Transforming AI from Generalist Tools to Specialised Teammates

Measuring Accuracy, Cost, and Efficiency Gains in Human-in-the-Loop Regulatory Workflows

Max Ashton-Lelliott* and Robert Wardrop**

September 15, 2025

*Senior Data Scientist, Regulatory Genome Development Ltd (RegGenome).
**Management Practice Professor of Finance & Founder of the Regulatory Genome Project, Cambridge Judge Business School.

The Regulatory Genome Project is a public-private research initiative hosted in the Cambridge Judge Business School that brings regulators and standard-setters together to co-develop regulatory obligation taxonomies as public goods. The private partner in the Regulatory Genome Project is RegGenome, a spin-out of CJBS that works with firms and regulatory bodies to convert legislative and regulatory texts into structured data that can be used directly by regulatory compliance and intelligence solutions.

Contents

Logline

This research quantifies the value of specialised SLMs in human-in-the-loop workflows, demonstrating a 38% relative accuracy gain over leading generalist LLM, Gemini 2.5 Pro, while running at 1/80th the cost and consuming an estimated 1/200th the energy.

Synopsis

This research contrasts state-of-the-art generalist LLMs with specialised SLMs to demonstrate the latter's value in a human-in-the-loop context. By evaluating models on accuracy, cost, speed, stability, and estimated energy consumption, the study shows how performance metrics translate into tangible workflow efficiencies. The analysis spans both established regulatory domains (AML) and emerging areas (cryptocurrency), revealing that specialisation advantages amplify in novel regulatory territories where the `RegGenome` model achieves a 72% relative improvement in complex classification. Specialised SLMs exhibited near-zero prediction variance, contrasting sharply with the instability observed in provider-hosted models. Furthermore, by operating at less than 1/80th cost and an estimated 1/200th the energy consumption, specialised SLMs enable rapid, cost-effective review cycles. The research also validates zero-shot generalisation capabilities through `RegGenome unified`, which provides viable first-pass classification on unseen taxonomies. These findings transform the AI from a simple tool into a viable teammate for regulatory professionals, reducing human review time by up to 40%.

Abstract

This research contrasts state-of-the-art generalist LLMs with specialised SLMs for regulatory document classification, specifically through the lens of a human-in-the-loop (HITL) workflow. In agentic AI systems where multiple models collaborate sequentially, each step requires maximum fidelity, as a single point of failure cascades through the entire pipeline. The central thesis argues that for AI to be a viable part of this workflow, it must provide not only accuracy but also the speed, stability, and resource efficiency necessary for effective collaboration. We present empirical results showing that the specialised SLM `RegGenome`, excels across five dimensions critical for HITL success:

- **Analyst Efficiency (Established Domains)**: Outperforming leading provider-hosted LLM Gemini 2.5 Pro by 16 percentage points (a 38% relative gain) on a 164-class AML regulatory taxonomy.
- **Amplified Advantage (Emerging Domains)**: Achieving 21 percentage points improvement (72% relative gain) over provider-hosted models on cryptocurrency regulation classification.
- **Workflow Velocity (Speed & Cost)**: Operating at 1/80th the cost and an estimated 1/200th the energy, enabling rapid, iterative review cycles that are economically unfeasible with provider-hosted models.
- **Predictable Collaboration (Stability)**: Exhibiting variance below measurement threshold (<0.001), contrasting with provider-hosted models showing cosine distances up to 0.170.
- **Adaptive Capability**: The `RegGenome unified` model achieves 0.38 Weighted F1-Score on unseen taxonomies without task-specific training, providing immediate utility for emerging regulatory frameworks.

A case study of a three-step document processing workflow demonstrates that these advantages compound multiplicatively, with the specialised approach successfully processing 64% more documents end-to-end. Estimates of human effort indicate up to 40% reduction in review time compared to manual annotation.

1 Introduction: The Importance of Trustworthy AI in Regulatory Technology

1.1 The Double-Edged Sword of LLMs in Regulation

The popularisation of Large Language Models (LLMs) has caused a frantic shift across numerous industries, with the legal sector poised for perhaps one of the most significant adoptions of the technology.[1] These models exhibit a capacity to process and generate human-like text, offering the potential to automate and/or improve a wide range of legal tasks, including document analysis, legal research, drafting, and summarisation. The potential to drive efficiency, reduce costs, and enhance the capabilities of legal professionals is immense. [Savelka, J., & Ashley, K. (2023), Sarkar, R., et al. (2021)]

However, this potential can be seen as a double-edged sword. The complexity that allows these models to achieve their vast gains also makes them extremely opaque, creating a "black-box" problem that is particularly acute in a field where precision and accountability are paramount. This apprehension is reflected in a significant industry-wide perception of a "trust deficit". According to industry reports, almost 60% of legal professionals are concerned about AI-generated "hallucinations" (the generation of plausible but factually incorrect information) and a similar proportion of survey respondents harbour security concerns. In the legal domain, where accuracy, accountability, and consistency are paramount, an AI system that misinterprets legal text, and does so in a constantly changing and unpredictable manner, is not merely unhelpful, it is a significant liability. [Ho, D. E. (2024), NLS Business Law Review. (2024)]

This challenge is compounded by the unique nature of regulatory language itself. Policy and regulatory text functions not only as legal writing but as a specialised dialect with distinct linguistic properties: hierarchical definitional structures, cross-referential frameworks, precise technical terminology, and standardised logical constructs. Unlike general legal language which often requires interpretive flexibility, regulatory language operates through codified, rule-based patterns that map consistently to compliance obligations. This structural regularity makes regulatory classification fundamentally different from broader legal AI tasks and, as we will demonstrate, makes it particularly amenable to specialisation through targeted fine-tuning.

The prevailing sentiment of concern towards potential AI errors logically leads to a strong preference for collaborative systems. One survey investigation how legal departments are using AI found that 66% of legal professionals believe AI should only play a supporting role, with human lawyers retaining ultimate control and responsibility and only 1% believed in full automation. It is within this human-in-the-loop (HITL) framework, where AI serves as a teammate rather than an autonomous agent, that the true value of the technology can be assessed. The central thesis of this paper is that for this partnership to be effective, the AI teammate must be not only accurate but also cost-effective and, most critically, stable. [Counselwell & Spellbook. (2025)]

1.2 Limitations of Accuracy: MAD Models

Conventional benchmarks for AI performance, such as accuracy or F1-score, are insufficient for evaluating the suitability of LLMs for regulatory applications. While important, they fail to capture a more fundamental prerequisite for trust: stability. An AI model is stable if it produces consistent and predictable outputs for identical inputs over multiple runs. This concept is intrinsically linked to the broader challenge of reproducibility, which has been identified as a growing *reproducibility crisis* within AI research. Many published results are difficult to replicate due to a variety of factors, including the inherent stochasticity of models, differences in hardware

[1] Recent reports confirm this trend, with a 2025 Forbes Advisor analysis noting that the UK legal sector has a 29.2% AI adoption rate, while a separate American Bar Association survey found that 31% of legal professionals personally use AI tools, though firm-wide integration remains more limited. [Forbes Advisor. (2025), ABA. (2025)]

and software environments, and a lack of transparency in training and evaluation methodologies. [Szalontai, B., et al. (2025), Andreas H., et al. (2025)]

Yet while this is billed as primarily an academic crisis, it has tangible implications for business. A model whose performance is not reproducible is, by definition, unreliable. In a regulatory setting, an unstable model that classifies the same document differently on separate occasions, or provides conflicting answers to the same query, undermines user trust and introduces unacceptable operational and legal risk. The challenge is not just that models can be wrong, but that their wrongness can be unpredictable. Consequently, any serious evaluation of AI for regulatory assessment use must prioritise the measurement and assurance of performance stability as a primary criterion for viability. [Ho, D. E. (2024), Blair-Stanek, A., & Van Durme, B. (2025)]

1.3 Research Objectives and Report Structure

This report presents the findings of a study designed to address these critical issues. The primary objectives of this research are fivefold:

1. To develop and apply a methodological framework for quantifying the performance instability of leading provider-hosted LLMs on a real-world, hierarchical regulatory document classification task.
2. To conduct a comprehensive comparison of these models against RegGenome's self-hosted SLMs, which are specialised and fine-tuned across the crucial dimensions of accuracy, cost-efficiency, stability, and energy consumption.
3. To investigate the impact of regulatory turbulence on model performance by examining classification accuracy across both established (AML) and emerging (cryptocurrency) regulatory domains, thereby testing the hypothesis that specialisation advantages amplify when dealing with novel regulatory frameworks.
4. To demonstrate through practical multi-agent workflow analysis how incremental accuracy improvements at individual steps compound into end-to-end efficiency gains in human-in-the-loop systems, providing quantitative evidence for the business value of specialised models.
5. To explore the viability of zero-shot generalisation (the ability to classify documents according to a new taxonomy without any specific training examples) as a strategic capability for adapting to the evolving nature of regulatory landscapes, and to articulate the broader strategic implications of self-hosting versus API-based approaches for enterprise adoption.

Our evaluation employs a two-tiered taxonomy structure. Level 1 (L1) comprises broad regulatory domains (e.g., 'Customer Due Diligence', 'Wire-transfers'), while Level 2 (L2) contains 164 granular obligations nested within these domains. This hierarchical structure mirrors real-world regulatory frameworks where specific requirements roll up into broader compliance categories. Examples of the data can be found in Appendix A.3.

Throughout this paper, we focus specifically on the classification of regulatory and policy text rather than on broader legal document processing. This distinction is important: while 'Legal LLMs' typically encompass contract analysis, case law interpretation, and negotiated language, RegGenome's work targets the structured, rule-based framework of regulatory compliance texts. This focus on codified policies and regulations requires different optimisation strategies than general legal language processing.

The remainder of this report guides the reader from the foundational problem of instability to the strategic implications of our findings.

- Section 2 introduces the methodology for measuring variance and presents the empirical results for API-based models.

- Section 3 provides a head-to-head comparison of all tested models, demonstrating the advantages of a specialised approach.
- Section 4 details our experiment in zero-shot learning and discusses its business value.
- Section 5 covers a case study that exemplifies the results of the previous sections and demonstrates the potential business value of specialisation.
- Section 6 combines these findings into a broader discussion of the strategic imperatives for AI in the policy making and regulatory sector.
- Finally, Section 7 offers a concluding summary of the research.

1.4 Definitions

From here on, we will use the following working definitions. First, with respect to model size we adopt the following definitions:

- An SLM is a LM that can fit onto a common consumer electronic device and perform inference with latency sufficiently low to be practical when serving the agentic requests of one user.
- An LLM is a LM that is not an SLM.

[Belek, P., et al. (2025)]

Secondly on hosting, we adopt the following definitions:

- A **self-hosted model** is a machine learning model deployed on user-controlled infrastructure (local hardware or private cloud) where the user maintains direct access to model weights, full control over inference parameters (sampling, quantisation, batch size), and autonomy over the computational environment and hardware specifications.
- A **provider-hosted model** (or **API model**) is a machine learning model deployed and maintained by a third-party service provider, accessed through standardised APIs with limited configuration control, where the provider manages the infrastructure, model versioning, and inference pipeline while restricting access to model weights and low-level parameters.

1.5 Evaluation Framework and Taxonomy

All evaluations in this study employ RegGenome's taxonomies. RegGenome's taxonomies are aligned with the taxonomies developed by the Cambridge Regulatory Genome Project in the Cambridge Judge Business School which collaborates with regulators to create taxonomies reflecting the standards and principles published by multi-lateral standard setting bodies. [2]

The evaluation exercise is a two-tiered regulatory document classification task. For the Anti-Money Laudering (AML) taxonomy, Level 1 (L1) consists of 19 high-level nodes representing broad regulatory domains, while Level 2 (L2) comprises 164 more granular child nodes representing specific obligations. Similarly, the Cryptocurrency taxonomy consists of 25 L1 nodes and 190 L2 nodes. We calculated standard classification metrics, Precision, Recall, and F1-Score, for each class. For summary reporting, the **Weighted F1-Score** serves as the primary single-number metric. This metric is particularly well-suited for this task as it accounts for class imbalance, a common characteristic of complex classification datasets where some classifications are far more prevalent than others, by weighting the F1-Score of each class by its support (the number of true instances for that class).

[2] In the case of Anti-Money Laundering regulations the most influential standards are published by the Financial Action Task Force (FATF).

Cost figures in the following sections are derived directly from API service provider billing for provider-hosted LLMs. For self-hosted SLMs, costs are calculated based on completion time and hardware requirements. All self-hosted tests utilized an NVIDIA A100-SXM4-80GB GPU on an a2-ultragpu-1g GCP VM instance. These self-hosted costs primarily illustrate scale, as they can vary by cloud provider and region, or reduce to electricity costs alone when using owned hardware.

Energy estimates accompany the cost analysis where relevant. For provider-hosted models, these represent informed estimates based on academic research, as providers rarely publish detailed energy consumption data (see Appendix A.6 for methodology). For self-hosted models, energy consumption is calculated directly from test duration and GPU power consumption.

Our self-hosted model setup prioritizes throughput through several optimizations. All self-hosted models use 4-bit quantized variants paired with the unsloth library for faster inference and reduced VRAM usage. We additionally leverage the vLLM library, which utilizes spare VRAM for accelerated token processing, complementing Unsloth effectively. [Kwon W. et al. (2023), Daniel Han, Michael Han and Unsloth team (2023)]

Not all self-hosted models support both unsloth and vLLM at time of writing, notably Gemma 3 and Mistral 3.2. For these models, we used only vLLM, which likely contributed to their higher cost and energy usage.

We deliberately exclude general legal benchmarks like LegalBench from our evaluation framework. LegalBench tests broad legal reasoning capabilities (e.g., hearsay detection, contract interpretation) that differ fundamentally from regulatory classification tasks. Our focus on structured policy and regulation text demands precision in taxonomy mapping rather than interpretive flexibility. This specialized domain requires benchmarks that reflect the hierarchical, multi-class nature of regulatory frameworks rather than binary legal reasoning tasks.

1.5.1 Model Selection Rationale

This study compares nearly all current state-of-the-art provider-hosted models (Claude 4.1 Opus, Claude 4 Sonnet, Gemini 2.5 Pro, Gemini 2.5 Flash, GPT-5, Grok-4), chosen for their strong performance on general benchmarks and widespread industry adoption.

For open-source baselines, we selected Mistral 3 Instruct (24B parameters), Qwen 3 (32B, with and without thinking mode enabled), Gemma 3 (27B), and Mistral 3.2 Instruct (24B). These smaller models were intentionally chosen to demonstrate that specialization can overcome raw model size disadvantages. This comparison is deliberate, if fine-tuning can elevate an older 24B parameter model (Mistral 3 Instruct) to outperform models with an estimated 10-50x more parameters, it validates specialization as a more efficient path than simply scaling model size. The selection also ensures reproducibility, as Mistral's open weights allow replication of our baseline, unlike provider-hosted models where version changes can invalidate comparisons.

Notable exclusions include DeepSeek, which was omitted because it is too large to be considered an SLM yet doesn't fit the provider-hosted model category despite offering API access, given its open-source weights. Similarly, Llama and Phi variants were excluded as their size configurations fall either much smaller or much larger than the models being tested in each category.

1.5.2 Evaluation Dataset Information

The evaluation dataset characteristics are detailed in the appendix, which includes examples of test data (A.3), taxonomy structure (A.5), and prompt format (A.4). The AML test dataset comprises 713 annotated pages of regulatory documents relating to AML from various jurisdictions worldwide. Each document receives a unique identifier upon addition to RegGenome's corpus, ensuring no overlap between training and test datasets.

These 713 pages contain 1,024 L1 tag assignments and 1,288 L2 tag assignments. This represents a multi-label, multi-class classification task where each page can receive multiple taxonomy labels. Models are instructed to generate predictions only at Level 2; the corresponding L1 assignment is then inferred. For example, tagging aml-cdd-enhanced automatically implies aml-cdd.

Input prompts typically contained approximately 7,000–9,000 tokens, with required completions ranging from 20–100 tokens, though reasoning-mode outputs occasionally produced longer responses.

We maintained consistency by using the same prompt structure (detailed in Appendix A.4) across all provider-hosted models, self-hosted open-source models, and specialized models. The only modifications occurred for open-source models that required separation of instruction and system prompts with special tokens. In these cases, as with proprietary models, the system prompt remained blank while the entire prompt was placed in the instruction or user input section.

2 Quantifying Performance Variance in Provider-hosted LLMs

Section Summary

Leading API models produce different answers to identical questions up to 100 times as often as self-hosted SLMs, creating unacceptable legal risk.

A key aspect of a trustworthy AI system is predictability. For a tool to be adopted by legal professionals, its behaviour must be as deterministic as possible. A model that correctly identifies a regulatory obligation one day but misses it the next is a liability. Even a model that is consistently incorrect has advantages, as its errors can be anticipated and managed. In contrast, a model that varies significantly has little value in this domain. This section establishes a framework for measuring this performance variance and applies it to leading LLMs, revealing a significant stability issue that grows with task complexity.

2.1 Framework for Measuring Instability

To move beyond anecdotal observations of model variability, we employ two distinct metrics to quantify different aspects of performance instability between model runs. These metrics operate on the F1-score vectors $\mathbf{u}$ and $\mathbf{v}$, where each vector represents the per-class F1-scores from a single evaluation run over a taxonomy with n classes.

2.1.1 Metric Definitions

The first metric, **Mean Absolute Difference (MAD)**, measures the average magnitude of performance fluctuation per class. It provides a direct and intuitive measure of raw performance variance. A MAD of 0.02, for example, indicates that on average, any given class's F1-score differs by 2 percentage points between runs. It is formally defined as:

$$\text{MAD}(\mathbf{u}, \mathbf{v}) = \frac{1}{n} \sum_{i=1}^{n} |u_i - v_i| \tag{1}$$

The second metric, **Cosine Distance**, measures the difference in the *pattern* of performance, independent of the overall magnitude. It is defined as 1 minus the cosine similarity of the two performance vectors. A distance near 0 indicates that the model's relative strengths and weaknesses are consistent across runs (i.e., it is good at the same classes and bad at the same classes). A larger distance reveals that the model's performance profile is shifting unpredictably. It is formally defined as:

$$\text{Cosine Distance}(\mathbf{u}, \mathbf{v}) = 1 - \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \sqrt{\sum_{i=1}^{n} v_i^2}} \tag{2}$$

The selection of these metrics aligns with a growing body of research focused on developing robust measures for LLM stability, such as the Total Agreement Rate (TAR@N), which assesses the percentage of identical answers across multiple runs. Our approach provides a more granular, per-class view of this instability. [Atil, B., et al. (2025)]

2.1.2 The Insight of Cosine Distance

It is important to note that while MAD and Cosine Distance seemingly serve the same purpose, they actually produce complementary, non-overlapping information. A model could have a non-zero MAD but a near-zero Cosine Distance. This would signify a "consistent inconsistency": the model's overall performance might fluctuate, but its relative performance across classes remains stable. Such a model, while not ideal, has predictable failure modes that can potentially be accounted for. Conversely, a high Cosine Distance indicates *erratic unreliability*. This is a

far less manageable problem, as it means not only is the model's performance inconsistent, but the very nature of its inconsistency is random. One cannot reliably predict which types of document it will handle well from one day to the next. This erratic behaviour suggests a fundamental lack of robust, generalisable understanding of the task domain. The model may be relying on superficial, stochastic correlations in the data rather than a deep, structured representation of the regulatory concepts it is meant to classify. The problem of inconsistency can be addressed in the short term by self-hosting and seeding the model's sampling methods as employed in this study. However, the long-term solution could arguably be to produce a model which actually *understood* the text consistency would then stem from correctness rather than enforced determinism via seeding. Regardless, in high-stakes environments such as law, even small degrees of variance are undesirable

2.2 Empirical Analysis of Provider-hosted Model Variance

We subjected several prominent provider-hosted LLMs, GPT-5, Claude Opus 4.1, Claude Sonnet 4, Grok-4, Gemini 2.5 Pro and Gemini 2.5 Flash, to five identical evaluation runs on the test set, with all deterministic parameters (e.g., temperature set to 0, top_p set to 0, top_k set to 1) and a seed where the provider's API made this possible. Each of the five runs employed identical prompts and hyperparameters. Where we report variance measurements of <0.001, this indicates variance below our measurement threshold rather than absolute determinism. Theoretical limits of floating-point precision and hardware variation make true zero variance unlikely; however, any residual variance here is negligible in practice. The results, summarised in Table 1 and Figure 1, reveal a concerning level of performance instability.

Table 1: Example of Prediction Variance Analysis of API-Based Models (Averaged over 5 runs)

Model	Taxonomy Level	Avg. MAD	Avg. Cosine Distance	Interpretation
Gemini 2.5 Pro	Level 1	0.0284	0.0046	Relatively stable magnitude, pattern
Gemini 2.5 Pro	Level 2	0.0248	**0.0514**	Stable magnitude, erratic pattern
Gemini 2.5 Flash	Level 1	0.0463	0.0073	Unstable magnitude, stable pattern
Gemini 2.5 Flash	Level 2	0.0210	0.0429	Stable magnitude, erratic pattern
Claude 4 Sonnet	Level 1	0.0130	0.0056	Relatively stable magnitude, pattern
Claude 4 Sonnet	Level 2	0.0005	0.0002	Very stable magnitude, pattern
Claude 4.1 Opus	Level 1	0.0297	0.0047	Relatively stable magnitude, stable pattern
Claude 4.1 Opus	Level 2	0.0419	**0.0856**	Unstable magnitude, very erratic pattern
Grok-4	Level 1	**0.0752**	0.0141	Very unstable magnitude, relatively stable pattern
Grok-4	Level 2	0.0492	**0.0878**	Unstable magnitude, very erratic pattern
GPT-5	Level 1	**0.0852**	0.0506	Very unstable magnitude and pattern
GPT-5	Level 2	0.0635	**0.1698**	Unstable magnitude, **extremely erratic pattern**

2.2.1 Analysis of Results

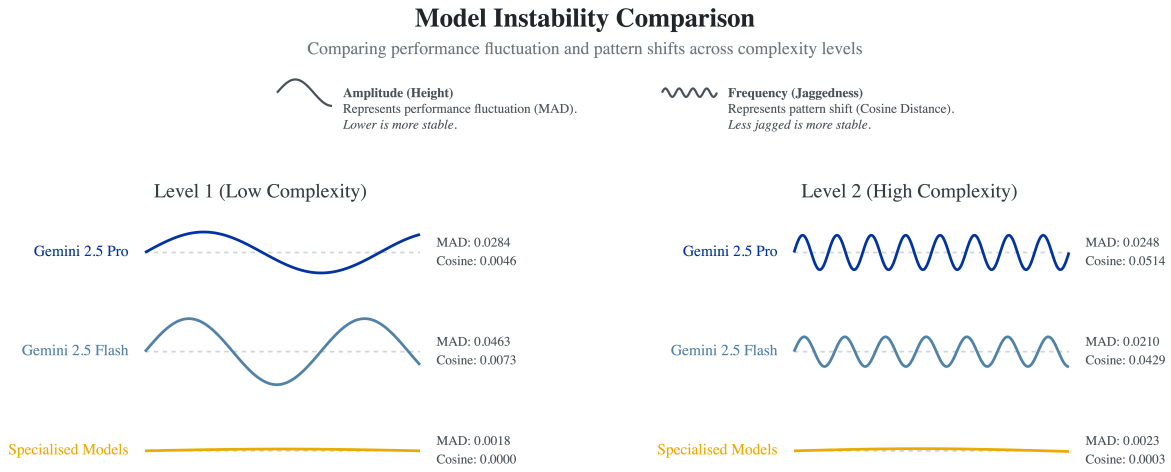
The data in Table 1 and Figure 1 show a clear difference across provider-hosted LLMs. For Gemini 2.5 Pro, while the average magnitude of error (MAD) is relatively low and stable across both taxonomy levels, the Cosine Distance experiences a more than tenfold increase from Level 1 to Level 2. This indicates that as the classification task becomes more granular and complex, moving from 19 broad categories to 164 specific ones the model's performance pattern becomes highly erratic. Its ability to consistently identify its own strengths and weaknesses breaks down precisely when the task demands more nuanced understanding.

Gemini 2.5 Flash exhibits a different but equally concerning profile. At Level 1, it suffers from a high MAD of 0.0463, meaning its F1-score for any given broad category fluctuates by an average of 4.6 percentage points between runs. While its pattern is more stable at this level (low Cosine Distance), the raw performance is unreliable. At Level 2, its behaviour mirrors that of the Pro model, with the Cosine Distance increasing significantly, signalling an unreliable performance pattern at the granular level.

One pattern that seems to emerge from these results is that the larger variants of the models seem to be noticeably more unstable. This is particularly true when comparing Claude Opus and Sonnet. Here in particular we see the smaller version exhibiting vastly less variance than the larger. Several factors could contribute. Larger models may require more distributed compute across heterogeneous hardware, increasing numerical variance. Alternatively, these larger models may invoke more "reasoning" functionality. This aligns with observations of Grok-4 and GPT-5 generating more reasoning tokens during this study, both also exhibit the highest instability.

A reliable teammate performs their role consistently. The erratic unreliability shown by the generalist models (high Cosine Distance) is akin to a teammate whose skills are unpredictable from one day to the next, making it impossible to build a trusted workflow.

Figure 1: A conceptual waveform visualisation of model instability.



2.2.2 The Deterministic Advantage of Self-Hosting

In stark contrast to the API-based models, and with the exception of Claude 4 Sonnet, all self-hosted, non-reasoning SLMs tested in this study, exhibited a variance of <0.001 for all runs. By controlling the entire software and hardware stack and setting a deterministic seed, near-perfect reproducibility can be achieved. This is not a minor technical detail but a fundamental architectural advantage. The instability observed in the API models is a direct consequence of their distributed, multi-tenant nature, where sources of non-determinism such as floating-point arithmetic differences across hardware, minor software updates, or even quantisation artifacts can introduce variability that is outside the user's control. For applications where trust and reliability are paramount, the ability to ensure near determinism is an advantage that only a self-hosted solution can provide. [Szalontai, B., et al. (2025), Andreas H., et al. (2025)]

3 A Comparative Analysis of Model Performance for Regulatory Document Classification

Section Summary

The `RegGenome` model delivers 38% better accuracy at 1/80th the cost than leading API models (Gemini 2.5 pro/flash). This translates to savings of up to $540,000 per taxonomy on the RegGenome corpus while reducing human review effort by up to 40%.

While stability is a prerequisite for trust, high performance is essential for utility. This section provides a comparison of the models, evaluating them not only on classification accuracy but also on the critical business dimensions of cost and speed. The results clearly demonstrate that the `RegGenome` model is superior across every significant metric.

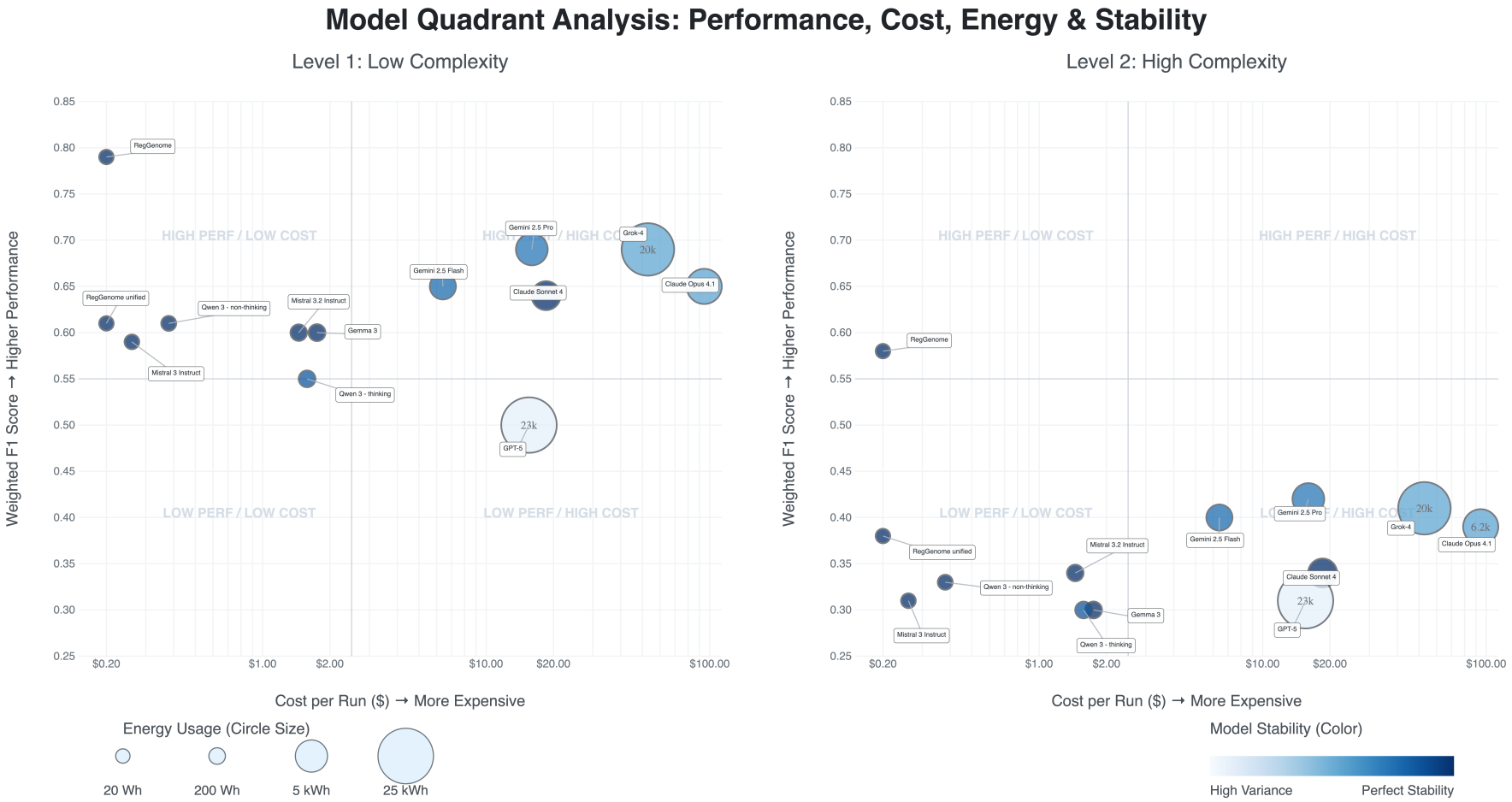
3.1 Performance and Cost-Efficiency Benchmarks

Table 2 and subsequent Figure 2 provide a view of model performance, juxtaposing accuracy with the practical considerations of inference time and cost, alongside the critical stability metrics discussed in the previous section.

Table 2: Model Performance and Cost-Efficiency Comparison on the RegGenome AML taxonomy.

Model	L1 Weighted F1	L2 Weighted F1	Energy Usage per run (Wh)	Cost per Run	Prediction Variance (Avg. L2 Cosine Distance)
Specialised Self-Hosted SLMs					
RegGenome	0.79	0.58	20.61	$0.20	<0.001
Provider-hosted LLMs					
Gemini 2.5 Pro	0.69	0.42	4456.25	$16.00	0.051
Grok-4	0.69	0.41	19964.00	$52.89	0.088
Gemini 2.5 Flash	0.65	0.40	2139.00	$6.40	0.043
Claude Opus 4.1	0.65	0.39	6238.75	$94.45	0.086
Claude Sonnet 4	0.64	0.34	3208.50	$18.53	<0.001
GPT-5	0.50	0.31	23172.50	$15.54	0.170
Self-hosted SLMs					
RegGenome unified	0.61	0.38	21.39	$0.20	<0.001
Mistral 3.2 Instruct	0.60	0.34	141.89	$1.45	<0.001
Qwen 3 - non-thinking	0.61	0.33	38.50	$0.38	<0.001
Mistral 3 Instruct	0.59	0.31	27.81	$0.26	<0.001
Gemma 3	0.60	0.30	175.40	$1.75	<0.001
Qwen 3 - thinking	0.55	0.30	158.29	$1.58	0.028

Figure 2: A quadrant analysis of model performance on the RegGenome AML taxonomy.



Note: Energy values (in kWh) are shown inside larger bubbles.

3.1.1 Analysis of Accuracy

The results presented in Table 2 and Figure 2 show a clear winner. The `RegGenome` model, RegGenome's specialised SLM, achieves a Weighted F1-Score of 0.58 on the complex L2 classification task. This is a 16 percentage point increase in F1-score, representing a 38% relative performance gain over the best API performer, Gemini 2.5 Pro (0.42). This is not an incremental gain, in a human-in-the-loop system, this 38% relative performance improvement directly translates to a significant reduction in the number of errors a human analyst must find and correct.

Furthermore, the performance of `RegGenome` (0.58) compared to its base model `Mistral 3 Instruct` (0.31) demonstrates the significant impact of fine-tuning. The base model, while a capable open-source foundation, lacks the domain-specific knowledge to perform this task effectively. The fine-tuning process elevates its performance by 27 percentage points (an 87% relative gain), confirming that specialisation is the key to unlocking high performance in this domain.

To illustrate the practical impact, this 16 percentage point improvement in F1-score signifies a substantial reduction in the combination of false positives and false negatives. In a human-in-the-loop system, this translates directly into a more manageable review queue for a human analyst, who would need to correct significantly fewer misclassifications. For a batch of 1,000 provisions, this could easily translate to a direct saving of several hours of expert work.

3.1.2 Analysis of Efficiency (Time & Cost)

The efficiency gains offered by the self-hosted SLM approach are even more pronounced. The `RegGenome` model completed the evaluation run in just over 3 minutes at a cost of $0.20. In contrast, the Gemini 2.5 Pro model cost $16.00. This makes the `RegGenome` model **1/80th the cost** on a per-run basis. These efficiency gains are table stakes for a viable HITL workflow. A large processing time for a single batch is operationally unfeasible for review cycles, whereas a 3-minute turnaround enables a fast, iterative workflow between the analyst and the model.

The 3-minute inference time scales linearly with document volume, enabling processing of approximately 250,000 pages per day on a single GPU instance. For organisations processing thousands of regulatory updates daily, horizontal scaling across multiple GPUs provides near-unlimited throughput. At peak load, a 4-GPU configuration could process around 1,000,000 pages daily at a fraction of the cost of API-based solutions, while maintaining consistent sub-5-minute latency for urgent classifications.

These figures represent only the direct operational costs. A complete Total Cost of Ownership (TCO) analysis would further favor the self-hosted approach by factoring in the significant "risk cost" associated with the provider-hosted models' instability and the strategic costs of relinquishing data control to a third-party API provider. For any application with meaningful volume, the economic case for self-hosting a specialised model is clear.

3.1.3 The Distinction Between Legal and Regulatory Language Processing

While our approach demonstrates strong performance on regulatory classification, it is important to distinguish our focus from broader "Legal AI" benchmarks. Services like VALS AI provide proprietary benchmarks for evaluating legal language models across tasks including contract analysis and case law interpretation. However, these benchmarks present several limitations for evaluating regulatory classification systems:

First, VALS employs GPT-4 as an evaluator for their benchmarks, a methodology that our stability analysis (Section 2) demonstrates is fundamentally unreliable, with evaluation variance potentially exceeding the performance differences being measured. Second, legal benchmarks focus primarily on interpretive legal tasks rather than the structured, taxonomic classification

required for regulatory compliance. Third, the benchmarks are inconsistently applied (e.g., models are evaluated on different domain subsets), which undermines reliable comparison.

The distinction is crucial: regulatory and policy language operates as a specialised dialect making performance on general "legal" benchmarks very difficult to rely on. This structural regularity is part of what enables RegGenome's approach to achieve such significant performance gains, while also explaining why these same techniques readily transfer to other highly-regulated domains (healthcare, environmental, HR compliance) that share similar linguistic properties, yet are not commonly assessed in benchmarks with a regulatory lens. [vals.ai]

3.2 Effect of Regulatory Turbulence

It is important to remember while investigating the abilities of these models that regulation is not a constant monolith, it fluctuates over time with new areas being folded in and older areas being altered and modernised. With this in mind we decided to investigate a second more modern regulatory theme, in this case cryptocurrency. The hypothesis going into this evaluation was that part of the reason the generalist models were able to classify to a reasonable, if not top of the line, standard was because they had a good degree of prior exposure to this sort of task. The RegGenome taxonomies are built with significant influence from existing accepted standards. In the case of AML which has been mostly standardised for decades this means that any LLM trained on wide swathes of the internet would have a good degree of exposure to these standards and potentially even some degree of classification experience against them. For more modern areas like cryptocurrencies for example, such standards are not readily available and so we would expect the relative performance gains on the specialised model to even more noticeably outstrip the generalist.

Table 3: Model Performance and Stability Comparison on the RegGenome Crypto taxonomy.

Model	L1 Weighted F1	L2 Weighted F1	Prediction Variance (Avg. L2 Cosine Distance)
Specialised Self-Hosted SLMs			
RegGenome	0.71	0.50	<0.001
Provider-hosted LLMs			
Grok-4	0.56	0.31	0.086
Gemini 2.5 Pro	0.57	0.29	0.117
Claude Opus 4.1	0.54	0.30	0.054
Claude Sonnet 4	0.55	0.28	0.023
Gemini 2.5 Flash	0.54	0.27	0.057
GPT-5	0.49	0.27	0.170
Self-hosted SLMs			
Qwen 3 - non-thinking	0.52	0.30	<0.001
Mistral 3 Instruct	0.52	0.30	<0.001
Gemma 3	0.51	0.28	<0.001
Mistral 3.2 Instruct	0.47	0.28	<0.001
Qwen 3 - thinking	0.44	0.20	0.022

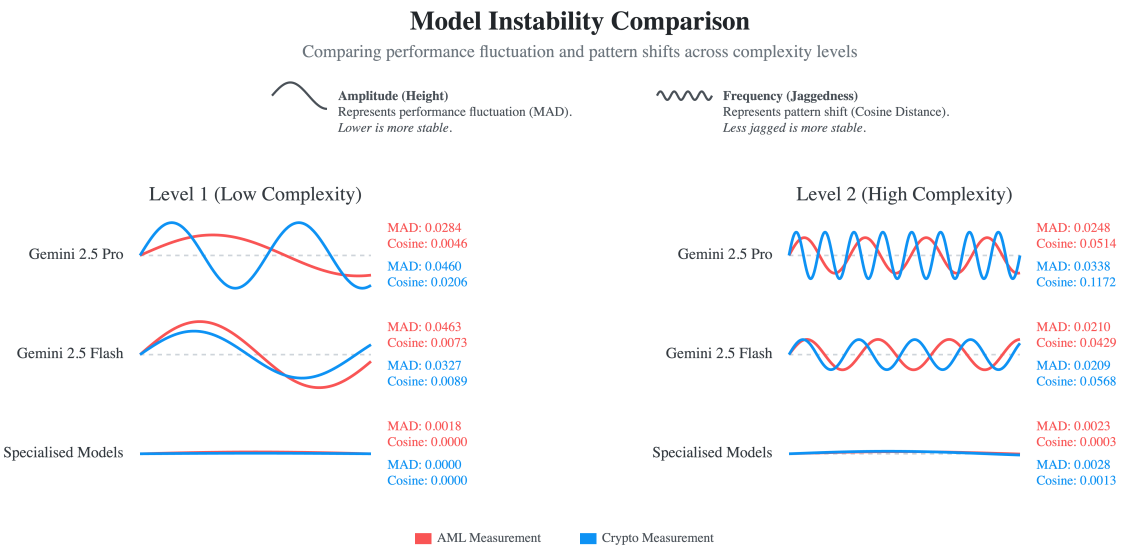
Efficiency metrics for the Crypto task were directionally consistent with the AML evaluation and are omitted for brevity.

3.2.1 Impact of Turbulence on Accuracy and Stability

As we can see in Table 3 these accuracy results reveal a different performance dynamic compared to the AML task. On the level 2 task we see Mistral 3 Instruct model actually outperforming both Gemini 2.5 pro and flash and Grok-4 takes over as the top-performing provider-hosted model. This indicates that while almost every model benchmark puts provider-hosted models well above the Mistral 3 and other open-source SLMs, they struggle immensely to generalise to newer areas of regulation. We also of course see that the `RegGenome` model outperforms all of these models by an even larger margin than in the previous assessment. This represents a

14-percentage-point (24.6% relative) increase at L1 and a 21-percentage-point (72.4% relative) increase at L2.

Figure 3: Conceptual waveform visualisation of an instability comparison for AML vs Crypto.



On the stability front as illustrated in Figure 3 we also see quite a significant change. Both Gemini 2.5 Pro and Flash experience increased erraticism as shown in their respective cosine scores when compared to their AML performance. However, what is most interesting is that the Pro model feels these impacts far more acutely than the Flash version with its instability more than doubling across both L1 and L2 in the crypto assessment. Additionally we see Claude 4 Sonnet which was previously largely equivalent to the self-hosted models' stability in the AML test become much more unstable here.

As in the AML test, one might speculate that instability is partly attributable to chain-of-thought implementations in larger provider-hosted models and in Qwen 3 (Thinking). Chain of thought in LLMs functions as an expanded context window for the model, it has a space to output relevant context tokens which might help it home in on the correct answer already present within its vast training data. However, because this task concerns an emergent area of regulation, it could be that the LLM has no relevant references in its training data and so the chain of thought serves as little more than as a maze of branching paths for the model leading to a sizeable increase in prediction instability.

Overall we can see that while the `RegGenome` model is the best performer in terms of both accuracy and stability across the board it is where regulation is at its most unstable that this approach is at its most essential.

3.3 The Power of Specialisation

The superior performance of `RegGenome` is a direct consequence of specialisation. Generalist models are trained on a vast and diverse corpus of internet text, which provides them with broad knowledge but leaves them ill-equipped to handle the unique linguistic and semantic structures of a specialised domain like law. Policy and regulatory language is not merely a different topic, it functions as a distinct dialect with its own vocabulary, syntax, and logical constructs. [Ho, D. E. (2024), NLS Business Law Review. (2024)]

By fine-tuning a smaller model on a high-quality, curated dataset of regulatory documents tagged by domain experts, the model learns the specific patterns, terminology, and conceptual relationships that are essential for accurate classification. This targeted approach is demonstrably more

effective than the brute-force method of simply increasing the size and breadth of a general model. Research has consistently shown that smaller, fine-tuned models can match or even exceed the performance of models many times larger on specific enterprise tasks. The performance gap observed in our results is so significant that it suggests a *performance cliff* for generalist models in this domain. Fine-tuning is not merely an optimisation, it is a transformative process to achieve professional-grade competence. [Fatemi, S., et al. (2024)]

4 Adapting to Regulatory Turbulence: Zero-Shot Generalisation for Dynamic Regulatory Taxonomies

Section Summary

New regulations arrive daily. The `RegGenome unified` model provides informed "first-pass" classification on brand-new regulations without months of training. This transforms regulatory response time from months to days, enabling rapid market expansion.

While supervised fine-tuning delivers market-leading performance on established tasks, its reliance on pre-labelled data presents a critical bottleneck in the legal and financial sectors. These domains are not static, they exist in a state of constant regulatory turbulence. Industry analyses consistently report that financial services face over 60,000 regulatory updates annually across 1,300+ regulatory bodies, generating millions of pages of new compliance requirements each year in various formats. Each significant regulatory shift can necessitate the creation of a new taxonomy or the modification of an existing one. Manually creating labelled data to train a new model for each change is an operational bottleneck.

An area with such high velocity and high volume necessitates an AI that can adapt to novel classification schemes before lengthy data annotation cycle is completed. The solution is zero-shot generalisation: the ability to classify content against a new taxonomy for which the model has no specific training examples. This section details our initial experiments to build and validate this crucial capability. [Sarkar, R., et al. (2021), Shambhavi, M., et al. (2024)]

4.1 `RegGenome unified`: Creating a Generalised Specialist

To overcome the "cold-start" problem of new regulations, we developed a "meta-learner" model, `RegGenome unified`. This model was trained not on a single taxonomy, but on the entirety of RegGenome's diverse, labelled data assets, which span multiple distinct regulatory frameworks.

The hypothesis was that by exposing the model to a wide variety of classification schemes, it would learn more than just the labels of a single task. It would begin to learn the underlying structural patterns, semantic relationships, and logic inherent to regulatory text itself. The goal was to create a model that could apply this generalised understanding to a completely unseen taxonomy, effectively learning "how to classify" as a transferable skill. This is a real-world test of zero-shot classification, a powerful capability that allows models to perform tasks without any task-specific labeled data. [Savelka, J., & Ashley, K. (2023)]

4.2 First-Pass Performance on Unseen Taxonomies

The results of this experiment, shown in Table 2, are highly encouraging. When evaluated on the unseen L2 taxonomy, the `RegGenome unified` model achieved a Weighted F1-Score of 0.38. While this does not match the 0.58 F1-score of the specialist `RegGenome` model fine-tuned for that specific task, it significantly outperforms the raw open-source baseline, `Mistral 3 Instruct` (0.31).

This 7 percentage point lift proves that the model successfully transferred knowledge from its diverse training to the new task. An F1-score of 0.38 provides an immediate, intelligent "first-pass" annotation. In the context of responding to one of the 200+ daily regulatory alerts, this capability transforms a manual, from scratch review process into a far more efficient human-in-the-loop workflow, providing a massive head start for analysts.

4.3 The Strategic Value of Zero-Shot Agility

The ability to achieve better than baseline performance on new taxonomies without training is not just a technical achievement, it is a core business advantage in a turbulent regulatory environment.

- **Agility in the Face of Regulatory Change:** When a new regulation is issued and a taxonomy modified or introduced, a zero-shot model allows for the immediate triaging and classification of relevant documents. This near-instant response capability is a powerful differentiator, enabling teams to understand and react to new compliance obligations in days, not months.
- **Overcoming the Annotation Bottleneck:** The primary barrier to deploying AI for new tasks is the cost and time of expert-led data annotation. Zero-shot capability drastically reduces this dependency. Instead of labeling from a blank slate, human experts can refine the model's first-pass output, creating a powerful, scalable system for developing new classification schemes with unprecedented efficiency. [Savelka, J., & Ashley, K. (2023)]
- **Scalable Market Expansion:** This capability allows for rapid and cost-effective expansion into new regulatory domains. The zero-shot model provides an instant baseline of understanding, de-risking entry into new jurisdictions and allowing the business to scale its intelligence offerings much more quickly.

4.4 Alternative Approaches and Future Work

While some practitioners might consider using LLM-generated synthetic training data as shortcuts to avoid manual annotation, these approaches face fundamental limitations for classification tasks. Using LLMs to generate training data introduces the risk of propagating the generalist model's errors and biases into the specialised system. Validating these hypotheses empirically represents an opportunity for future research as it is out of scope for the current work. Preliminary analysis suggests that human-expert annotation remains the gold standard for creating high-quality training data in regulatory domains.

5 Case Study: The Compounding Cost of Errors in a Multi-Agent Workflow

Section Summary

In this example workflow with multiple AI steps, RegGenome's approach processes 64% more documents successfully than generalist models. Small accuracy gains compound into massive efficiency improvements at scale.

To move from theoretical benchmarks to practical application, this section examines the impact of model choice on a realistic, multi-step workflow. In modern process automation, tasks are often broken down into a series of dependent steps, each handled by an agent. In such a system, the final output is only correct if *every single step* in the chain succeeds. The overall reliability of the workflow is the product of the reliability of its individual components, meaning that even small inaccuracies at each stage compound into a significant overall failure rate.

5.1 A Three-Step Automation Scenario

Consider a common document processing task for a regulatory / compliance team, broken into a three-agent workflow:

1. **Agent 1 (Identify):** The AI reads a document and identifies pages containing high-level regulatory subject matter. This step uses the L1 classification model to determine if a page is relevant.
2. **Agent 2 (Extract):** If a page is identified as relevant, a second AI agent extracts the specific sentence or clause containing the core obligation. For this, we assume the use of a high-performing generalist model is suitable for text extraction tasks.
3. **Agent 3 (Classify):** Finally, the extracted text is classified against a granular taxonomy using the L2 classification model.

For a document to be processed correctly through this pipeline, it must pass all three stages successfully. A failure at any point, for example, a missed identification in Step 1, a faulty extraction in Step 2, or a misclassification in Step 3 renders the entire process for that document a failure from the perspective of the end-user.

Figure 4 visualises the outcome of this workflow when processing a batch of documents, comparing a pipeline built on the best-performing generalist LLM (Gemini 2.5 Pro) with one using the specialist `RegGenome` SLM. The diagram clearly illustrates the effect of compounding failure. Despite the generalist models performing adequately at each individual stage, the cumulative effect of their lower accuracy results in far fewer correctly processed documents. The specialised pipeline, with its higher accuracy at the critical classification stages, successfully processes 18 more documents, representing a **64% relative increase in end-to-end throughput** at **1/80th the cost**. This improvement validates recent proposals that specialised SLMs represent the future of agentic AI, particularly when agents perform narrow, repetitive tasks rather than requiring general conversational abilities. [Belek, P., et al. (2025)]

Furthermore, we see the benefits increase even more when we move away from well understood regulatory domains like AML into more emergent areas, in this case Crypto. Comparing the best-performing generalist LLM (Grok-4) and the `RegGenome` SLM, we see a **112% relative performance gain** more than doubling the number of correctly processed documents. This is doubly important from the perspective of business costs, as emergent areas are unlikely to be able to be dealt with by in-house capability in the first instance and therefore necessitate the hiring of expensive outside expertise. While the introduction of the `RegGenome` SLM does not completely negate this cost we will investigate in the next section what tangible effect having more correctly annotated documents has on the man-hours required for experts to review.

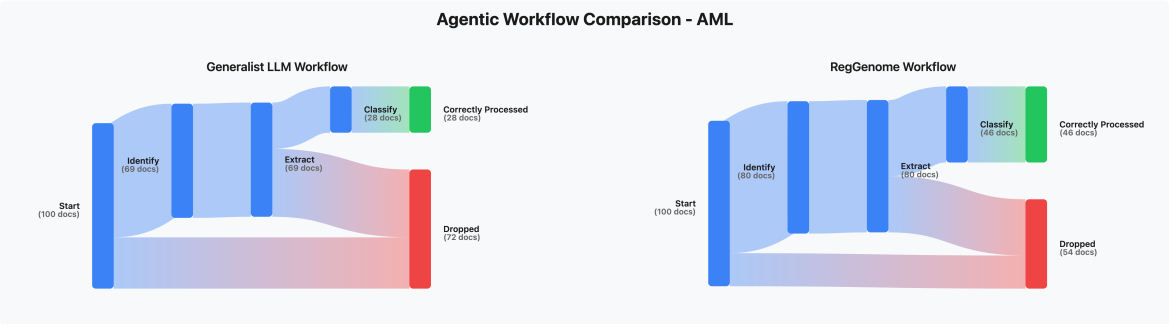


Figure 4: A Sankey diagram visualising the compounding inaccuracy in two multi-step workflows for the AML taxonomy. The RegGenome workflow successfully processes 18 more documents out of 100 compared to the Generalist LLM. This represents a 64% relative increase in performance vs the generalist.

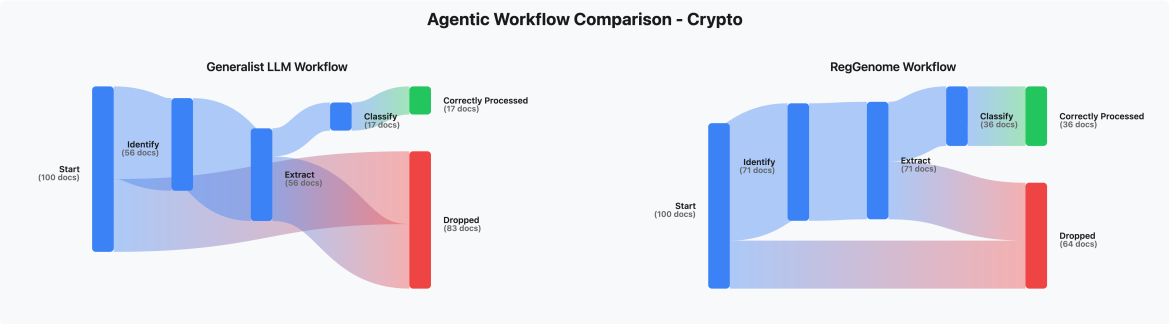


Figure 5: A Sankey diagram visualising the compounding inaccuracy in two multi-step workflows for the Crypto taxonomy. The RegGenome workflow successfully processes 19 more documents out of 100 compared to the Generalist LLM. This represents a 112% relative increase in performance vs the generalist.

5.2 Translating Throughput to Human Effort

The true value of this increased throughput becomes clear when we analyze the impact on the human-in-the-loop. A higher success rate from the AI directly reduces the manual effort required by analysts. We consulted the RegGenome annotation team to estimate the time required to complete the AML test set annotation under three conditions: from scratch, reviewing the generalist model's output, and reviewing the specialist model's output.

The results, summarised in Table 4, clearly demonstrate the effectiveness of this approach. Although the generalist model provides a marginal benefit over fully manual annotation, its long processing time and lower accuracy mean the overall time saving is minimal. In contrast, the `RegGenome` SLM creates a step-change in efficiency. Not only reducing the required human review time by ∼40% compared to manual annotation, but its near-instantaneous processing time makes the entire workflow faster and more responsive.

Table 4: Human-in-the-Loop Efficiency: A Comparison of Total Workflow Time

Workflow Scenario	AI Processing Time	Human Review Time	Total Workflow Time	Relative Time Saving (vs. Manual)
Manual Annotation (from scratch)	N/A	30 - 36 hrs	30 - 36 hrs	-
HITL with Generalist (Gemini 2.5 Pro)	∼5 hrs	24 - 27 hrs	29 - 32 hrs	∼10%
HITL with Specialist (RegGenome)	**∼5 mins**	**18 - 21 hrs**	**18 - 21 hrs**	**∼40%**

This analysis proves that for a HITL system to be truly effective, it must be evaluated on total end-to-end efficiency. The **RegGenome** model is not just more accurate or cheaper in isolation,

it enables a fundamentally more efficient and collaborative relationship between the analyst and the AI, saving days of expert-level work and transforming the viability of AI-assisted review at scale.

6 Discussion and Strategic Implications

Section Summary

Building specialised AI in house creates three defensive moats: data sovereignty, predictable costs, and proprietary capabilities competitors can't replicate. This aligns with 96% of legal professionals who demand AI as a controlled assistant, not an autonomous replacement.

The empirical results presented in this report carry significant strategic implications for RegGenome and the broader technology industry. They provide a clear, data-driven validation for a strategy centered on specialisation, in-house development, and data asset cultivation. The choice between using a provider-hosted model and building a specialised one is not just an economic decision about cost, it is a fundamental choice about a company's long-term competitive posture.

6.1 The Three Key Advantages

For executive leadership, the findings of this study can be simplified into three proven competitive advantages that result from the specialised approach:

1. **Superior Performance:** Fine-tuned models are demonstrably more accurate, which directly reduces the time and cost of human review and correction, leading to a higher quality final product.
2. **Unmatched Reliability:** The self-hosted models are vastly more stable and reproducible. This predictability is essential for human users, who can trust the AI's output and build efficient, repeatable workflows around it, eliminating the risk of an unreliable AI partner. The instability makes the generalist model a poor teammate, requiring constant supervision and creating more work, not less.
3. **Scalable, Cost-Effective Collaboration:** RegGenome's approach is orders of magnitude cheaper and faster, making high-volume, human-in-the-loop collaboration economically feasible. This sustainable cost structure enables us to deploy human-AI teams at a scale that is impossible with expensive APIs.

This combination of higher quality, higher reliability, and lower cost represents a powerful and defensible market position.

6.2 The Strategic Necessity for Self-Hosting and Specialisation

The decision to develop models in-house is a strategic one that builds a deep, defensible moat around the business. This strategy is again demonstrated by three key pillars:

6.2.1 The Total Cost of Ownership (TCO) Argument

The per-run cost comparison in Table 2 only scratches the surface. A full TCO analysis reveals an even more compelling case for self-hosting. While there are upfront and ongoing costs associated with hardware, maintenance, and expert personnel, these costs are predictable and manageable. In contrast, API-based costs are variable and can escalate uncontrollably with usage, in fact to process RegGenome's whole corpus of documents for even a single taxonomy with Gemini 2.5 Pro would cost around **$540,000**, for results that are provably less accurate and vastly less consistent. For any application with high-volume, mission-critical usage, the break-even point where self-hosting becomes more economical is reached quickly.

6.2.2 The Data Privacy and Security Moat

In the legal and financial sectors, data confidentiality is a foundational requirement. Entrusting sensitive client or proprietary regulatory data to a third-party API creates an inherent and unavoidable security risk. It exposes the organisation to the provider's security practices, potential data breaches, and policies regarding data usage for model training. Self-hosting on private, controlled infrastructure maximizes control over data residency and processing. This commitment to security is a powerful differentiator and an important consideration for operating in high-stakes domains. [NLS Business Law Review. (2024)]

6.2.3 The Persistence of Specialisation Advantage

Many are understandably concerned with wasted effort when opting for a custom solution. With all of the continued hype around constantly improving models, many are concerned about the "leap-frog" risk, the idea that in the time it takes to create your own solution the models will have improved to such a degree that it renders the effort unnecessary. However, future models, despite continued improvements, cannot eliminate the need for specialised solutions due to fundamental architectural constraints:

1. **Inherent Stochasticity:** Large models' probabilistic nature introduces unavoidable randomness, even with hyperparameters set to be as deterministic as possible.
2. **Opaque Updates:** Provider-hosted models undergo unpublished modifications that can silently alter behavior.
3. **Environmental Variance:** Hardware differences, quantisation, and distributed processing create reproducibility challenges.
4. **Benchmark Overfitting:** Recent research shows diminishing real-world gains despite benchmark improvements. [Xu, C., et al. (2024)]
5. **Shrinking moat:** The performance gap between open-source SLMs and provider-hosted LLMs models continues to narrow, allowing specialised approaches to leverage improving foundations while maintaining their domain specific advantages.

These limitations suggest that specialisation will remain the optimal path for mission-critical applications.

6.2.4 Control, Customisation, and Long-Term Innovation

Choosing to self-host is a choice to become a producer of AI, not merely a consumer. This provides complete control over the technology stack, from the hardware to the model architecture. This control enables deep customisation and the development of unique, proprietary capabilities, such as `RegGenome unified`, that are impossible to create within the confines of a closed, third-party ecosystem. Having complete control allows the company to respond to market needs and to build a core intellectual property asset that grows more valuable over time.

6.2.5 Cross-Domain Applicability

Although this research focuses on policy and regulatory text, the framework readily adapts to other domains characterised by specialized language and subject to an extensive body of internal policies and procedures that follow established structures, hierarchies and cross-referencing. Healthcare regulations (HIPAA, FDA guidelines), HR compliance (employment law, workplace safety), and environmental standards share similar characteristics and stability requirements. The key enabler is domain-specific taxonomy development, once established, our fine-tuning approach allows organisations to produce transferable, leverageable learnings and competencies and create SLMs for any domain where consistency and accuracy outweigh interpretive flexibility.

6.3 Aligning with the Professional Consensus

The strategy of developing a specialised, reliable AI teammate aligns perfectly with the consensus view of legal professionals themselves. The industry does not envision a future of full automation, but rather one of augmented intelligence. Industry reports found that a striking 96% of legal professionals believe AI making final decisions on a legal, tax or trade matter would be a step too far.

This confirms that the demand is not for a replacement, but for a tool that can reliably handle high-volume, low-judgment tasks, freeing up human experts to focus on strategy, nuanced analysis, and client relationships. Therefore, the most pressing technological challenge is not the pursuit of autonomous AI, but the improvement of the human-AI partnership. By proving that a SLM can be a more accurate, stable, and efficient teammate than a generalist LLM, this research demonstrates the most viable path to building AI that professionals will actually trust and adopt at scale.

7 Conclusion

For the high-stakes domain of regulatory workflows, the prevailing methodology of relying on generalist provider-hosted LLMs is suboptimal and fraught with risk. These models, while powerful in a general context, exhibit unacceptable levels of performance instability, lower accuracy, and a prohibitive cost structure when applied to nuanced tasks.

The research provides a clear, empirically validated blueprint for a superior approach. By focusing on the development of specialised SLMs that are fine-tuned on high-quality, proprietary data and self-hosted on controlled infrastructure, we achieve three key strategic advantages: superior accuracy, reliability, and radical cost-efficiency. This approach transforms AI from a volatile external dependency into a stable assistant. It augments human experts and provides significant efficiency gains, focusing on collaboration rather than complete automation.

Furthermore, RegGenome's initial successful exploration of zero-shot generalisation demonstrates a viable path to overcoming the challenge of adapting to new and evolving regulatory taxonomies, reinforcing the long-term strategic value of building an intelligence asset. The future of AI in regulation belongs not to the largest model, but to those who build their own intelligence assets that transform regulatory complexity into competitive advantage.

Acknowledgements

The authors thank the RegGenome team for their invaluable contributions to this research. We also acknowledge the thoughtful feedback from several individuals who read early drafts of this paper including: Bryan Zhang, Raghu Rau, Tim Lane, Robert Dilworth, Giovanni Bandi, Michael Silva, Mike Dewar and Lotte Schou Zibell.

References

[Blair-Stanek, A., & Van Durme, B. (2025)] Blair-Stanek, A., and Van Durme, B. (2025). *LLMs Provide Unstable Answers to Legal Questions*. arXiv. https://arxiv.org/pdf/2502.05196

[Savelka, J., & Ashley, K. (2023)] Savelka, J., & Ashley, K. (2023). *GPT-3.5 for Zero-Shot Semantic Annotation of Legal Texts*. Frontiers in Artificial Intelligence. https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1279794/full

[Ho, D. E. (2024)] Ho, D. E. (2024). *Hallucinating the Law: Legal Mistakes of Large Language Models are Pervasive*. Stanford HAI. https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive

[Andreas H., et al. (2025)] Andreas Hochlehnert and Hardik Bhatnagar and Vishaal Udandarao and Samuel Albanie and Ameya Prabhu and Matthias Bethge (2025).*A Sober Look at Progress in Language Model Reasoning: Pitfalls and Paths to Reproducibility* arXiv. https://arxiv.org/pdf/2504.07086

[Atil, B., et al. (2025)] Berk Atil and Sarp Aykent and Alexa Chittams and Lisheng Fu and Rebecca J. Passonneau and Evan Radcliffe and Guru Rajan Rajagopal and Adam Sloan and Tomasz Tudrej and Ferhan Ture and Zhe Wu and Lixinyu Xu and Breck Baldwin (2025). *Non-Determinism of "Deterministic" LLM Settings*. arXiv. https://arxiv.org/pdf/2408.04667

[NLS Business Law Review. (2024)] NLS Business Law Review. (2024). *The Disadvantages and Limitations of Using Large Language Models in the Field of Law*. National Law School of India University. https://www.nls.ac.in/research/projects/the-disadvantages-and-limitations-of-using-large-language-models-in-the-field-of-law/

[Shambhavi, M., et al. (2024)] Mishra, Shambhavi and Ahmed, Tanveer and Mishra, Vipul and Srivastava, Priyam and Sayeed, Abuzar and Gupta, Umesh (2024). *Rhetorical Role Detection in Legal Judgements Using Zero-Shot Learning*. ResearchGate. https://www.researchgate.net/publication/377392111_Rhetorical_Role_Detection_in_Legal_Judgements_Using_Zero-Shot_Learning

[Sarkar, R., et al. (2021)] Sarkar, Rajdeep and Ojha, Atul Kr. and Megaro, Jay and Mariano, John and Herard, Vall and McCrae, John P.(2021). *Few-shot and Zero-shot Approaches to Legal Text Classification: A Case Study in the Financial Sector*. ACL Anthology. https://aclanthology.org/2021.nllp-1.10/

[Szalontai, B., et al. (2025)] Szalontai, Balázs and Márton, Balázs and Pintér, Balázs and Gregorics, Tibor (2025). *Investigating Reproducibility Challenges in LLM Bugfixing*. Preprints.org. https://www.preprints.org/manuscript/202505.2321/v1

[Fatemi, S., et al. (2024)] Sorouralsadat Fatemi and Yuheng Hu and Maryam Mousavi (2024). *A Comparative Analysis of Instruction Fine-Tuning LLMs for Financial Text Classification*. arXiv. https://arxiv.org/abs/2411.02476

[ABA. (2025)] American Bar Association. (2025). *2025 Legal Technology Survey Report*. ABA Publishing. https://www.americanbar.org/groups/law_practice/resources/law-technology-today/2025/the-legal-industry-report-2025/

[Counselwell & Spellbook. (2025)] Counselwell & Spellbook. (2025). spellbook *AI in Legal Departments: 2025 Benchmarking Report*. Spellbook. https://www.spellbook.legal/blog/counselwell

[Forbes Advisor. (2025)] Forbes Advisor. (2025). *UK Artificial Intelligence (AI) Statistics And Trends In 2025*. Forbes. https://www.forbes.com/uk/advisor/business/software/uk-artificial-intelligence-ai-statistics/

[Xu, C., et al. (2024)] Cheng Xu and Shuhao Guan and Derek Greene and M-Tahar Kechadi (2024). *Benchmark Data Contamination of Large Language Models: A Survey*. arXiv. https://arxiv.org/pdf/2406.04244

[vals.ai] Vals AI. *Proprietary Benchmark Evaluation*. Vals AI. https://www.vals.ai/home

[Belek, P., et al. (2025)] Peter Belcak and Greg Heinrich and Shizhe Diao and Yonggan Fu and Xin Dong and Saurav Muralidharan and Yingyan Celine Lin and Pavlo Molchanov (2025). *Small Language Models are the Future of Agentic AI*. arXiv. https://arxiv.org/pdf/2506.02153

[Daniel Han, Michael Han and Unsloth team (2023)] Daniel Han, Michael Han and Unsloth team (2023). *Unsloth*. GitHub. http://github.com/unslothai/unsloth

[Kwon W. et al. (2023)] Woosuk Kwon and Zhuohan Li and Siyuan Zhuang and Ying Sheng and Lianmin Zheng and Cody Hao Yu and Joseph E. Gonzalez and Hao Zhang and Ion Stoica (2023). *Efficient Memory Management for Large Language Model Serving with PagedAttention*. arXiv. https://arxiv.org/pdf/2309.06180

[Elsworth C. et al. (2025)] Cooper Elsworth and Keguo Huang and David Patterson and Ian Schneider and Robert Sedivy and Savannah Goodman and Ben Townsend and Parthasarathy Ranganathan and Jeff Dean and Amin Vahdat and Ben Gomes and James Manyika. (2025). *Measuring the environmental impact of delivering AI at Google Scale*. arXiv. https://arxiv.org/pdf/2508.15734

[Epoch AI. (2025)] Epoch AI. (2025). *Energy Analysis of ChatGPT and Token-to-Wh Baselines*. Epoch AI. https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use

[Google DeepMind. (2025)] Google DeepMind. (2025). *Gemini 2.5 Pro with Deep Think: Budgeted Reasoning*. Google AI Blog. https://blog.google/products/gemini/gemini-2-5-deep-think

[Anthropic. (2025)] Anthropic. (2025). *Extended Thinking API Documentation*. Anthropic. https://docs.anthropic.com/en/docs/build-with-claude/extended-thinking

[Jegham, N. et al. (2025)] Nidhal Jegham and Marwan Abdelatti and Lassad Elmoubarki and Abdeltawab Hendawi (2025). *How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference*. arXiv. https://arxiv.org/pdf/2505.09598

[xAI. (2024)] xAI. (2024). *Colossus Cluster Infrastructure Specifications*. xAI. https://x.ai/colossus

[Jegham, N. et al. (2025)] Yunho Jin and Gu-Yeon Wei and David Brooks (2025). *The Energy Cost of Reasoning: Analyzing Energy Usage in LLMs with Test-time Compute*. arXiv. https://arxiv.org/pdf/2505.14733

A Appendix

A.1 Example Classification Reports

Table 5: Example L2 Classification Report for `RegGenome`

Class ID	Class Name	Precision	Recall	F1-Score	Support
...	...	...	...	...	...
C003	aml-activitybasedsanction-commercial	0.80	0.67	0.73	6
C004	aml-activitybasedsanction-general	0.30	0.38	0.33	8
...	...	...	...	...	...
C165	aml-tfs-sanctionspermit	0.68	0.74	0.71	23
C166	aml-wiretransfer-information	0.79	0.87	0.83	31
...	...	...	...	...	...
micro avg		0.63	0.58	0.60	1288
macro avg		0.51	0.47	0.46	1288
weighted avg		0.62	0.58	0.58	1288

A.2 Example Prediction Variance Reports

Table 6: Example Raw F1-Scores for Gemini 2.5 Pro (L2) across 5 Runs

Class ID	Run 1 F1	Run 2 F1	Run 3 F1	Run 4 F1	Run 5 F1
C001	0.45	0.42	0.48	0.44	0.46
C002	0.61	0.68	0.63	0.65	0.62
C003	0.22	0.19	0.25	0.21	0.23
...	...	...	...	...	...
C164	0.88	0.91	0.85	0.89	0.90

A.3 Example of training / testing data

Page of text from a regulatory document:

SCHEDULE 5
EVIDENCE AND INFORMATION
Article 12
1. (1) Without prejudice to any other provision of this Order, or to any provision of any other law, the Governor may request any person in or resident in the Territory to furnish any information in their possession or control, or to produce any document in their possession or control, which the Governor may require for the purposes of article 12 of this Order; and any person to whom such a request is made shall comply with it within such time and in such a manner as may be specified in the request.
(2) Nothing in the foregoing sub-paragraph shall be taken to require any person who has acted as counsel or solicitor for any person to furnish or produce any privileged information or document in their possession in that capacity.
(3) Where a person is convicted of an offence under paragraph 4(a) of this Schedule of failing to furnish information or produce a document when requested to do so, the court may make an order requiring them, within such a period as may be specified in the order, to furnish the information or provide the document.
(4) The power conferred by this paragraph to request any person to produce documents shall

include power to take copies of or extracts from any document so produced and to request that person, or, where that person is a body corporate, any other person who is a present or past officer of, or is employed by, the body corporate, to provide an explanation of any of them.
(5) The furnishing of any information or the production of any document under this paragraph shall not be treated as a breach of any restriction imposed by law.
2. (1) If any judge, justice of the peace or magistrate is satisfied by the information on oath given by any police officer, constable or person authorised by the Governor to act for the purposes
of this paragraph either generally or in a particular case that:
(a) there is a reasonable ground for suspecting that an offence under this Order, or with respect to any matters regulated by this Order, an offence relating to customs, has been or is being committed and that evidence of the commission of the offence is to be found on any premises specified in the information, or in any vehicle, ship or aircraft so specified, or
(b) any documents which ought to have been produced under paragraph 1 and have not been produced are to be found on any such premises or in any such vehicle, ship or aircraft, he or she may grant a search warrant authorising any police officer or constable, together with any persons named in the warrant and any other police officers or constables, to enter the premises specified in the information or, as the case may be, any premises upon which the vehicle, ship or aircraft so specified may be, at any time within one month from the date of the warrant and to search the premises, or as the case may be, the vehicle, ship or aircraft.
(2) Any authorised person who has entered any premises or any vehicle, ship or aircraft by virtue of the warrant issued in accordance with sub-paragraph (1) may do all or any of the following things:
(a) inspect and search those premises or the vehicle, ship or aircraft for any material which they have reasonable grounds to believe may be evidence in relation to an offence referred to in this paragraph;
(b) seize anything on the premises or on the vehicle, ship or aircraft which they have reasonable grounds for believing is evidence in relation to an offence referred to in this paragraph;
(c) seize anything on the premises or on the vehicle, ship or aircraft which they have reasonable grounds to believe are to be produced in accordance with paragraph 1; or
(d) seize anything that is necessary to be seized in order to prevent it being concealed, lost, damaged, altered or destroyed.

Annotations provided by a trained regulatory analyst:
“aml-assetfreeze-information”, “aml-assetfreeze-investigation”

A.4 Example prompt structure

You have a page of text from a legislative or financial services document and a series of tags that can be applied to this page along with their definitions.
Your Task:
• Classify the text using the most relevant tags from the list provided.
• Focus on Main Themes: Identify the primary themes, regulatory requirements, and significant content areas directly addressed in the text.
Assign Relevant Tags:
• Assign tags that represent substantial and substantive content, rather than incidental mentions or keywords.
• If the text covers multiple significant themes, you may assign multiple tags, but ensure each tag reflects a substantial aspect of the text.
• When multiple tags seem applicable, choose the one that best represents the main purpose of the provision as interpreted by the analysts.
Handle Overlapping Tags Carefully:
• If a provision could fit under multiple tags, prefer the tag that represents the broader or more foundational concept, unless the text explicitly emphasizes the narrower aspect.

• Avoid assigning overly general tags when a more specific tag is appropriate.
Consider Context and Nuance:
• Do not assign tags based solely on the presence of certain words or phrases; consider the context and overall meaning of the text.
• Be attentive to implicit themes and the broader implications of the provisions.
Output Format:
• Your response should be in the form of a JSON with the key "classification" and value being a list of tags to be applied. For example:
• In the single tag case: "classification":["aml-abcpolicy-assess"].
• In the multiple tag case: "classification":["aml-abcpolicy-assess", "aml-abcpolicy-controls"].
TAGS: [RegGenome taxonomy tags and definitions]
TEXT: [Text from a page of a regulatory document]

A.5 Examples of RegGenome Taxonomy Tags and Descriptions

aml-assetfreeze-information = Power of competent authority to obtain information from financial institutions in relation to sanctioned individuals or entities.
aml-assetfreeze-injunctions = Power of the regulatory authorities to enforce injunctions in relation to sanctions.

A.6 Energy Estimation Methodology for Provider-Hosted Models

A.6.1 Purpose and Scope

This appendix presents a transparent methodology for estimating the energy consumption of provider-hosted frontier LLMs under deliberately intensive workloads: 10,000-token inputs (prefill-dominant), approximately 1,000 "thinking" tokens (test-time compute), and roughly 60 visible output tokens. Given the proprietary nature of provider architectures and serving infrastructure, we present ranges based on explicit assumptions and consistent measurement boundaries, drawing from recent empirical studies and production measurements where available.

A.6.2 Measurement Framework

System Boundaries and Overhead Accounting We distinguish between IT-only energy consumption (accelerators and host systems during active compute) and full-stack energy consumption (IT infrastructure plus idle capacity, networking, and facility overhead via Power Usage Effectiveness). Following Google's production-instrumented study of Gemini Apps, which reports median text prompt consumption of 0.24 Wh end-to-end, we provide IT-centric central values while incorporating realistic facility PUE factors (1.1–1.3) to account for total system overhead. [Elsworth C. et al. (2025)]

Workload Decomposition Autoregressive inference comprises distinct operational phases: prefill (parallel processing of input tokens, compute-bound) and decode (sequential token generation from KV cache, memory-bound). For models with test-time compute capabilities, we treat extended "thinking" as additional decode work potentially involving parallel hypothesis generation and self-refinement. Total energy is therefore modeled as:

$$E_{total} = E_{prefill} + E_{decode} + E_{TTC} \tag{3}$$

where E_{TTC} represents the test-time compute overhead.

Baseline Calibration For modern deployments on Hopper-class GPUs, we adopt Epoch AI's empirical estimate of approximately 2.5 Wh for 10,000-token prefill as our baseline for efficient LLM stacks. This anchors all models to consistent long-context geometry before applying model-specific multipliers. [Epoch AI. (2025)]

A.6.3 Model-Specific Estimations

Google Gemini 2.5 Series For Gemini 2.5 Flash, optimised for speed and efficiency with adjustable thinking budgets, we estimate 3.0 Wh (range: 2.5–3.5 Wh IT-only), applying a modest 1.2× multiplier to account for limited decode and minimal thinking overhead. Gemini 2.5 Pro which employs parallel hypothesis exploration and extended inference, requires a higher 2.5× multiplier, yielding 6.25 Wh (range: 5.5–7.5 Wh IT-only). [Google DeepMind. (2025)]

Anthropic Claude 4 Series Claude Sonnet 4, positioned for balanced performance with optional extended thinking, receives a 1.8× multiplier over baseline, resulting in 4.5 Wh (range: 4.0–5.5 Wh IT-only). Claude Opus 4.1, targeting complex reasoning tasks with larger thinking budgets, requires a 3.5× multiplier, yielding 8.75 Wh (range: 7.5–10.0 Wh IT-only). [Anthropic. (2025)]

OpenAI GPT-5 External studies report average consumption of approximately 18.35 Wh for 1,000 tokens with peaks approaching 40 Wh, derived from measured response times and assumed server power on Hopper-class systems. For our heavier workload specification, we adopt 25–40 Wh as the plausible range. [Jegham, N. et al. (2025)]

xAI Grok 4 While xAI's Colossus cluster operates at supercomputer scale (200,000 GPUs, 280–300 MW capacity), per-query energy depends critically on concurrency and workload distribution. Applying comparable reasoning multipliers to our baseline, we estimate 18–30 Wh for the specified workload, acknowledging substantial uncertainty without direct instrumentation. Additionally, during the study it was noted that the xAI API does not allow one to limit the number of thinking tokens the model used as others did, meaning this model in particular is likely to be at the higher end or even exceed the estimate provided here. [xAI. (2024)]

A.6.4 Validation and Limitations

Our estimates align with independent measurements showing that long-context prefill dominates energy consumption for heavy prompts, with test-time compute introducing material multipliers ranging from 1.2× to 3.5× depending on thinking budget and parallelism. The consistency between our projections and external GPT-5 measurements provides additional validation of the methodology. [Jegham, N. et al. (2025)]

Key limitations include:

- **Architectural Opacity:** Active expert counts for mixture-of-experts models, exact test-time compute implementations, and batching policies remain proprietary.
- **Boundary Sensitivity:** IT-only versus full-stack reporting can differ by substantial factors depending on utilisation and facility efficiency.
- **Workload Specificity:** Estimates apply specifically to the heavy workload profile; median consumer workloads may consume orders of magnitude less energy.

A.6.5 Summary of Central Estimates

These estimates provide order-of-magnitude guidance for comparing the energy efficiency of provider-hosted models against self-hosted alternatives, while acknowledging the inherent uncertainty in proprietary system configurations.

Table 7: Energy Consumption Estimates for Provider-Hosted Models (IT-centric)

Model	**Central Estimate (Wh)**	**Range (Wh)**
Gemini 2.5 Flash	3.0	2.5–3.5
Gemini 2.5 Pro	6.25	5.5–7.5
Claude Sonnet 4	4.5	4.0–5.5
Claude Opus 4.1	8.75	7.5–10.0
GPT-5	32.5	25–40
Grok 4	28	18–30