7529	0.9026	0.8090	0.8532
Baseline: Retrieval + Roberta + DeBERTa_Aug	0.7667	0.8329	0.7984	0.8062	0.9034	0.8520
Baseline: Retrieval + Roberta_Aug + Deberta_Aug	0.8175	0.7649	0.7903	0.8549	0.8582	0.8566
Feature Variant: BGE_Rerank_XGB13D	0.7764	0.6694	0.7189	0.8407	0.7944	0.8169
Feature Variant: BGE_Rerank_XGB10D	0.6584	0.8531	0.7432	0.7390	0.8915	0.8081
Feature Variant: BGE_Rerank_XGB5D	0.7690	0.7062	0.7362	0.8339	0.8070	0.8203
Ensemble Meta Variant:						
Advanced Variant: Triple Ensemble	0.8072	0.8035	0.8053	0.8379	0.8838	0.8603

4.2. Empirical Analysis on Task 2.1 Binary Detection

The comparative matrix presented in Table 4 outlines the performance trajectories across alternative pipeline variations. Our complete configuration, DeBERTa_XGB_QwenAudit, achieves a Document F_1-score of **0.8071**. When isolating components over the system baseline, removing the retrieval mechanism (*Ablation:* Without Retrieval Component) bounds the Document F_1-score to a baseline of **0.4602**, indicating that the fine-tuned discriminative encoder relies heavily on the Top-3 localized context fusion to establish stable cross-sentence alignment boundaries under dense biomedical lexicons.

Conversely, our ablation analysis reveals that the continuous generative auditing block and the tree-based calibration layer function primarily as marginal boundary smoothing operations within this specific dataset distribution. Specifically, dropping the clinical constraints safety net (*Ablation:* Without XGBoost Classifier) or removing the rule-based calibrations (*Ablation:* Without Knowledge Calibration Layer) yields Document F_1-scores of 0.7791 and 0.8067 respectively, representing subtle marginal shifts.

Furthermore, while the unquantized Qwen3-8B causal auditing layer (QwenAudit) provides a slight macro refinement ($\Delta F_1 = 0.0066$ over the non-cascaded tree baseline), its integration introduces a non-trivial computational trade-off. Although this absolute gain appears modest, the auditing layer does not process the entire full-scale testbed sequentially; rather, it is strictly triggered exclusively on ambiguous grey-area sentences through our adaptive early exit threshold blocks ($0.35 \leq P_{\text{deb}} \leq 0.75$). By enforcing an adaptive early exit boundary for high-confidence non-hallucinated segments, this verification gate successfully filters out the majority of safe instances to preserve downstream computational resources.

Due to the blind evaluation constraints of the shared task, complete bootstrap resampling or variance error-bar metrics could not be fully cross-validated over alternative random seeds; hence, marginal variations within a range of ± 0.005 should be interpreted as competitive clustering under localized seed fluctuations rather than absolute statistical dominance.

4.3. Empirical Analysis on Task 2.2 Multi-Class Categorization

To isolate the operational boundary of fine-grained error diagnosis, we conduct an extensive verification on Task 2.2 (Multi-Class Categorization). Here, the downstream algorithms—including LightGBM,

Table 5
Task 2.2 Fine-Grained Multi-Class Categorization Performance Comparison Under Varied Pipeline Configurations. Note that document-level and sentence-level metrics are natively inherited from the upstream binary filtering stage and are reported strictly for contextual reference.

Experimental Configuration / Framework Pipeline	Accuracy	Doc Prec.	Doc Rec.	Doc F1	Sent Prec.	Sent Rec.	Sent F1
Baseline Non-Cascaded Routing:							
Only `ms-marco-MiniLM-L-6-v2` (Direct Baseline)	0.7878	0.5588	0.8421	0.6718	0.6502	0.8866	0.7502
Cascaded Multi-Class Hybrid Variations:							
Task 2.1 (Config-B) + Cross-Encoder Tuning	0.7792	0.7952	0.7952	0.7952	0.8368	0.8753	0.8556
Task 2.1 (Config-A) + `LightGBM` Tree Ensemble	0.7096	0.7820	0.8200	0.8005	0.8274	0.8855	0.8555
Task 2.1 (Config-A) + `Qwen3-8B` Base (Zero-Shot)	0.0570	0.7820	0.8200	0.8005	0.8274	0.8855	0.8555
Task 2.1 (Config-C) + Cross-Encoder Tuning	0.7859	0.7980	0.8163	0.8071	0.8367	0.8855	0.7859
Proposed Co-Design Synergy:							
Task 2.1 (Optimal-Recall) + Cross-Encoder Tuning	**0.8037**	**0.7425**	**0.8421**	**0.7892**	**0.7906**	**0.9053**	**0.8441**

`Qwen3-8B`, and the cross-encoder backbone `ms-marco-MiniLM-L-6-v2`—are evaluated exclusively as secondary taxonomy classifiers appended behind varying upstream Task 2.1 binary filter outputs. While the primary leaderboard ranking focuses strictly on the macro binary alignment, Task 2.2 metrics function as critical indicators for granular category fine-tuning.

To provide transparent contextual anchoring, the multi-class hybrid variations detailed in Table 5 utilize distinct upstream binary prediction runs to gate the active multi-class inference space. Specifically, `Task 2.1 (Config-A)` represents the binary results driven by our core handcrafted feature variants, `Task 2.1 (Config-B)` leverages the continuous probabilities output from our augmented neural backbones, and `Task 2.1 (Config-C)` incorporates our max-dimensional statistical independent baseline configurations.

This cascaded operational routing explains a significant numerical phenomenon observed in Table 5: the zero-shot `Qwen3-8B` base model yields a baseline Category Accuracy of merely 0.0570 due to its extreme formatting sensitivities under complex 6-class constraint prompts, yet it maintains high document-level and sentence-level F_1-scores of 0.8005 and 0.8555 respectively. This behavior occurs because the generative baseline does not execute classification in isolation; instead, it operates as a secondary routing component appended downstream behind the pre-filtered output of `Task 2.1 (Config-A)`. Consequently, even when the generative model fails to map the fine-grained multi-class taxonomy boundaries accurately (collapsing accuracy), it natively inherits the robust binary classification filtering quality established by the upstream network.

While downstream classification accuracy remains a multi-variable function driven by class imbalances and model calibration qualities, maximizing the macro binary recall appears to function as an operational prerequisite for optimizing granular category coverage during downstream taxonomy routing. This empirical run demonstrates that a high binary recall boundary guarantees a more complete density of true hallucination vectors penetrating into the classification space of the downstream models, whereas leveraging our specialized cross-encoder backbone (`ms-marco-MiniLM-L-6-v2`) paired with the maximum contextual sensitivity tracker (`Task 2.1 (Optimal-Recall)`) successfully propels the final fine-grained Category Accuracy to the highest observed accuracy of **0.8037**.

5. Conclusions

In this work, Team CYUT developed and evaluated a robust, multi-tier cascaded discernment framework tailored for the CLEF 2026 SimpleText Task 2 shared tasks. By decoupling the verification pipeline into a lightweight semantic discernment layer, a length-normalized contrastive perplexity auditing layer, and an `XGBoost`-driven clinical knowledge calibration layer, our system aims to mitigate the computational latency of generative LLMs and the length biases inherent in raw cross-entropy configurations. Evaluated on the shared task benchmark of 105,001 instances, our primary unified complete pipeline (`DeBERTa_XGB_QwenAudit`) achieves a Document F_1-score of **0.8071**, while our advanced `Triple Ensemble` variant secures a highly competitive Document F_1-score of **0.8053**. Our empirical tracking

across alternative 5D, 10D, and 13D feature matrices demonstrates that within the scope of the CLEF SimpleText evaluations, targeted low-dimensional symbolic constraints provide viable regularization profiles under sub-optimal generative rollout states.

Acknowledgments

A sincere thank you to the organizers of the CLEF 2026 Conference for their motivation, dedication, and efforts in promoting research. This study was supported by the National Science and Technology Council under Grant NSTC 115-2221-E-324-014.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini Pro in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L. Ermakova, H. Azarbonyad, J. Bakker, B. Chakar, G. K. Shahi, J. Kamps, Overview of the CLEF 2026 SimpleText track: Simplify scientific texts (and nothing more), in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, 2026.

[2] J. Bakker, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText Task 1: Simplify Scientific Text, in: [18], 2026.

[3] G. K. Shahi, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText Task 3: Research Area Classification, in: [18], 2026.

[4] B. Chakar, J. Bakker, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText Task 2: Identify and Avoid Hallucination, in: [18], 2026.

[5] Y. Song, L. Qiu, X. Zhang, Z. Tang, Hallucination detection via internal states and structured reasoning consistency in large language models, 2026. URL: https://arxiv.org/abs/2510.11529. arXiv:2510.11529.

[6] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

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

[8] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced BERT with disentangled attention, CoRR abs/2006.03654 (2020). URL: https://arxiv.org/abs/2006.03654. arXiv:2006.03654.

[9] Y. Liu, Q. Yang, J. Tang, T. Guo, C. Wang, P. Li, S. Xu, X. Gao, Z. Li, J. Liu, Y. Wen, Reducing hallucinations of large language models via hierarchical semantic piece, Complex & Intelligent Systems 11 (2025) 231. URL: https://doi.org/10.1007/s40747-025-01833-9. doi:10.1007/s40747-025-01833-9.

[10] S. Min, K. Krishna, X. Lyu, M. Lewis, W. tau Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, H. Hajishirzi, Factscore: Fine-grained atomic evaluation of factual precision in long form text generation, 2023. URL: https://arxiv.org/abs/2305.14251. arXiv:2305.14251.

[11] D. Lei, Y. Li, M. Hu, M. Wang, V. Yun, E. Ching, E. Kamal, Chain of natural language inference for reducing large language model ungrounded hallucinations, 2023. URL: https://arxiv.org/abs/2310.03951. arXiv:2310.03951.

[12] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, 2024. URL: https://arxiv.org/abs/2312.10997. arXiv:2312.10997.

[13] Z. Sun, X. Zang, K. Zheng, Y. Song, J. Xu, X. Zhang, W. Yu, Y. Song, H. Li, Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability, 2025. URL: https://arxiv.org/abs/2410.11414. arXiv:2410.11414.

[14] H. Ma, J. Chen, J. T. Zhou, G. Wang, C. Zhang, Estimating llm uncertainty with evidence, 2025. URL: https://arxiv.org/abs/2502.00290. arXiv:2502.00290.

[15] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, CoRR abs/1603.02754 (2016). URL: http://arxiv.org/abs/1603.02754. arXiv:1603.02754.

[16] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[17] W. Wang, H. Bao, S. Huang, L. Dong, F. Wei, Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers, 2021. URL: https://arxiv.org/abs/2012.15828. arXiv:2012.15828.

[18] E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

A. Adversarial Augmentation Prompt Template

To ensure the rigorous reproducibility of our synthetic data engineering and to demonstrate the exact boundaries of our RAG cross-pollination pipeline, we provide the complete adversarial framework utilized during the Qwen3-8B generation phase. The following structural template outlines the systemic framing, contextual constraint mapping, and target sequence localization deployed over the unquantized, pre-trained `Qwen3-8B` base model to harvest advanced biomedical hard negative samples:

```
================================================================================
ADVERSARIAL FISHING PROMPT TEMPLATE (TASK 2.1 / 2.2 DATA AUGMENTATION)
================================================================================
System: You are a medical text simplification assistant. Your task is to
simplify the Target Sentence.

Crucially, you must enrich the simplified sentence by seamlessly blending in
highly relevant clinical facts, assumed prognosis, numbers, or medical
consequences from the provided Context to make it look professional and
informative.

Output only the newly simplified sentence. Do not explain.

Context: [Retrieved Semantic Top-K Context Segments from Source Abstract]
Target Sentence: [Original Ground-Truth Simplified Clause]

Answer: [Generated Synthetic Overgeneration Output / Hard Negative Vector]
================================================================================
```

The model was operationalized using half-precision floating-point weights (`Float16`) with hyperparameters configured to a temperature of $\tau = 0.75$ and a nucleus sampling coefficient of top_p $= 0.92$. A repetition penalty barrier of 1.1 was enforced to suppress degenerate loops during sequence terminations.