VietCheckMed: Explainable Regulatory Compliance Checking for Medical Advertisements on Vietnamese Social Media

Nguyen Thanh Tam[1,3*], **Khanh Quoc Tran**[2,3,4*], **Dat Thanh Pham**[2,3,4], **Truong Phu Le**[1,3], **Nguyen Hoang Gia Han**[1,3], **Binh T. Nguyen**[1,3,4†]

[1]University of Science, Ho Chi Minh City, Vietnam
[2]University of Information Technology, Ho Chi Minh City, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
[4]AISIA Research Lab, Vietnam

Abstract

Regulatory compliance checking for online medical advertisements poses a critical public safety challenge distinct from traditional fact-checking, particularly in low-resource languages. Existing automated systems are ill-suited for the authorization-based, evidence-grounded, and explainable reasoning this task demands. To address this gap, we introduce `VietCheckMed`, a novel retrieval-augmented framework, and `VietAestheticAds`, the first large-scale, expert-validated benchmark for this task, comprising **8,329 advertisements** paired with an authoritative regulatory corpus of **9,978 facilities**. Comprehensive experiments demonstrate that our evidence-grounded approach is essential, substantially outperforming powerful unassisted LLM baselines by over 0.3805 F1-score. A detailed analysis reveals that the primary remaining challenges are nuanced failures in semantic and logical reasoning, defining a clear frontier for future research. To promote advances in regulatory technology and responsible AI, our dataset, code, and evaluation scripts will be made publicly available. This work contributes a foundational methodology and a vital public resource for developing responsible AI in high-stakes regulatory domains.

Code — https://github.com/kh4nh12/VietCheckMed

1 Introduction

The proliferation of digital platforms has created an unprecedented challenge for regulatory oversight, particularly in high-stakes domains like public health. While much of the scientific community's attention has been focused on combating medical misinformation and assessing the veracity of therapeutic claims, a more insidious threat is the advertising of medical services by entities that lack the legal authority to perform them (Lazer et al. 2018). This problem shifts the core task from traditional fact-checking (Vlachos and Riedel 2014) to the more complex challenge of regulatory compliance checking: verifying an advertiser's credentials and offered services against official, often opaque, legal frameworks. This task is especially daunting in regions like Vietnam, where a complex regulatory landscape intersects with a low-resource language and culturally nuanced advertising strategies, making it a challenging testbed for robust AI solutions. The complex, multi-faceted nature of this task is illustrated in Figure 1, and formally defined in Appendix A.

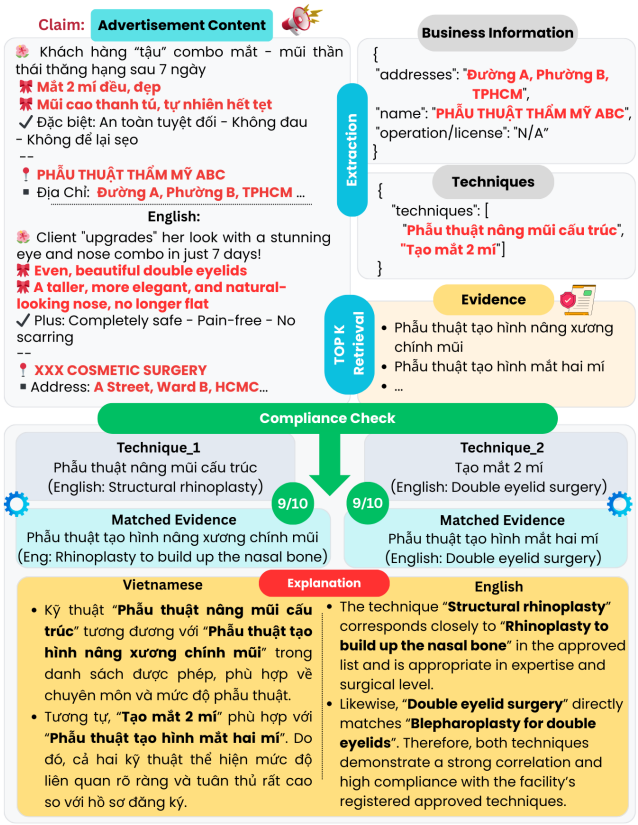

Figure 1: An example of the explainable regulatory compliance checking task.

Existing automated solutions are ill-equipped to tackle this challenge due to three fundamental gaps. First, the **Domain Gap**: Mainstream Automated Fact-Checking (AFC) systems are designed to verify public claims against open-web evidence (Guo, Schlichtkrull, and Vlachos 2022) like news articles or Wikipedia, not to parse and reason over the

*These authors contributed equally.
†Corresponding author.

structured legal corpora that define regulatory compliance (Schuster, Fisch, and Barzilay 2021). Second, the **Resource and Nuance Gap**: State-of-the-art Large Language Models (LLMs) often falter when applied to low-resource languages like Vietnamese, which have unique linguistic properties and lack large-scale, domain-specific annotated corpora (Nguyen and Tuan Nguyen 2020). These models also struggle to interpret the persuasive and culturally-specific rhetoric common in social media advertising. Third, and most critically, is the **Explainability Gap**: For regulatory enforcement and public trust, a simple "compliant" or "non-compliant" verdict is insufficient. Stakeholders require clear, actionable, and evidence-backed explanations for any judgment (Atanasova et al. 2020), a feature that the "black-box" nature of many AFC systems fails to provide.

To address these critical gaps, this paper introduces `VietCheckMed`, the first framework for Vietnamese explainable, evidence-grounded regulatory compliance checking. Our work makes three primary contributions:

1. **A Novel Framework for Grounded Regulatory Reasoning.** We propose a new Explainable Automated Fact-Checking (XAFC) architecture that systematically extracts and verifies business and technical information from unstructured advertisements against an authoritative regulatory corpus, pioneering a granular, technique-level compliance assessment.
2. **The First Large-Scale, Validated Benchmark for Vietnamese Regulatory Compliance.** We construct and will make publicly available `VietAestheticAds`, a new dataset of over 8,300 medical advertisements coupled with a curated corpus detailing the official registrations of over 9,000 medical facilities. To ensure high fidelity and address the inherent risks of purely automated annotation, the dataset has undergone a rigorous validation process involving human domain experts, establishing a reliable gold-standard benchmark for reproducible research in this critical new area.
3. **A Comprehensive Empirical Analysis of LLMs for Regulatory Adherence.** We conduct the first extensive benchmark of leading LLMs, revealing crucial insights into their capabilities for nuanced, evidence-based reasoning in a low-resource, high-stakes domain.

2 Literature Review

Our research is situated at the confluence of four key areas: Automated Fact-Checking (AFC), Explainable AI (XAI), domain-specific regulatory analysis, and low-resource Natural Language Processing (NLP). The novelty of our work lies in the unique challenges that arise at this intersection.

The field of AFC has matured around the task of verifying the veracity of public claims against large-scale, open-domain corpora (Vlachos and Riedel 2014). The canonical AFC pipeline involves claim detection, evidence retrieval, and stance detection to yield a verdict (Guo, Schlichtkrull, and Vlachos 2022). Seminal datasets such as FEVER (Thorne et al. 2018) and its successors like FEVEROUS (Aly et al. 2021) have been instrumental in driving this paradigm. Methodologically, the field has heavily invested in techniques like Retrieval-Augmented Generation (RAG) to ground model outputs in retrieved textual evidence, thereby mitigating the risk of factual hallucination (Lewis et al. 2020). However, the reliance on open-web corpora introduces significant challenges in evidence credibility and noise (Popat et al. 2018). More fundamentally, this paradigm is predicated on an assumption that the "ground truth" exists within a vast, unstructured, and non-authoritative body of text. This assumption breaks down when the task shifts from verifying general-knowledge claims to verifying compliance against a finite, structured, and authoritative legal corpus, where precision and fidelity to the source are paramount.

In parallel, the demand for transparency in high-stakes domains has propelled the growth of XAFC. The objective of XAFC is to augment veracity verdicts with human-understandable justifications. Many effective methods in XAFC provide "shallow" explanations, typically by extracting the specific sentences from a source document that served as evidence (Thorne et al. 2018; Atanasova et al. 2020). While valuable for tracing a verdict to a source, this form of explanation is insufficient for regulatory contexts. A regulator requires a "deep" explanation that articulates an inferential chain. This contrasts with more advanced, program-guided reasoning approaches that can deconstruct a complex claim into verifiable sub-claims but are still focused on veracity rather than authorization (Pan et al. 2023). Current XAFC benchmarks, such as Factcheck-Bench (Wang et al. 2024b), are vital for assessing the factuality of LLM outputs but do not yet focus on evaluating this deeper, multi-step reasoning required for regulatory analysis.

We posit that regulatory compliance checking is a fundamentally distinct task from factual verification. The objective is not to determine if a claim is factually true, but whether an entity is **legally authorized** to offer an advertised service. This shifts the core challenge from a veracity-based problem to an authorization-based one. While prior work has addressed domain-specific AFC, such as in healthcare (Schuster, Fisch, and Barzilay 2021), the focus has remained on the veracity of claims (e.g., "Is this medical claim true?"). The question of provider authorization has remained largely unexplored. To situate our contribution, Table 1 provides a comparative overview of existing fact-checking datasets, highlighting the unique focus of `VietAestheticAds`.

The challenges of regulatory compliance are magnified in a low-resource linguistic context like Vietnamese. The language is characterized by tonal complexities and syntactic ambiguities that challenge standard NLP models (Nguyen and Tuan Nguyen 2020). This is exacerbated by a scarcity of large-scale annotated datasets for specialized tasks. While Vietnamese fact-checking datasets and knowledge graphs are emerging (Hoa et al. 2025; Le et al. 2024; Duong, Ho, and Do 2023), they are focused on veracity in news and open-domain text. The confluence of these challenges: the need for automated verification, the demand for deep explanations, the unique logic of regulatory compliance, and the intricacies of a low-resource language, represents a critical, unaddressed frontier. This work pioneers research in this area, providing the first methodological framework and validated benchmark to fill this explicit gap.

Dataset	Task Type	Domain	Language	# Claims	Evidence Corpus Size
FEVER (2018)	Veracity	Open	English	185k	5.4M Wikipedia Articles
FEVEROUS (2021)	Veracity	Open	English	87k	5.4M Wikipedia Articles
VitaminC (2021)	Veracity	Health	English	36k	100K Wikipedia Articles
ViWikiFC (2024)	Veracity	Open	Vietnamese	21k	73 Vietnamese Wikipedia
ViFactCheck (2025)	Veracity	News	Vietnamese	7k	1,000 News Articles
VietAestheticAds (Ours)	**Authorization**	**Regulatory**	**Vietnamese**	**8.3k**	**9,978 Facility Registrations**

Table 1: A comparative overview of fact-checking datasets. VietAestheticAds is unique in its focus on the **Authorization** task and its use of official registries as an evidence corpus, a key distinction from existing veracity-based benchmarks.

3 The VietCheckMed Framework

To address the multifaceted challenge of explainable regulatory compliance checking, we designed and implemented VietCheckMed, a cascaded, retrieval-augmented reasoning pipeline. The framework's architecture, depicted in Figure 2, systematically transforms an unstructured input advertisement, A, into a structured, evidence-grounded compliance judgment, O. It decomposes this complex task into four distinct stages: (1) entity identification and linking, (2) claim extraction and normalization, (3) evidence retrieval, and (4) evidence-grounded reasoning and explanation. The core LLM functionalities are implemented using powerful instruction-tuned models, with specific prompts and implementation details provided in Appendix B.

3.1 Stage 1: Entity Identification and Linking

The first stage grounds the advertisement in the real world by linking it to a specific legal entity. An instruction-tuned LLM parses the raw advertisement text A to extract a set of key business identifiers, which we denote as the entity profile $E_B = \{\text{name, address, permit\_id}\}$. This module is designed for high recall, capturing multiple potential names or addresses from the often colloquial and unstructured text of social media posts. The output is a structured object to ensure reliable integration with the subsequent retrieval stage.

3.2 Stage 2: Claim Extraction and Normalization

Concurrent to Stage 1, a second LLM-based module identifies the set of all distinct medical services or techniques advertised in A, which we denote as the extracted claims $E_T = \{t_1, t_2, \ldots, t_n\}$. A critical function of this stage is not just extraction but also *normalization*. To facilitate this, the LLM is provided with a curated **Glossary of Medical Terms** as part of its context. This allows the model to map informal, persuasive marketing language (e.g., "age-reversing therapy") to a canonical medical vocabulary (e.g., "mesotherapy"), which is essential for accurate mapping against the formal language of the regulatory corpus.

3.3 Stage 3: Evidence Retrieval

This stage is responsible for retrieving the authoritative ground truth for the entity identified in Stage 1. It queries our **Regulatory Corpus**, $\mathcal{R}$, using the entity profile E_B to retrieve the precise set of legally permitted scopes and techniques, S_{E_B}. This is achieved through a prioritized hybrid search strategy to maximize both precision and recall:

1. **Identifier Search:** A high-precision keyword search is first attempted using the unique "permit_id" from E_B.
2. **Semantic Search:** If no match is found, a fallback vector search is performed using the "name" and "address" fields. This provides robustness against variations in naming and spelling.

For facilities with extensive permissions, a final semantic filtering step selects the top-k techniques from S_{E_B} that are most relevant to the extracted claims E_T. This creates a focused evidence set, $\hat{S}_{E_B} \subseteq S_{E_B}$, balancing signal and noise for the final reasoning stage.

3.4 Stage 4: Evidence-Grounded Reasoning and Explanation

The final stage performs the core compliance assessment by synthesizing outputs from the previous stages. Critically, this stage employs an **adaptive reasoning process** that tailors its context based on the facility type to reflect distinct regulatory structures. For facilities with granular, itemized permissions (e.g., Healthcare clinics), the model performs direct semantic matching between claimed and permitted techniques. Conversely, for those governed by broader scopes of practice (e.g., Beauty Salons), the model's reasoning is augmented with our codified **Medical Regulatory Knowledge** base (K) to ensure correct legal interpretation.

The LLM is prompted to perform a two-level analysis: (a) a granular, technique-level assessment for each claim $t_i \in E_T$ against the evidence $\hat{S}_{E_B}$, and (b) a holistic, advertisement-level assessment that synthesizes the granular findings. The final output, O, is a structured JSON object containing these multi-level compliance scores, along with a detailed, human-readable explanation in Vietnamese that justifies each verdict by explicitly referencing the provided evidence and regulatory knowledge.

4 The VietAestheticAds Benchmark

A core contribution of this work is the creation and validation of VietAestheticAds, a new benchmark designed to facilitate research in explainable regulatory compliance checking. Unlike traditional fact-checking datasets, which focus on verifying the *veracity* of claims, VietAestheticAds is the first large-scale resource designed for the task of verifying the *authorization* of advertised services against a formal regulatory corpus in a low-resource language. This section details the principles, curation pipeline, and analytical properties of this new benchmark.

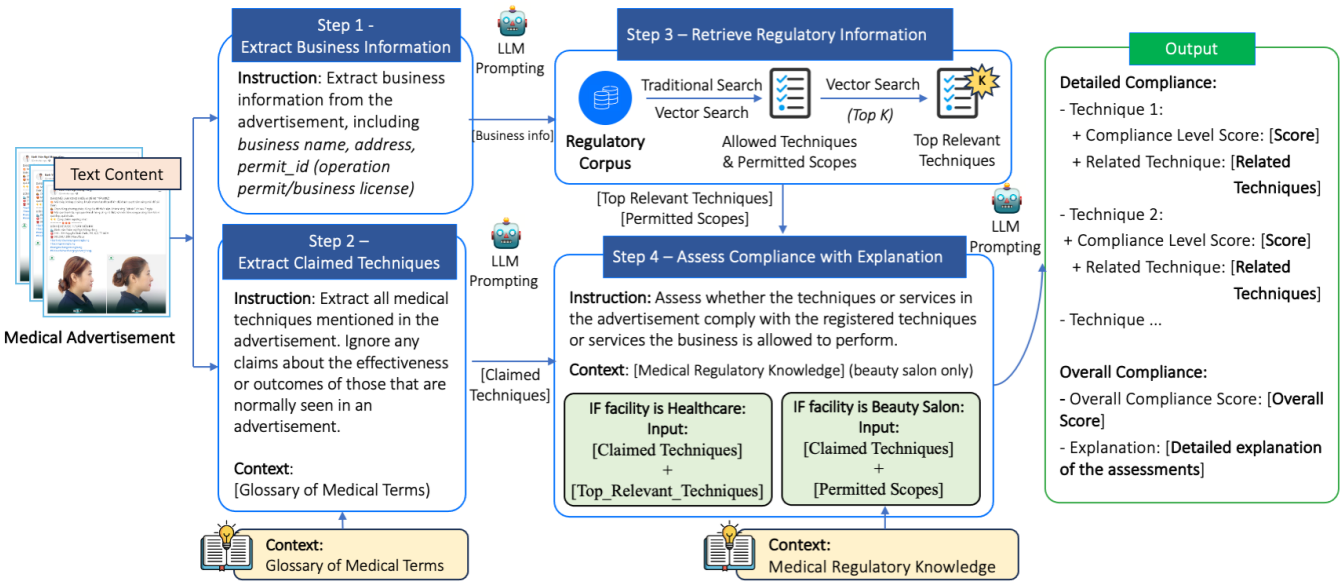

Figure 2: VietCheckMed Pipeline. All the tasks that need LLM are annotated with the Robotic icon. Contexts are also provided to help the model understand the knowledge in the medical domain.

4.1 Dataset Curation and Design Principles

The construction of `VietAestheticAds` was guided by key principles to ensure its utility as a challenging scientific instrument. We prioritized **Authoritative Grounding**, ensuring every advertisement is linkable to a verifiable, official regulatory source. We focused on **Linguistic Realism** by capturing the full spectrum of language, from formal legal terminology to informal marketing slang. Finally, we engineered for **Task Complexity** by including posts with both single and multiple claims, enabling a robust evaluation of compositional reasoning.

This principled design was realized through a systematic curation pipeline. The benchmark consists of two aligned corpora: an **Advertisement Corpus** of 8,329 posts from public social media pages, and a **Regulatory Corpus** containing the official registration details for 9,978 medical facilities. For this foundational study, the dataset is geographically focused on Ho Chi Minh City (HCMC). This was a strategic decision, as HCMC represents the largest, most data-rich market for the industry in Vietnam and was the only jurisdiction with the necessary public regulatory data available during our research. A full justification for this methodological choice is provided in Appendix C.

4.2 A Hybrid Human-AI Annotation Protocol

To label the dataset at scale while ensuring the highest level of data fidelity, we developed a hybrid annotation protocol. Our process diverges from conventional dataset creation by being grounded in the noisy reality of existing advertisements. The protocol involved a pilot phase to refine comprehensive guidelines (Appendix D), followed by a large-scale initial annotation using an instruction-tuned LLM. A 20% subset then underwent meticulous human review to create the final Gold Standard (GS). To scientifically validate the reliability of our dataset, this GS was then validated in a final study with two external domain experts, who achieved substantial inter-annotator agreement (Krippendorff's $\alpha = 0.91$). Our GS labels demonstrated near-perfect agreement with the expert consensus ($\kappa = 0.97$), confirming the benchmark's validity.

4.3 Analytical Properties and NLP Challenges

`VietAestheticAds` is partitioned into development (50%), and testing (50%) sets. An analysis of its properties reveals characteristics that pose specific NLP challenges. For instance, non-compliant advertisements are statistically longer and more lexically diverse, challenging a model's capacity for **salient information detection** and **robustness to lexical diversity**.

Score Range	Percentage (%)
0.8 – 1.0 (High Similarity)	0.43%
0.6 – 0.8 (Moderate Similarity)	9.23%
0.0 – 0.6 (Low Similarity)	90.34%

Table 2: Distribution of Semantic Similarity Scores between Advertised Techniques and Permitted Scopes.

The core difficulty lies in bridging the semantic gap between ad text and legal text. We quantified this by analyzing the semantic similarity between advertised techniques and their corresponding permitted scopes in non-compliant cases. The analysis shows that 90.34% of these cases exhibit low semantic similarity (Table 2). This empirically proves that simple keyword matching is insufficient; the benchmark demands that models perform **deep semantic reasoning** to succeed. A comprehensive analysis and datasheet are provided in Appendices E and F.

Model	With gold evidence				Without gold evidence		
	F1	P	R	Δ F1	F1	P	R
Beauty							
Large-Scale Models (>50B)							
Llama 4 Maverick	0.8042	0.8663	0.7504	0.1406	0.6636	0.8498	0.5444
Llama 3.3 †	**0.8698**	**0.8812**	0.8586	0.1150	**0.7548**	0.7671	0.7429
Qwen3	0.7303	0.5845	0.9728	0.0121	0.7182	0.5737	0.9602
High-Efficiency Models (<30B)							
Deepseek V3	0.7351	0.5827	**0.9955**	0.0043	0.7308	0.5758	**1.0000**
Gemma-2	0.7931	0.6901	0.9322	0.0560	0.7371	0.6282	0.8917
Qwen2.5	0.7496	0.6056	0.9835	−0.0029	0.7525	0.6509	0.8917
Mistral Small	0.7550	0.8467	0.6812	0.2082	0.5468	0.8280	0.4082
Healthcare							
Large-Scale Models (>50B)							
Llama 4 Maverick	0.5031	0.9127	0.3473	0.4846	0.0185	0.8125	0.0093
Llama 3.3	0.6624	0.9163	0.5187	0.6453	0.0171	**1.0000**	0.0086
Qwen3 †	**0.7171**	0.7550	0.6828	0.0824	**0.6347**	0.4743	**0.9591**
High-Efficiency Models (<30B)							
Deepseek V3	0.6616	0.5095	**0.9433**	0.1129	0.5487	0.4694	0.6602
Gemma-2	0.6824	0.7911	0.6000	0.6008	0.0816	0.5980	0.0438
Qwen2.5	0.3037	**0.9301**	0.1815	0.0619	0.2418	0.5788	0.1528
Mistral Small	0.1927	0.8824	0.1081	0.1757	0.0170	0.6316	0.0086

Table 3: Comparative Performance of Baseline LLMs on Compliance Classification, grouping models by scale. The **Grounding Impact (Δ F1)** column quantifies the absolute performance drop when models operate without evidence. The best result in each column is bolded. The overall top-performing model in each category is marked with †.

5 Experiments and Results

We conducted a series of experiments to evaluate `VietCheckMed` and benchmark modern LLMs on our novel regulatory compliance task. Our evaluation investigates three key aspects: (1) the effectiveness of leading LLMs on this new challenge, (2) the impact of advertisement complexity and evidentiary context on performance, and (3) the contributions of our framework's components and the nature of primary failure modes.

5.1 Experimental Setup

Baselines and Setups We benchmarked a diverse suite of leading language models, including variants from the Llama (Touvron et al. 2023), Gemma (Team et al. 2024), Qwen (Team 2024), DeepSeek (Liu et al. 2024), and Mistral (Jiang et al. 2023). To ensure reproducibility, all experiments used deterministic decoding (temperature to 0.0) and a top-k retrieval of k=10, a value informed by a sensitivity analysis (Appendix G). Full specifications for each baseline and our experimental configuration are detailed in Appendix H.

Evaluation Metrics The primary task is compliance classification. Performance is measured using macro-average Precision, Recall, and F1-Score (Powers 2020). For the quality of explanations, we introduce and apply the **FIDES-Score**, a novel multi-dimensional metric that assesses explanations on key principles of explainability, including fidelity, justification soundness, and clarity (Zheng et al. 2023). The FIDES-Score is implemented via a validated LLM-as-a-Judge framework; its full methodology and validation against human experts are detailed in Appendix I.

5.2 Main Results and Analysis

Comparative Performance and the Primacy of Evidentiary Grounding Our main experiments (Table 3) reveal a clear performance differential among leading LLMs and offer a crucial insight into the nature of this task. The superior performance of instruction-tuned models like Llama 3.3 and Gemma-2 suggests that the fine-grained, rule-based reasoning required for regulatory compliance benefits significantly from advanced instruction-following capabilities. Interestingly, some models like DeepSeek-V3 exhibit a "high-sensitivity, low-precision" profile, which could be suitable for a first-pass, automated screening system designed to flag all potentially non-compliant content for human review. In contrast, the more balanced profile of Llama 3.3 makes it a better candidate for a system where precision is paramount.

However, the central and most unequivocal finding of this analysis is the absolute dependency on evidentiary grounding. Across all models, removing access to the regulatory corpus causes a catastrophic performance collapse, evidenced by an F1-score drop exceeding 64 points for Llama 3.3 in the Healthcare domain. This result provides dispositive proof that for authorization-based reasoning tasks, where truth is defined by a specific, non-public, and dy-

Model	Single Extracted Technique			Multiple Extracted Techniques		
	F1-score	Precision	Recall	F1-score	Precision	Recall
Beauty Domain						
Deepseek V3	0.7322	0.5796	**0.9937**	0.7378	0.5856	0.9971
Llama 4 Maverick	0.7979	0.7270	0.8842	0.8096	0.8517	0.7714
Llama 3.3	**0.8906**	0.8892	0.8921	**0.8504**	**0.8735**	0.8286
Gemma-2	0.8584	0.8246	0.8952	0.7456	0.6072	0.9656
Qwen3	0.7231	0.5683	0.9936	0.7372	0.6007	0.9539
Qwen2.5	0.7581	0.6242	0.9651	0.7423	0.5902	**1.0000**
Mistral Small	0.7519	**0.9426**	0.6254	0.7574	0.7853	0.7314
Healthcare Domain						
Deepseek V3	0.6705	0.5168	**0.9542**	0.6477	0.4980	**0.9262**
Llama 4 Maverick	0.5169	0.9242	0.3588	0.4809	0.8934	0.3290
Llama 3.3	0.7080	0.9178	0.5763	0.5829	0.9134	0.4280
Gemma-2	0.6706	0.7757	0.5905	**0.7021**	0.8174	0.6152
Qwen3	**0.7778**	0.8104	0.7477	0.6169	0.6609	0.5784
Qwen2.5	0.3597	0.9095	0.2242	0.2053	**1.0000**	0.1144
Mistral Small	0.1615	**0.9740**	0.0880	0.2389	0.8065	0.1402

Table 4: Impact of Input Complexity on Classification Performance. Best score in each column per metric is bolded.

namic body of legal fact, a model's latent parametric knowledge is fundamentally inadequate. A retrieval-augmented architecture is, therefore, not an enhancement but the prerequisite for any viable solution.

Analysis of Performance Disparity Across Categories A notable performance gap exists between the "Beauty" and "Healthcare" categories, a discrepancy we attribute to several factors related to domain complexity. First, "Healthcare" regulations exhibit higher **legal-linguistic complexity**; the terminology is more precise and the scope of permitted procedures is more granular, increasing the risk of misinterpretation. Second, the larger **scale and diversity of the evidence corpus** for healthcare facilities creates a more challenging "needle-in-a-haystack" retrieval problem. Finally, "Healthcare" claims show greater **semantic ambiguity**, as advertisers often use euphemistic language to describe medical procedures. This analysis suggests regulatory compliance is not a monolithic task. Performance is highly sensitive to domain-specific characteristics, a key insight for developing future specialized AI systems.

The Impact of Input Complexity on Reasoning Coherence To probe the models' reasoning capabilities further, we analyzed their performance as a function of input complexity, specifically, the number of distinct techniques advertised in a single post (Table 4). This tests a model's ability to maintain "reasoning coherence" under an increased cognitive load. For a robust model like Llama 3.3, the F1-score shows only a slight decrease when handling multiple techniques, suggesting strong compositional generalization.

Conversely, a model like Gemma-2 exhibits a notable precision-recall trade-off. Its shift towards higher recall at the expense of precision suggests a potential failure mode where the model, unable to precisely resolve multiple distinct claims against their respective evidence, defaults to a more 'cautious' heuristic of flagging any potential ambiguity. This highlights the critical challenge of maintaining stable reasoning under varying input complexity, a key factor for the reliability of any real-world deployment. Beyond these quantitative scores, the

Benchmarking Explanation Quality Beyond classification accuracy, a critical component of our framework is its ability to generate faithful, human-understandable explanations. To evaluate this, we benchmarked the generated explanations from each baseline using our FIDES-Score metric, which assesses fidelity, justification soundness, and clarity.

Model	F	J	C	FIDES
Llama 4 Maverick	4.1	3.9	4.3	4.1
Llama 3.3	**4.5**	**4.4**	**4.6**	**4.5**
Qwen3	3.8	3.5	4.1	3.8
Deepseek V3	3.4	3.2	3.9	3.5
Gemma-2	4.2	4.0	4.4	4.2
Qwen2.5	3.6	3.3	4.0	3.6
Mistral Small	3.9	3.8	4.2	4.0

Table 5: Explanation Quality Benchmark using the FIDES-Score. The overall best performance is in bold. Abbreviations used are: F (Fidelity), J (Justification), and C (Clarity).

The results, presented in Table 5, reveal that explanation quality varies significantly across models. We find that performance on this task correlates strongly with advanced instruction-following capabilities. Llama 3.3, the top-performing model in classification, also generates the highest quality explanations, achieving an overall FIDES-Score of 4.5. This suggests that the same capabilities that enable accurate reasoning also contribute to a model's abil-

ity to articulate that reasoning process faithfully.

Interestingly, we note that Gemma-2, while not the top classifier, produces explanations of a significantly higher quality than several larger models, making it a compelling choice for applications where both performance and efficiency are concerns. These findings underscore the importance of evaluating explainability as a critical dimension of model performance in high-stakes regulatory domains.

5.3 Analysis and Discussion

For our in-depth analyses, we selected **Llama 3.3**, justified by its SOTA, balanced performance in our benchmarks (Table 3), and its role as a representative foundation model.

Modular Ablation Study To quantify the value of our architectural decisions, we performed a modular ablation study, systematically removing key components of the `VietCheckMed` framework (Table 6). The results reveal a clear hierarchy of importance. The removal of the **evidence retrieval** module remains catastrophic, causing a 0.3805 drop in the Macro F1 score. This confirms that for authorization-based tasks, verifiable, retrieved evidence is the non-negotiable cornerstone of reliable reasoning. The impact is particularly stark in the 'Healthcare' domain, where performance collapses to near zero, underscoring its reliance on detailed, retrieved permission lists.

	F1-Score		
Configuration	**Beauty**	**Healthcare**	**Δ F1**
Full Framework	**0.8586**	**0.6624**	**0.0000**
w/o Medical Glossary	0.6480	0.6380	0.1175
w/o Reg. Knowledge	0.5383	–*	0.3203
w/o Retrieval	0.7429	0.0171	0.3805

*Reg. Knowledge is only used for the Beauty domain.

Table 6: Results of the Ablation Study. The table shows the F1-Score for each category and the overall Performance Drop (Δ F1) from removing key components. The color intensity visually represents the magnitude of the impact.

The ablation of other components reveals more nuanced dependencies. Removing the **Medical Glossary** significantly degrades performance across both domains, confirming the universal importance of claim normalization at the start of the pipeline. Crucially, the ablation of the **Regulatory Knowledge** base validates our adaptive reasoning design. As per our methodology, this component is exclusively applied to interpret the broader scopes of practice for Beauty Salons. Consequently, its removal causes a severe performance drop in the 'Beauty' domain (from 0.8586 to 0.5383 F1), while, as expected, having no impact on the 'Healthcare' domain. This result demonstrates the importance of augmenting reasoning with explicit legal rules for specific, complex regulatory contexts.

Qualitative Error Analysis Our manual analysis of 400 incorrect predictions reveals residual errors stem from nuanced reasoning, not mechanical failures. The primary error categories are **Semantic Mismatch**, a failure of comprehension with domain-specific terms, and **Inference Hallucination**, a failure of logical fidelity where models generate conclusions not entailed by the evidence. Other issues include **Evidence Retrieval Failure**. Critically, a **0% Technique Extraction Error** validates our robust claim normalization module. This pinpoints the research frontier: improving the core semantic and logical reasoning of LLMs, even when fully grounded in evidence. A full set of examples for each error category is provided in Appendix J.

Error Category	**Percentage (%)**
Semantic Mismatch	42%
Inference Hallucination	35%
Evidence Retrieval Failure	23%
Technique Extraction Error	0%

Table 7: Distribution of Error Types.

Synthesizing the Analyses Our experiments collectively demonstrate that while powerful, modern LLMs are insufficient for this regulatory task without significant architectural support (Table 3). Performance is critically dependent on domain and input complexity (Table 4) but is most influenced by our framework's scaffolding (Table 6). Our analysis confirms that retrieval is paramount, while claim normalization and explicit legal rules are also significant contributors. Finally, our error analysis illuminates the research frontier: improving the nuanced semantic and logical fidelity of LLMs, even when fully grounded in evidence.

6 Conclusion and Future Work

This paper introduced `VietCheckMed`, a novel framework for the critical and understudied task of explainable regulatory compliance checking for medical advertisements. We make three primary contributions: (1) a modular, retrieval-augmented framework for evidence-grounded reasoning; (2) `VietAestheticAds`, the first large-scale, expert-validated benchmark for this task; and (3) a comprehensive empirical analysis that reveals the capabilities and limitations of modern LLMs for regulatory adherence. Our work provides a foundation methodology and resource for advancing regulatory technology and responsible AI.

Our limitations define a clear research roadmap. We will expand the dataset's geographic scope beyond Ho Chi Minh City and investigate methods to mitigate LLM-induced annotation bias (Wang et al. 2024a). Our error analysis shows the primary challenge is improving model reasoning; we will therefore enhance semantic understanding through domain-specific fine-tuning and bolster logical fidelity by exploring advanced prompting strategies (Wei et al. 2022; Yao et al. 2023a). Ultimately, our vision is to evolve `VietCheckMed` into an autonomous agentic system for real-time monitoring, grounding our work in recent advances in LLM agents (Yao et al. 2023b; Mialon et al. 2023).

Ethics Statement

Our research was conducted with a strong commitment to ethical standards in data collection, privacy, and the responsible dissemination of artifacts. The `VietAestheticAds` dataset was sourced entirely from publicly available information, including official government health portals and public-facing social media pages, in alignment with relevant legal frameworks for data use in Vietnam. We have taken meticulous steps to protect privacy by systematically redacting or masking all business-identifying information, such as phone numbers and street addresses, in our released data.

To ensure transparency and foster future research, the complete, anonymized `VietAestheticAds` dataset, along with our evaluation scripts and the source code for the `VietCheckMed` framework, will be made publicly available upon publication under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. A comprehensive Datasheet, providing granular details on all aspects of the dataset's motivation, curation, and intended use, is provided in Appendix E.

References

Aly, R.; Guo, Z.; Schlichtkrull, M. S.; Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Cocarascu, O.; and Mittal, A. 2021. The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task. In Aly, R.; Christodoulopoulos, C.; Cocarascu, O.; Guo, Z.; Mittal, A.; Schlichtkrull, M.; Thorne, J.; and Vlachos, A., eds., *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, 1–13. Dominican Republic: Association for Computational Linguistics.

Atanasova, P.; Simonsen, J. G.; Lioma, C.; and Augenstein, I. 2020. Generating Fact Checking Explanations. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7352–7364. Online: Association for Computational Linguistics.

Duong, H. T.; Ho, V. H.; and Do, P. 2023. Fact-checking vietnamese information using knowledge graph, datalog, and kg-bert. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(10): 1–23.

Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10: 178–206.

Hoa, T. T.; Duy, T. Q.; Tran, K. Q.; and Van Nguyen, K. 2025. ViFactCheck: A New Benchmark Dataset and Methods for Multi-domain News Fact-Checking in Vietnamese. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 308–316.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. 2018. The science of fake news. *Science*, 359(6380): 1094–1096.

Le, H. T.; To, L. T.; Nguyen, M. T.; and Van Nguyen, K. 2024. Viwikifc: Fact-checking for Vietnamese Wikipedia-based textual knowledge source. *arXiv preprint arXiv:2405.07615*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Mialon, G.; Dessì, R.; Lomeli, M.; Nalmpantis, C.; Pasunuru, R.; Raileanu, R.; Rozière, B.; Schick, T.; Dwivedi-Yu, J.; Celikyilmaz, A.; et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Nguyen, D. Q.; and Tuan Nguyen, A. 2020. PhoBERT: Pre-trained language models for Vietnamese. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1037–1042. Online: Association for Computational Linguistics.

Pan, L.; Wu, X.; Lu, X.; Luu, A. T.; Wang, W. Y.; Kan, M.-Y.; and Nakov, P. 2023. Fact-Checking Complex Claims with Program-Guided Reasoning. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6981–7004. Toronto, Canada: Association for Computational Linguistics.

Popat, K.; Mukherjee, S.; Yates, A.; and Weikum, G. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 22–32. Brussels, Belgium: Association for Computational Linguistics.

Powers, D. M. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Schuster, T.; Fisch, A.; and Barzilay, R. 2021. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 624–643. Online: Association for Computational Linguistics.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Team, Q. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 809–819. New Orleans, Louisiana: Association for Computational Linguistics.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vlachos, A.; and Riedel, S. 2014. Fact Checking: Task definition and dataset construction. In Danescu-Niculescu-Mizil, C.; Eisenstein, J.; McKeown, K.; and Smith, N. A., eds., *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 18–22. Baltimore, MD, USA: Association for Computational Linguistics.

Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Kong, L.; Liu, Q.; Liu, T.; and Sui, Z. 2024a. Large Language Models are not Fair Evaluators. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9440–9450. Bangkok, Thailand: Association for Computational Linguistics.

Wang, Y.; Gangi Reddy, R.; Mujahid, Z. M.; Arora, A.; Rubashevskii, A.; Geng, J.; Mohammed Afzal, O.; Pan, L.; Borenstein, N.; Pillai, A.; Augenstein, I.; Gurevych, I.; and Nakov, P. 2024b. Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 14199–14230. Miami, Florida, USA: Association for Computational Linguistics.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023b. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623.