Legal Semantic Engineering: Reconciling Probabilistic Generation with Rigid Normative Constraints

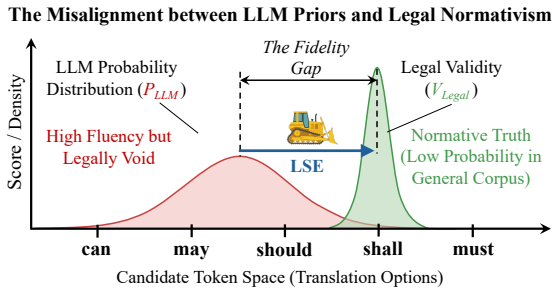
Anonymous ACL submission

Abstract

Legislative translation demands the precise preservation of normative intent. However, Large Language Models (LLMs) frequently suffer from "Deontic Drift"—a systemic failure where models prioritize probabilistic fluency over rigid normative mandates. By analyzing a five-jurisdiction benchmark via Zipf-Mandelbrot modeling, we characterize this failure as a structural distributional mismatch: the high-concentration mandatory monopoly of source legal terms diverges significantly from the granular, dispersed probability distributions of target languages. To bridge this gap, we propose **Legal Semantic Engineering (LSE)**, a framework that introduces vertical hierarchical control as a robust alternative to horizontal multi-agent collaboration. Through an Anchoring-Shaping-Polishing (ASP) pipeline, LSE explicitly decouples normative logic validation from stochastic text generation. Experiments on a trilingual legislative benchmark demonstrate that LSE is highly robust to backbone variations; implementations using DeepSeek, GPT, and Gemini all significantly surpass strong horizontal agent baselines. Furthermore, our analysis unveils the *gain-interference-rescue* dynamics, quantitatively illustrating the necessary trade-offs between linguistic fluency and legal fidelity.

1 Introduction

While the rapid evolution of Large Language Models (LLMs) (OpenAI, 2023; Touvron et al., 2023; Anil et al., 2023; DeepSeek-AI et al., 2025; Yang et al., 2025) has revolutionized Neural Machine Translation (NMT) (Sutskever et al., 2014; Cho et al., 2014; Vaswani et al., 2017; Gu et al., 2019; Liu et al., 2020; Fan et al., 2021; He et al., 2024; Wu et al., 2025) with unprecedented linguistic fluency (Kocmi et al., 2024; Zhang et al., 2025a), their deployment in high-stakes legislative domains remains precarious (Niklaus et al., 2025). Unlike general prose, legislative texts are not merely descriptive; they are performative utterances that establish, modify, or extinguish legal obligations (Austin, 1975; Cao, 2007; Cheng and Sin, 2011). In this context, the "stochastic" nature of LLMs (Bender et al., 2021) becomes a liability rather than an asset, as the primary objective transcends surface-level smoothness to achieve normative fidelity—the precise preservation of sovereign power across jurisdictional boundaries (Godfrey and Burdon, 2024; Ariai et al., 2025; Graziadei, 2025).



Input: As a legal translation specialist, translate the following text from Chinese to English. 科学技术研究开发机构和高等学校 ***应当*** 加强对科学技术人员的科学技术知识 ···

LLM's Output: *Scientific and technological R&D institutions and higher education institutions* ***should*** *strengthen the cultivation of scientific ...*

Reference: *Scientific and technological R&D institutions and higher education institutions* ***shall*** *strengthen the cultivation of scientific ...*

Figure 1: The red curve (P_{LLM}) illustrates *Deontic Drift*, where models prioritize high-fluency but legally incorrect tokens (e.g., "should"). The proposed LSE framework (blue arrow) bridges this *Fidelity Gap* by recalibrating the output to ensure the precise preservation of legal force.

The fundamental challenge stems from a deep **structural asymmetry** between source and target legal systems (Šarčević, 1997, 2000; Cao, 2007; Al-Saeed and Abdulwahab, 2023). Chinese legislative discourse employs a high-concentration "mandatory monopoly," where a single operator "应当" (*yīngdāng*) universally signifies strict obligation (Cheng and Sin, 2011; Gong et al., 2020), which is predominantly translated into "shall" in the of-

ficial translations. In contrast, the English target distribution is granularly fragmented into a spectrum of modals (e.g., *shall, must, should, may*), each carrying distinct deontic weights (Chan, 2011; Garzone, 2013). This mismatch creates a "Deontic Ambiguity" trap (Lyu et al., 2024): while "shall" carries the force of law, the statistically more common "should" often implies mere recommendation (Ballesteros-Lintao et al., 2016; Sancheti et al., 2022).

As illustrated in Figure 1, this misalignment is not a random error but a systemic consequence of the Fidelity Gap (Bender et al., 2021; Dobriban, 2025). An LLM's probability distribution (P_{LLM}) is shaped by general corpora (Bender et al., 2021), where colloquial couplings (e.g., "should") dominate. Consequently, the model is statistically gravitated toward "High Fluency but Legally Void" candidates (Zhang et al., 2025b). When the peak of legal validity (V_{Legal}) lies in the low-probability tail of the general distribution, the model remains "trapped" by its likelihood objective, lacking the specialized prior to navigate the normative gradient of the token space (Palmer, 2001; Ballesteros-Lintao et al., 2016; Wu and Aji, 2023).

Standard mitigation strategies, such as Supervised Fine-Tuning (SFT) or Prompt Engineering (Brown et al., 2020; Ouyang et al., 2022), often prove insufficient. Since these methods still fundamentally optimize for statistical likelihood, they struggle to override the deep-seated probabilistic priors ingrained during pre-training (Bender et al., 2021). To bridge the Fidelity Gap, we require a mechanism that moves beyond soft guidance to perform hard normative intervention (Hu et al., 2019; Xu et al., 2023), effectively shifting the decoding objective from "what is probable" to "what is valid".

In response, we propose **Legal Semantic Engineering (LSE)**, a framework designed to recalibrate the LLM's output toward normative truth (Xu et al., 2023; Hu et al., 2025). Unlike standard decoding strategies, LSE imposes vertical hierarchical control. By operationalizing an Anchoring-Shaping-Polishing (ASP) pipeline, our method acts as a regulatory mechanism that decouples logic validation from text generation, ensuring that the "Mandatory Monopoly" of the source text is structurally projected onto the target (Šarčević, 1997; Šarčević, 2016; Chalkidis et al., 2020; Meng et al., 2025).

Our contributions are three-fold:

- **Systemic Diagnosis:** We provide the first quantitative characterization of the "Fidelity Gap" using Zipf-Mandelbrot modeling, empirically substantiating the structural incompatibility between the concentrated source norms and granular target distributions.
- **Methodological Framework:** We introduce Legal Semantic Engineering (LSE), a framework that introduces vertical hierarchical control as a robust alternative to horizontal agentic collaboration, strictly enforcing immutable constraints via the Anchoring-Shaping-Polishing (ASP) pipeline.
- **Behavioral Insights:** We reveal the Gain-Interference-Rescue (GIR) dynamics, quantifying the non-linear "interference cost" necessary to prioritize legal validity over linguistic smoothness, offering a new behavioral metric for evaluating high-stakes generation.

2 Systemic Diagnosis: Quantifying Structural Asymmetry

To transition from qualitative observation to engineering intervention, we formalize the challenge of legal transfer as a problem of distributional alignment. We posit that the "Deontic Drift" observed in LLMs is not a stochastic error but a deterministic failure to reconcile the model's learned probability priors with the rigid normative intensity of the source legal system.

2.1 Experimental Grounding: The Pentagonal Benchmark

To decouple linguistic artifacts from jurisprudential structures, we expand our diagnostic scope to a pentagonal benchmark encompassing five legal jurisdictions. This specific constellation was selected to rigorously isolate the impact of legal lineage from linguistic typology, with detailed environmental profiles and typological breakdowns provided in Appendix C. As illustrated in Figure 2, this design covers the genealogical evolution of Civil Code:

1. **CN** (China): The source environment, representing a high-intensity legislative style characterized by rigid unitary operators (a "Mandatory Monopoly").
2. **DE** (Germany) & **JP** (Japan): Serving as the "Germanic" control group, these jurisdictions

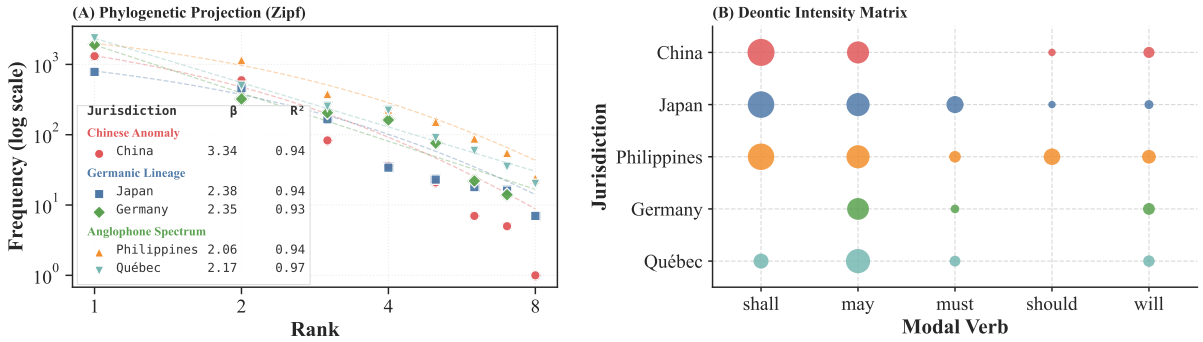


Figure 2: Phylogenetic Projection of Legal Norms. (A) Rank-frequency distributions fitted with the Zipf-Mandelbrot law. The slopes clearly distinguish the steep **Chinese Anomaly** from the flatter **Germanic** and **Anglophone** lineages. (B) Deontic intensity matrix (bubble area $\propto$ frequency) confirming the structural bifurcation between the source's mandatory monopoly and the target's distributed modality.

examine whether structural steepness is a feature of the Civil Law lineage rather than a language-specific artifact.

3. **PH** (Philippines) & **QC** (Québec): Representing "Anglophone Civil Law," these regions provide the critical target baseline—Civil Law logic articulated within the English linguistic topology.

2.2 Formalization: The Power-Law of Normative Force

We quantify the systemic distribution of legal force by modeling the rank-frequency profile of deontic operators. While natural language distributions are typically characterized by the Zipf-Mandelbrot law (Mandelbrot, 1961) ($f(r) \propto (r+\gamma)^{-\beta}$), our empirical analysis of the legal lexicon reveals a strong fit to power-law distributions in the high-rank deontic domain. Consequently, we employ the standard form (assuming $\gamma \to 0$) to explicitly capture the decay rate:

$$f(r) = \frac{\alpha}{r^{\beta}} \tag{1}$$

Here, the exponent β serves as the structural signature of the legal system. A high β (Steep Slope) indicates a "Mandatory Monopoly," where legislative power is concentrated in a singular lexical operator (e.g., "应当"). Conversely, a low β (Flat Slope) reflects a "Granular Spectrum," where normative force is distributed across a diverse lexical field.

2.3 Phylogenetic Interpretation

The projection in Figure 2 reveals a striking correlation between mathematical distribution and legal phylogeny. The **Germanic Lineage** (DE & JP) exhibits high consistency ($\beta \approx 2.36$), quantitatively reflecting the historical reception of the German Civil Code by Japan. In contrast, the **Chinese Anomaly** stands as a structural outlier with the steepest descent ($\beta \approx 3.34$), visually separated from the flatter **Anglophone Spectrum** ($\beta \approx 2.1$). This confirms that Chinese legislation relies on a single-point reliance mechanism for obligation, whereas English systems utilize distributed redundancy.

2.4 The Mechanism of Drift: A Probability Mass Shift

This distributional disparity demonstrates that translating Chinese law into English is not a bijective mapping, but a projection from a low-variance distribution to a high-variance distribution. We argue that deontic drift is a statistically expected behavior in probabilistic models.

LLMs, driven by likelihood maximization, tend to perform probability smoothing. When mapping the rigid, concentrated source signal (CN, Type-3.3) into the dispersed target space (EN, Type-2.1), the model tends to redistribute the sharp probability mass of the source mandate (the "peak") into the flatter "plains" of the target distribution. The result is a text that is linguistically smooth (high probability in general corpora) but legally diluted—the normative intensity of "shall" is lost to the statistical gravity of high-frequency tokens like "should."

3 Methodology: Legal Semantic Engineering via Hierarchical Control

The diagnosis in Section 2 establishes that "Deontic Drift" is a deterministic consequence of the structural entropy mismatch between the source (Type-3.3) and target (Type-2.1) manifolds. Stan-

dard multi-agent frameworks, such as *TransAgents* (Wu et al., 2024b, 2025), rely on horizontal collaboration (i.e., peer debate) (Liang et al., 2024). While effective for creative tasks requiring diverse feedback, we argue this topology is fundamentally flawed for legislative transfer due to three systemic risks:

- **Semantic Drift & Averaging Effect:** In horizontal negotiation, agents tend to converge on the "mean" of the probability distribution (Wu and Aji, 2023). This sacrifices the rigid "Mandatory Monopoly" of legal terms for statistically safer, colloquial options (e.g., compromising on "should" instead of "shall"), creating a "High Fluency, Low Fidelity" trap.
- **Reciprocal Hallucination:** Without a fixed reference, agents sharing similar pre-training biases (e.g., GPT-4o peers) often cross-validate each other's errors, reinforcing incorrect deontic choices (Zhang et al., 2025b).
- **Convergence Issues:** Iterative debate often suffers from diminishing returns, consuming excessive tokens while trapping the system in local optima far from the professional legal manifold (Liang et al., 2024).

To resolve these issues, LSE shifts the paradigm from *horizontal democracy* to *vertical hierarchical control*. As shown in Figure 3, we implement a strictly unidirectional **Anchoring-Shaping-Polishing (ASP)** pipeline, where downstream generation is mathematically constrained by upstream normative locks (Hu et al., 2019; Xu et al., 2023).

3.1 Layer 1: Terminology Anchoring (Conceptual Gain)

The Defense of External Validity. Standard NMT relies on implicit parametric knowledge, which is susceptible to catastrophic forgetting or hallucination (Koehn and Knowles, 2017). However, legislative translation is distinct: it is strictly governed by authoritative external constraints (e.g., official gazettes, standardized term bases) (Šarčević, 1997; Cao, 2007). LSE explicitly integrates these resources not as auxiliary prompts, but as non-negotiable boundary conditions.

- **Deterministic Extraction & Mapping:** The system first identifies legal entities (E) in the source text X. Instead of probabilistic sampling, we enforce a discrete mapping function $f : E \to T_{ref}$, where T_{ref} is a verified terminology set.
- **Lexical Lock:** The mapped terms are frozen as immutable constants. Unlike "soft constraints" in prompt engineering which the model may override for fluency, these anchors define a reduced search space $S' \subset S$, preventing the model from exploring synonymous but legally invalid alternatives, e.g., forbidding *"unforeseeable events"* when *"force majeure"* is required (Meng et al., 2025). This ensures that the **Conceptual Gain** is secured before any syntactic generation begins.

3.2 Layer 2: Syntactic Shaping (Normative Interference)

Once terminology is anchored, the next challenge is preserving the deontic intensity. General LLMs, driven by likelihood maximization (P_{LLM}), naturally drift towards high-frequency modals, e.g., "should" (Bender et al., 2021).

- **Deontic Protocol Injection:** We intervene in the decoding process by imposing a deontic protocol. For instance, a source operator like "应当" (strict obligation) is forcibly mapped to the target token "shall", explicitly pruning high-probability candidates like "must" or "should" from the candidate pool (Cheng and Sin, 2011; Gong et al., 2020). This mechanism operates analogously to the hypothesis pruning in phrase-based SMT (Koehn, 2010), where low-fidelity paths are discarded regardless of their language model probability.
- **Structural Rigidity:** This layer generates a "syntactic skeleton"—a sentence structure built exclusively to support the L1 anchors and L2 deontic modals.
- **Normative Interference:** We acknowledge that this rigorous constraint satisfaction introduces a "fluency penalty", creating a text that is logically sound but linguistically stiff. We define this temporary drop in smoothness not as an error, but as the necessary energy cost (**Interference**) required to override the model's colloquial priors (Hu et al., 2019).

3.3 Layer 3: Discourse Polishing (Manifold Alignment)

The final stage resolves the tension between legal rigidity and linguistic readability.

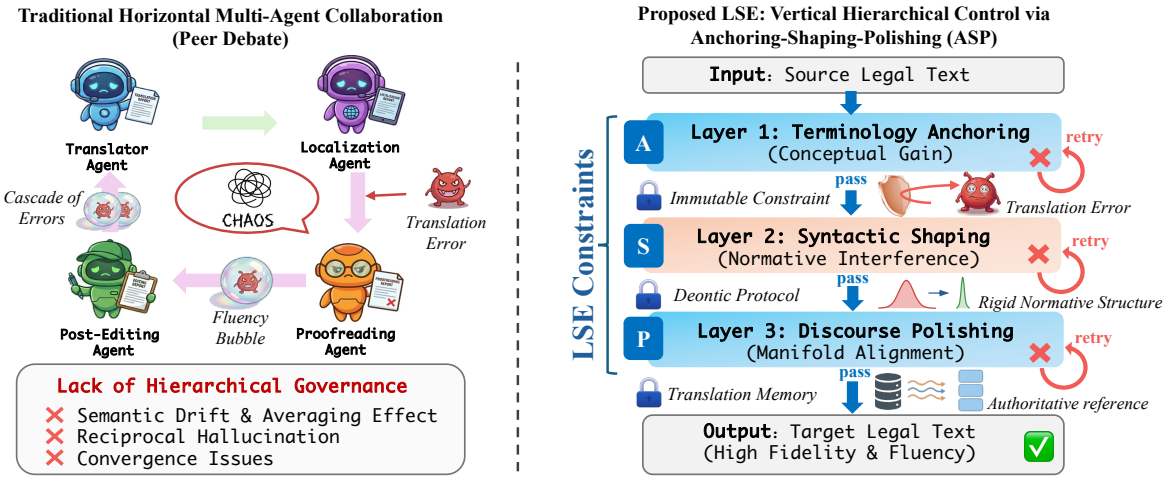


Figure 3: Horizontal peer debate (left) vs. vertical LSE (right). Horizontal systems lack governance, leading to semantic drift and reciprocal hallucination. In contrast, LSE employs a unidirectional **Anchoring-Shaping-Polishing (ASP)** pipeline. The "locks" enforce immutable constraints at terminology, syntactic, and discourse levels to prevent legal mandate dilution.

- **Constrained Smoothing:** The L3 agent receives the syntactic skeleton from Layer 2. Its optimization objective is to maximize $P_{fluency}(Y|X_{skeleton})$ subject to the constraint that $X_{skeleton}$ remains topologically invariant.

- **Manifold Alignment:** The agent is permitted to adjust surface-level features (connectives, prepositions, word order) to align the text with the target language's professional norms, but is digitally forbidden from altering the lexical locks (L1) or deontic modals (L2). This **Rescue** phase ensures the final output achieves high metrics in both fidelity and fluency, effectively bridging the Fidelity Gap (Garzone, 2013; Biel et al., 2019; Reheman et al., 2025).

3.4 Implementation: Decoupled Optimization via Cascading Agents

To execute the LSE framework, we deploy an LLM-based Multi-Agent System (MAS) structured as a cascading refinement architecture. This design acts as a structural corrective to the lack of hierarchical governance prevalent in traditional multi-agent systems (as diagnosed in Figure 3).

- **Vertical Topology vs. Horizontal Debate:** Standard frameworks like *TransAgents* rely on horizontal peer debate, which often leads to *Convergence Issues* and *Reciprocal Hallucination* where agents cross-validate errors (Wu et al., 2024c; Zhang et al., 2025a). In contrast, our system operates on a strictly vertical axis. By enforcing a unidirectional flow ($L1 \rightarrow L2 \rightarrow L3$), we eliminate the risk of circular consensus loops, ensuring that generation is an additive process of refinement rather than a destructive process of negotiation.

- **Decoupled Optimization of State Spaces:** We structurally decouple the optimization objectives to prevent "Contextual Dilution", assigning orthogonal goals to each layer:
  - **L1 (Accuracy Space):** Optimizes for *Exact Match* of entities against external authoritative TermBases, minimizing Semantic Drift.
  - **L2 (Logic Space):** Optimizes for *Constraint Satisfaction* of deontic rules, prioritizing validity over probability.
  - **L3 (Fluency Space):** Optimizes for *Perplexity Reduction* (Linguistic Smoothness), but strictly bounded by the constraints inherited from L1 and L2.

- **Injection via Immutable Prompts:** Vertical control is legally enforced through prompt injection. The output O_{i-1} of the upstream agent is not treated as "context" but as an immutable ground truth injected into the downstream agent's system prompt. For instance, the L3 agent is explicitly instructed: *"Optimize the fluency of the text, but you are digitally forbidden from altering the modal verbs locked by the previous layer."* This method-

Category	System / Model	Chinese → English (Zh-En)				Chinese → Japanese (Zh-Ja)			
		BLEU↑	BERTScore↑	COMET↑	GEMBA↑	BLEU↑	BERTScore↑	COMET↑	GEMBA↑
Commercial NMT	ALIBABA/BAIDU	32.88	93.25	82.97	-	46.79	91.04	91.50	-
General LLMs	QWEN3-8B	29.91	94.73	84.16	94.71	45.74	90.64	90.75	**92.15**
	DEEPSEEK-V3	30.34	94.81	84.86	94.73	45.63	91.21	91.52	91.60
	GPT-4O	31.68	93.72	84.57	94.13	44.04	91.20	92.60	91.40
	GEMINI-2.5-PRO	36.79	95.58	85.88	94.36	49.38	91.36	92.65	91.14
Multi-Agent	TRANSAGENTS	30.72	93.09	83.23	94.10	44.95	**94.58**	91.65	91.25
LSE (ours)	QWEN3-8B	**35.08**‡	**94.99**‡	84.16	**94.84**†	**48.30**‡	91.90‡	**92.85**‡	91.35
	DEEPSEEK-V3	**35.14**‡	**95.09**‡	84.84	**94.80**†	**47.11**‡	92.15‡	**92.75**‡	**91.85**‡
	GPT-4O	**37.38**‡	**95.93**‡	**86.06**‡	**94.35**‡	**48.88**‡	92.20‡	**92.93**‡	**91.70**‡
	GEMINI-2.5-PRO	**39.29**‡	**96.51**‡	**86.85**‡	**94.50**‡	**52.12**‡	92.55‡	**92.95**‡	**91.25**†

Table 1: Main results on the legislative translation benchmark. LSE consistently achieves SOTA performance across backbones. **Boldface** indicates LSE outperforms the respective backbone baseline. Statistical significance markers indicate improvements over the respective general LLM baselines: ‡ for $p < 0.001$ and † for $p < 0.05$. **Note on GEMBA:** While LSE drives massive gains in fidelity (BLEU/COMET), GEMBA scores remain stable, reflecting the trade-off between rigid normative validity and generic probabilistic smoothness.

ology effectively transforms the "stochastic parrots" (Bender et al., 2021) into verifiable legal engineers.

4 Experimental Setup

To validate the efficacy of the LSE framework, we conduct a comprehensive evaluation addressing three core research questions:

- **RQ1 (Effectiveness):** Can LSE effectively mitigate Deontic Drift and outperform SOTA baselines (including horizontal multi-agent systems) in high-stakes legislative translation?
- **RQ2 (Mechanism):** Does the "Gain-Interference-Rescue" (GIR) dynamic empirically emerge as hypothesized? Specifically, is the *Interference Cost* (a temporary drop in fluency) a necessary trade-off for ensuring normative validity?
- **RQ3 (Human Alignment):** Can vertical control resolve the *Fidelity-Fluency Inversion* typically observed in probabilistic LLMs, where high readability often masks legal inaccuracy?

Dataset. We utilize a TRILINGUAL LEGISLATIVE BENCHMARK, a curated collection of 35 Chinese official statutes aligned with official English translations and expert-verified Japanese translations. To ensure cross-domain robustness, the corpus covers three distinct categories: Core Codes (e.g., Civil, Penal), Administrative Regulations (e.g., Environmental, Education), and Specialized Commercial Laws (e.g., Intellectual Property, Maritime). We implement a strict split: 30 statutes are utilized for TermBase/Translation Memory construction (extracting external constraints for Layer 1 and Layer 3), while the remaining 5 statutes are reserved exclusively for testing (see Appendix A).

Baselines. We compare LSE against three distinct approaches: (1) **Commercial NMT**: SOTA systems from Google, Microsoft, Alibaba, and Baidu. (2) **General LLMs**: GPT-4O (OpenAI, 2023), GEMINI-2.5-PRO (Anil et al., 2023), and DEEPSEEK-V3 (DeepSeek-AI et al., 2025) operating in a standard zero-shot setting. (3) **Horizontal Multi-Agent**: TRANSAGENTS (Wu et al., 2024c), representing the SOTA peer-debate framework without hierarchical constraints.

Metrics. Beyond standard automatic metrics including BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2019), COMET (Rei et al., 2020), and GEMBA (Kocmi and Federmann, 2023). We also employ expert human evaluation, which is conducted on a 1-5 Likert scale assessing two distinct dimensions: Normative Fidelity and Fluency (refer to Appendix G for scoring guidelines).

5 Main Results and Analysis

5.1 Vertical Control vs. Horizontal Collaboration

As shown in Table 1, LSE consistently outperforms baselines across all metrics.

Superiority over Horizontal Agents. The most critical comparison reveals that LSE significantly outperforms the horizontal TRANSAGENTS baseline. On the Zh-En task, our best configuration (Gemini-backed, see Table 1) achieves 39.29 BLEU, surpassing TRANSAGENTS (30.72) by a

Configuration	Chinese → English					Chinese → Japanese				
	Automated Metrics		Human Evaluation (1-5)			Automated Metrics		Human Evaluation (1-5)		
	BLEU	COMET	Fidelity	Fluency	Total Avg	BLEU	COMET	Fidelity	Fluency	Total Avg
Baseline	31.68	84.57	$3.53_{\pm 0.93}$	$3.76_{\pm 1.26}$	$3.61_{\pm 0.75}$	44.03	92.60	$4.16_{\pm 0.45}$	$4.26_{\pm 0.56}$	$4.19_{\pm 0.35}$
+ Anchoring (L1)	35.52	84.76	$3.68_{\pm 0.89}$	$3.81_{\pm 1.17}$	$3.72_{\pm 0.71}$	47.10	92.80	$4.45_{\pm 0.41}$	$4.58_{\pm 0.50}$	$4.49_{\pm 0.32}$
+ Shaping (L2)	35.78	84.79	$3.61_{\pm 0.88}$	$3.73_{\pm 1.12}$	$3.65_{\pm 0.69}$	47.51	92.76	$4.37_{\pm 0.43}$	$4.61_{\pm 0.53}$	$4.45_{\pm 0.34}$
+ Full LSE (L3)	**37.38**‡	**86.06**‡	**3.87**‡$_{\pm 0.86}$	**3.91**‡$_{\pm 1.22}$	**3.88**‡$_{\pm 0.70}$	**48.88**‡	**92.93**‡	**4.52**‡$_{\pm 0.40}$	**4.72**‡$_{\pm 0.49}$	**4.59**‡$_{\pm 0.31}$

Table 2: Ablation results (GPT-4o backbone). While automated metrics show linear growth, human evaluation reveals the *GIR Dynamics*. The Syntactic Shaping layer (L2) induces a distinct interference cost, a drop in Fidelity and Fluency (underlined) due to normative rigidity, which is rescued by Layer 3.

decisive margin of +8.57 points. Qualitative analysis suggests that while horizontal agents suffer from reciprocal hallucination (compromising on safe but legally weak terms), LSE's vertical locks successfully enforce the mandatory monopoly of the source text.

Cross-Model Robustness. LSE demonstrates remarkable robustness. Implementations across different backbones (QWEN-8B,DEEPSEEK-V3, GPT-4O, GEMINI-2.5-PRO) all show significant gains over their respective zero-shot baselines, with improvements ranging from +2.5 to +5.1 BLEU. This confirms that the performance boost stems from the ASP architecture rather than the parametric knowledge of any single model.

Statistical Significance. Paired t-tests confirm that the performance gains are robust. The full LSE framework achieves extremely strong significance on precision-based metrics like BLEU and CHRF across all tested backbones. Notably, even for metrics with marginal numerical gains like GEMBA, LSE demonstrates statistically significant improvements , proving the reliability of our hierarchical control in correcting subtle deontic errors that are often overlooked by general-purpose evaluators.

5.2 The Dynamics of Control: Analyzing GIR Effects

To understand the cost of rigidity, we track layer-wise performance evolution using the ablation study (Table 2) and a micro-case study (Table 3).

Phase 1: Conceptual Gain (+Anchoring). Injecting Terminology (L1) yields an immediate gain, boosting BLEU from 31.68 to 35.52 (Zh-En). This confirms that lexical precision is the foundation of legal fidelity.

Phase 2: Normative Interference (+Shaping). Crucially, when adding the Syntactic Layer (L2), we observe the interference cost. While automated metrics remain relatively static (35.52 → 35.78), human evaluation reveals the hidden cost of rigidity: both Fidelity and Fluency scores drop (e.g., Fluency falls from 3.81 → 3.73). This empirical dip validates our hypothesis: to enforce strict deontic modals, the system must actively disrupt the LLM's probability manifold, resulting in valid but linguistically stiff structures.

Phase 3: Discourse Rescue (Full ASP). Finally, the Discourse Polishing (L3) layer leverages Translation Memory to harmonize the text. This restores fluency (rising to 3.91) and fidelity (peaking at 3.87), creating the distinctive ✓ trajectory. This proves that L3 successfully reconciles the structural rigidity of L2 with linguistic naturalness.

Stage	Output Fragment	State Analysis
Baseline	*strengthen innovative spirit...*	✗ **Drift**: Wrong Term
Layer 1	*...strengthen **innovation** spirit...*	✓**Gain**: Term Fixed
Layer 2	*...are required to strengthen...*	**Interference**: Rigid Syntax
Layer 3	***shall** strengthen...*	✓**Rescue**: Canonical Form

Table 3: Micro-case study. Layer 2 introduces necessary "interference" (validity over fluency), which is subsequently rescued by Layer 3.

5.3 Resolving the Fidelity-Fluency Inversion

Baseline LLMs exhibit a dangerous misalignment, acting as "stochastic parrots" (Bender et al., 2021) that prioritize readability (Fluency ≈ 3.76) over legal equivalence (Fidelity ≈ 3.53). This phenomenon confirms the *Fidelity-Fluency Inversion* hypothesis, where stylistic smoothness often masks normative errors (Zhang et al., 2025a). LSE successfully bridges this fidelity gap by elevating fidelity from the mediocre range to the high-precision tier (3.87-4.52). Crucially, this gain does not compromise readability. While the Syntactic Layer (L2) introduces necessary rigidity (the *Interference Cost*), the RAG-enhanced Discourse Layer (L3) restores fluency, ensuring the output is both legally binding and linguistically professional.

6 Related Work

6.1 MT in High-Stakes Legal Domains

Traditional neural machine translation (NMT) has long struggled with domain-specific terminological consistency (Chalkidis et al., 2020; Geng et al., 2021; Niklaus et al., 2025). The advent of LLMs has shifted the challenge from lexical coverage to normative fidelity. As noted in recent evaluations of the "LLM Era," while models achieve human-parity in fluency, they frequently suffer from a fidelity-fluency inversion in specialized domains (Kocmi et al., 2024; Zhang et al., 2025a). In high-stakes legislative translation, stylistic smoothness often masks systemic erosions of legal validity (Ariai et al., 2025). This phenomenon is deeply rooted in the functional equivalence requirements of legal discourse (Šarčević, 1997; Cao, 2007). Unlike general prose, legislative translation demands the precise preservation of legal force across distinct jurisprudential frameworks (Palmer, 2001; Halliday, 1994). Recent studies on Chinese and Japanese legislative corpora highlight that probabilistic sampling often leads to deontic drift, where the mandatory intent of the sovereign is diluted by the model's preference for colloquial modals (e.g., *should* vs. *shall*) (Chan, 2011; Ballesteros-Lintao et al., 2016). Our work leverages Zipf-Model to quantitatively characterize this structural asymmetry (Section 2), providing a formal grounding for the necessity of intervention.

6.2 Constrained Generation and Agentic Control

While modern LLMs prioritize likelihood maximization, our approach revisits the rigid search space control foundational to Statistical Machine Translation (SMT) (Koehn, 2010). Previous work in lexically constrained decoding attempted to enforce terminology via grid beam search or soft masking (Hu et al., 2019; Gu et al., 2018). LSE extends this lineage by elevating constraints from the lexical level to the structural and discursive levels, effectively re-implementing the distinct feature functions of log-linear models (Koehn et al., 2003) as explicit agentic layers. Multi-Agent Systems (MAS) have demonstrated superiority in complex reasoning through role specialization (Yao et al., 2023; Li et al., 2023). However, existing translation agents typically adopt a **horizontal organization** (e.g., peer debate or iterative refinement) (Wu et al., 2024c, 2025). While effective for creative refinement, we argue that horizontal consensus lacks a definitive ground truth mechanism for normative tasks, potentially leading to "hallucinated consensus" (Zhang et al., 2025b). Our LSE framework addresses this by introducing a **vertical control hierarchy**. Building on multi-granularity control concepts (Lyu et al., 2024), LSE ensures that statistical outputs are filtered through deterministic legal locks (Wu et al., 2024a), functionally operationalizing legislative intent as a computable constraint rather than a negotiated outcome.

6.3 Behavioral Interpretability in Generation

Explainability is foundational to the responsible deployment of NLP in law (Ariai et al., 2025). Unlike conventional interpretability research that focuses on internal neuron probing (Belinkov, 2022), our study adopts a **behavioral interpretability** perspective. We introduce the *Gain-Interference-Rescue* (GIR) dynamics to quantify the trade-off between symbolic constraints and probabilistic sampling. This non-linear trajectory offers a tangible metric for assessing model reliability (Dobriban, 2025). This distinction is crucial to counter the epistemic narratives surrounding AI (Chamoun et al., 2025). As cautioned by the "stochastic parrots" critique (Bender et al., 2021), perceived reasoning is often a statistical illusion. By visualizing the *Interference Cost* required to correct deontic drift, LSE provides empirical evidence that high-fidelity legal generation requires active deviation from the model's natural probability manifold, shifting from black-box reliance to verifiable engineering.

7 Conclusion

We address the challenge of legislative translation by formalizing the "Fidelity Gap" through Zipf-Mandelbrot modeling. To mitigate the observed "Deontic Drift", we propose Legal Semantic Engineering (LSE), replacing the prevailing horizontal agent debate with strict vertical hierarchical control. Through the Anchoring-Shaping-Polishing pipeline, we demonstrate that deterministic logic can be successfully injected into stochastic generation. The observed GIR dynamics confirm that the pursuit of fidelity imposes a measurable cognitive load on the model, which must be structurally managed. Ultimately, this work bridges the structural wisdom of classical alignment with the generative power of modern LLMs, ensuring that the "Safety" we seek is grounded in mathematical rigor.

8 Limitations

While LSE establishes a new benchmark for fidelity in legislative translation, we acknowledge certain limitations in its current instantiation, which we frame as opportunities for future development.

Inference Latency vs. Verification Necessity. The hierarchical nature of the ASP pipeline (three sequential rounds) inevitably incurs higher inference latency (approx. $3.5\times$) compared to single-pass generation. However, we argue that in legislative translation, **precision is non-negotiable**, whereas real-time speed is secondary. The high societal cost of mistranslation justifies this computational overhead.

Dependence on External Knowledge Modules. Critics might view our reliance on TermBases and Rule Tables as a regression to manual engineering. However, we clarify that these resources are **automatically constructed** via unsupervised extraction algorithms. Due to space constraints, the full details of these extensive data mining pipelines are not exhaustively presented here. Crucially, this decoupled design is intentional: it serves as the foundational architecture for our developing DEEPTRANS$^{\text{Studio}}$. This human-interactive workbench extends the fully automated LSE pipeline into a "human-in-the-loop" system, allowing professional translators to intervene in the three control layers, thereby combining automated rigor with human professional agency.

The "Stiffness" of Constrained Output. The *Syntactic Shaping* layer (L2) can occasionally produce text that feels structurally rigid (the "Cost of Rigidity"). While the *Discourse Polishing* layer (L3) mitigates this, there remains a delicate trade-off between strict normative isomorphism and target-language naturalness. Future iterations will employ Direct Preference Optimization (DPO) to better harmonize these objectives without relaxing legal constraints.

9 Ethical Considerations

The deployment of LSE in legislative translation introduces critical ethical responsibilities. We address these through three primary lenses:

Deontic Bias and Rights. As identified in our study, LLMs prone to "Deontic Drift" may inadvertently dilute strict legal obligations (e.g., Chinese *yīngdāng*) into softer recommendations (e.g., English *should*). In legal contexts, such stylistic shifts are not merely linguistic errors but ethical risks, as they may lead to the misinterpretation of mandatory rights and duties, potentially disadvantaging underrepresented groups who rely on the precise enforcement of statutory protections.

Accountability and the Non-Substitution Principle. LSE is designed as a *human-in-the-loop* assistant for professional jurisconsults and legislative drafters. We explicitly emphasize that AI-generated translations, regardless of their fidelity scores, cannot claim legal authority or sovereign force. The responsibility for the final text must reside with qualified human legal experts. We advocate for a regulatory framework where the developer ensures transparency of the "Legal Locks," while the legal practitioner remains the ultimate arbiter of legal intent.

Algorithmic Transparency vs. Data Bias. While LSE implements vertical control to mitigate the "stochastic parrot" effect, the underlying models are trained on general-scale data and may inherit historical biases present in legal corpora. To promote equitable legal AI, we advocate for periodic bias audits of the TermBase (Layer 1) and the RAG-enhanced Discourse Layer (L3) to ensure that LSE does not propagate archaic or discriminatory legal terminology from colonial-era or outdated jurisprudential records.

References

A. A. M. AlSaeed and M. M. Abdulwahab. 2023. Functional equivalence in legal translation: Legal contracts as a case study. *Global Journal of Politics and Law Research*, 11(3):72–150.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, and 33 others. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Farid Ariai, Joel Mackenzie, and Gianluca Demartini. 2025. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *ACM Comput. Surv.*, 58(6).

J. L. Austin. 1975. *How to Do Things With Words*, 2nd edition. Harvard University Press, Cambridge, MA.

Rachelle Ballesteros-Lintao, Maria Regina P. Arriero, Judith Ma. Angelica S. Claustro, Kristina Isabelle U. Dichoso, Selenne Anne S. Leynes, Maria Rosario R. Aranda, and Jean Reintegrado-Celino. 2016. Deontic meanings in philippine contracts. *International Journal of Legal Discourse*, 1:421 – 454.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Ł. Biel, J. Engberg, R.M. Ruano, and V. Sosoni. 2019. *Research Methods in Legal Translation and Interpreting: Crossing Methodological Boundaries*. ISSN. Taylor & Francis.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

D. Cao. 2007. *Translating Law*. Topics in Translation. Multilingual Matters.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904. Association for Computational Linguistics.

Eric Chamoun, Nedjma Ousidhoum, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2025. Social good or scientific curiosity? uncovering the research framing behind NLP artefacts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25310–25346, Suzhou, China. Association for Computational Linguistics.

Clara Ho-yan Chan. 2011. The use and translation of chinese legal terminology in the property laws of mainland china and hong kong: Problems, strategies and future development. *Terminology*, 17(2):249–273.

Le Cheng and King Kui Sin. 2011. A sociosemiotic interpretation of linguistic modality in legal settings. *Semiotica*, 2011(185):123–146.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Edgar Dobriban. 2025. Statistical methods in generative ai. *Preprint*, arXiv:2509.07054.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Giuliana Garzone. 2013. Variation in the use of modality in legislative texts: Focus on shall. *Journal of Pragmatics*, 57:68–81.

Saibo Geng, Rémi Lebret, and Karl Aberer. 2021. Legal transformer models may not always help. *arXiv preprint arXiv:2109.06862*.

Nicholas Godfrey and Mark Burdon. 2024. Fidelity in legal coding: applying legal translation frameworks to address interpretive challenges. *Information & Communications Technology Law*, 33:153 – 176.

Mingyu Gong, Winnie Cheng, and Le Cheng. 2020. Development of deontic modality in chinese civil laws: A corpus study. *Pragmatics and Society*, 11(3):337–362.

M. Graziadei. 2025. Legal translation and the quest for authenticity. *Int J Semiot Law*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11179–11189.

Michael Alexander Kirkwood Halliday. 1994. *An Introduction to Functional Grammar*, 2nd edition. Edward Arnold, London.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinghao Hu, Leilei Gan, Wenyi Xiao, Kun Kuang, and Fei Wu. 2025. Fine-tuning large language models for improving factuality in legal question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4410–4427, Abu Dhabi, UAE. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Philipp Koehn. 2010. *Statistical machine translation*. Cambridge University Press.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: communicative agents for "mind" exploration of large scale language model society. *CoRR*, abs/2303.17760.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Chenyang Lyu, Minghao Wu, and Alham Fikri Aji. 2024. Beyond probabilities: Unveiling the misalignment in evaluating large language models. *CoRR*, abs/2402.13887.

Benoit B. Mandelbrot. 1961. On the theory of word frequencies and on related markovian models of discourse. *Structure of Language and its Mathematical Aspects*, 12:190–219.

Lingyi Meng, Maolin Liu, Hao Wang, Yilan Cheng, Qi Yang, and Idlkaid Mohanmmed. 2025. Building from scratch: a multi-agent framework with human-in-the-loop for multilingual legal terminology mapping. *Artificial Intelligence and Law*.

Joel Niklaus, Jakob Merane, Luka Nenadic, Sina Ahmadi, Yingqiang Gao, Cyrill A. H. Chevalley, Claude Humbel, Christophe Gösken, Lorenzo Tanzi, Thomas Lüthi, Stefan Palombo, Spencer Poff, Boling Yang, Nan Wu, Matthew Guillod, Robin Mamié, Daniel Brunner, Julio Pereyra, and Niko Grupen. 2025. SwiLTra-bench: The Swiss legal translation benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14894–14916, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Frank Robert Palmer. 2001. *Mood and Modality*, 2nd edition. Cambridge University Press, Cambridge.

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

Abudurexiti Reheman, Hongyu Liu, Junhao Ruan, Abudukeyumu Abudula, Yingfeng Luo, Tong Xiao, and JingBo Zhu. 2025. Enhancing neural machine translation through target language data: A kNN-LM approach for domain adaptation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10053–10065, Vienna, Austria. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Abhilasha Sancheti, Aparna Garimella, Balaji Vasan Srinivasan, and Rachel Rudinger. 2022. Agent-specific deontic modality detection in legal language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11563–11579, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Susan Šarčević. 2016. Basic principles of term formation in the multilingual and multicultural context of eu law. In *Language and Culture in EU Law*, pages 183–206. Routledge.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint*, abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *CoRR*, abs/2307.03025.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George F. Foster, and Gholamreza Haffari. 2024a. Adapting large language models for document-level machine translation. *CoRR*, abs/2401.06468.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024b. Importance-aware data augmentation for document-level neural machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–752, St. Julian's, Malta. Association for Computational Linguistics.

Minghao Wu, Jiahao Xu, and Longyue Wang. 2024c. TransAgents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141, Miami, Florida, USA. Association for Computational Linguistics.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Preprint*, arXiv:2405.11804.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *CoRR*, abs/2309.11674.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Ran Zhang, Wei Zhao, and Steffen Eger. 2025a. How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10961–10988, Albuquerque, New Mexico. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025b. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

S. Šarčević. 1997. *New Approach to Legal Translation*. Kluwer Law Internat.

Susan Šarčević. 2000. Legal translation and translation theory: A receiver-oriented approach. In *International Colloquium,'Legal translation, theory/ies, and practice', University of Geneva*, pages 17–19.

A Datasheet for Datasets

A.1 Motivation

Purpose To rigorously evaluate LLMs in high-stakes legislative translation, specifically addressing "Deontic Drift," and to diagnose structural asymmetries across legal systems.

Creation Curated by the authors. Unlike static web crawls, this benchmark involves manual synchronization with legislative amendments effective as of late 2025.

B Dataset Composition and Diversity

To avoid redundancy while addressing concerns regarding domain generalization, we provide a categorical breakdown of the TRILINGUAL LEGISLATIVE BENCHMARK used in this study.

As shown in Figure 4, the corpus is not limited to the Civil Code but maintains a balanced distribution across four major legal domains:

- **Fundamental (14.3%):** The "Constitutional Core" ensuring basic rights.
- **Commercial & IP (34.3%):** High-variance domains testing the model's handling of conditional rights (e.g., Patent Law).
- **Administrative (34.3%):** Procedural laws focusing on compliance (e.g., Data Security Law).
- **Social & Labor (17.1%):** Protective statutes focusing on obligations.

This structural diversity ensures that the reported statistical significance ($p < 0.001$) reflects robust generalization rather than domain overfitting.

B.1 Collection and Quality Control

Sources (1) **Zh-En**: Official government legislative databases, manually updated to reflect 2025 amendments. (2) **Zh-Ja**: Primary sources include the Japan External Trade Organization (JETRO), recognized for authoritative legal translation.

Curation Process We implemented a "Human-in-the-Loop" pipeline. Bilingual legal experts performed secondary verification to correct version lags (synchronizing old translations with new amendments) and unify terminological standards across the 35 statutes.

Preprocessing Texts underwent sentence-level alignment via heuristic matching followed by manual correction. Metadata and formatting artifacts were removed.

B.2 Distribution and License

IP Rights The underlying statutory texts are sourced from Public Domain or Open Government portals[1]. For the Chinese-Japanese portion, while initial references include third-party translations (e.g., JETRO[2]), the dataset provided herein consists of expert-verified secondary rectifications and original sentence-level alignments. These annotation layers, which correct deontic drift and unify terminological standards, constitute the authors' intellectual property and are released for academic research under a Fair Use doctrine for linguistic analysis.

Disclaimer This dataset is for computational research only and should not serve as a substitute for official legal counsel.

License **CC-BY-NC 4.0** (Attribution-NonCommercial).

URL will be released upon acceptance.

C Legal and Linguistic Context of Target Regions

In this section, we provide a granular breakdown of the specific legal frameworks and linguistic typologies characterizing the five studied jurisdictions. These factors serve as the environmental variables in our analysis of the "Source Anomaly" (China) and the "Germanic Lineage" (Germany, Japan).

Table 6 outlines the macroscopic features of each legal system, while Table 4 provides the quantitative Zipf-Mandelbrot parameters derived from our corpus analysis.

D Computational Cost Analysis

To assess the economic viability of the LSE framework, we recorded the total token consumption and API costs for reproducing the full experimental suite (including the 500-sample test set and ablation studies).

As shown in Table 5, the choice of backbone model significantly impacts deployment feasibility.

[1] http://www.npc.gov.cn

[2] https://www.jetro.go.jp/world/asia/cn/ip/law/

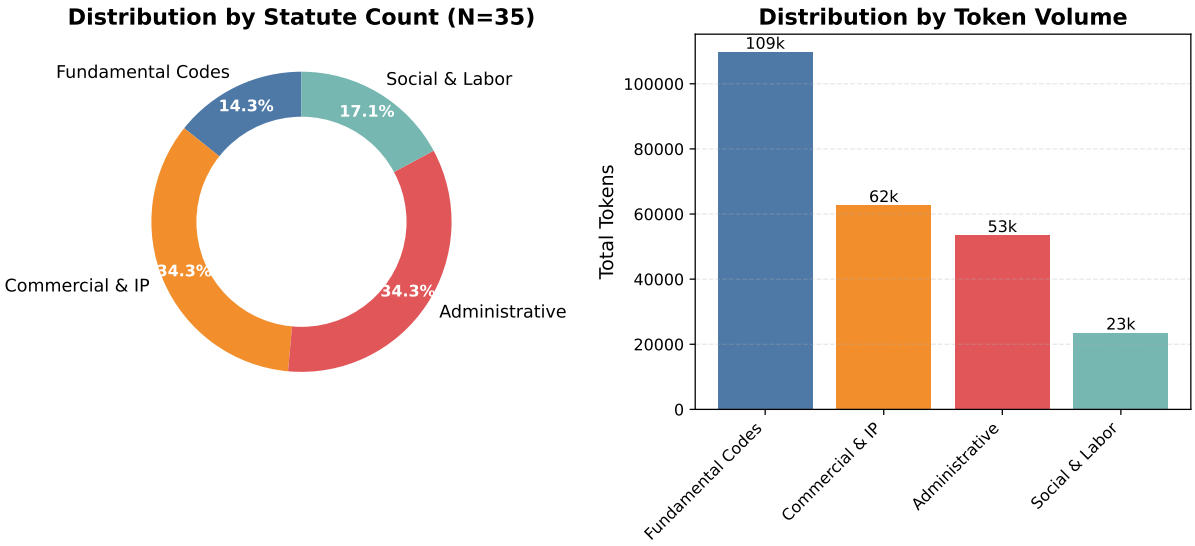


Figure 4: **Benchmark Diversity Analysis.** While *Fundamental Codes* (e.g., Civil Code) constitute the volume-heavy core (Right), the dataset explicitly incorporates a high number of *Commercial* and *Administrative* statutes (Left) to ensure the LSE framework generalizes across diverse normative registers.

Phylogenetic Cluster	**Jur.**	α **(Scale)**	β **(Slope)**	R^2 **(Fit)**
Source Anomaly (High-Mandate)				
China		2835.89	**3.341**	0.939
Germanic Lineage (Precision Civil Law)				
Germany		2185.52	2.350	0.932
Japan		1301.39	2.376	0.941
Anglophone Target (Baselines)				
Québec		2630.49	2.167	**0.971**
Philippines		3168.29	**2.057**	0.940

Table 4: **Zipf-Mandelbrot parameters** ($f = \alpha/r^{\beta}$) across the pentagonal benchmark. The data reveals strict alignment within the Germanic lineage ($\Delta\beta < 0.03$) and a massive structural gap between the Source (CN) and Target (QC/PH) manifolds.

Backbone Model	**Cost**	**Tokens**	**Efficiency**
GEMINI-2.5-PRO	$55.80	5.69 M	High Perf. / High Cost
GPT-4O	$25.41	9.37 M	Balanced
DEEPSEEK-V3	**$11.80**	**24.92 M**	**High Efficiency**
QWEN3-8B	$4.23	21.11 M	Low Cost

Table 5: **Computational Cost and Token Consumption Profile.** Comparison of total expenses for the full evaluation pipeline.

While critics argue that multi-agent hierarchies increase costs, our deployment with DEEPSEEK-V3 processed nearly 2.6× more tokens than GPT-4O while costing less than half the price.

E Qualitative Analysis of Translation Divergence

To complement the quantitative metrics, we present specific examples of terminological and syntactic divergence. These examples illustrate the "Deontic Drift" where legal force is altered during translation (Table 7), and the lexical inconsistencies encountered in key legal concepts (Table 8).

Figure 5 further visualizes the deontic modality distribution mentioned in the main text, highlighting the divergence in how "obligation" is encoded across different legal traditions.

F Implementation and Visualization

To operationalize the LSE framework and facilitate reproducibility, we implemented the system as a modular Python-based library with a decoupled visualization frontend. The source code is available at `https://anonymous.4open.science/r/deeptrans-studio-EFE6`.

Motivation for DeepTrans Studio: Breaking the Walled Garden. Current commercial Computer-Assisted Translation (CAT) tools (e.g., SDL Trados, MemoQ) operate as closed-source walled gardens. They typically rely on rigid translation memory (TM) algorithms and offer limited extensibility for integrating modern Large Language Models (LLMs). Furthermore, their high licensing costs create significant barriers for academic research.

To democratize access to high-stakes translation technology, we developed **DeepTrans Studio**, an

Figure 5: **Comparative analysis of deontic modality profiles. (A) Normalized Usage:** Frequency of modal verbs per 10,000 words (log scale). **(B) Modality Profile:** Row-normalized heatmap showing the proportional share of each modal verb. The analysis highlights the structural divergence between *shall*-centric traditions (e.g., China) and *may*-centric traditions (e.g., Germany).

Region	Legal System	Linguistic Typology	Impact on Corpus & Translation
China	**Civil Law** (Socialist)	**Isolating** (Sino-Tibetan)	High semantic density with rigid statutory terminology. Lack of inflection results in a steeper word frequency drop-off (High β).
Japan	**Civil Law** (German-influenced)	**Agglutinative** (Japonic)	High-context legal culture. Complex scripts (Kanji, Kana) and honorifics create a moderate vocabulary curve.
Germany	**Civil Law** (BGB - Codified)	**Fusional** (Indo-European)	Extreme precision and nominal compounding (e.g., *Rechtsschutz...*) generate a heavy "long tail" of low-frequency terms.
Philippines	**Mixed System**	**Agglutinative** (Austronesian)	Bilingual environment (English/Filipino) with code-switching leads to a vast array of unique tokens (Low β).
Québec	**Bijuralism**	**Fusional** (Indo-European)	Strict linguistic protectionism results in a highly standardized lexicon and the most stable statistical fit ($R^2 \approx 0.97$).

Table 6: **Comparative Matrix of Legal Systems and Linguistic Typologies.** This table summarizes the diverse environmental constraints acting on the translation models.

Jur.	Source Art.	Legal Text (Excerpt)	Deontic Marker
***Case 1**: Principle of Good Faith*			
CN	Art. 7	"... a person of the civil law **shall**, in compliance with..."	*shall* (Imperative)
QC	Art. 6	"Every person **is bound to** exercise his civil rights..."	*is bound to* (Obligational)
DE	Sec. 242	"An obligor **has a duty to** perform according to..."	*has a duty to* (Descriptive)
***Case 2**: Age of Majority*			
CN	Art. 17	"A natural person aged 18 or above **is** an adult."	*is* (Static Definition)
QC	Art. 153	"Full age or the age of majority **is** 18 years."	*is* (Static Definition)
DE	Sec. 2	"Majority **begins** at the age of eighteen."	*begins* (Temporal Process)

Table 7: **Qualitative Analysis of Deontic Drift.** Comparison of modal markers in semantically equivalent provisions. Note how Germany's temporal framing (*begins*) contrasts with static definitions (*is*) in China and Québec.

Concept	Jur.	Term Employed (English)	Context Match
Legal Capacity	**CN**	Capacity for enjoying civil-law rights	Functional description vs. abstract concept.
	DE	Legal capacity	Direct equivalent.
	QC	Juridical personality	distinct Civil Law abstraction.
Force Majeure	**CN**	Force majeure	Standard international term.
	DE	Force majeure	Standard international term.
	QC	Superior force	Linguistic protectionism (Calque).

Table 8: **Terminological Divergence.** Key legal concepts exhibit distinct lexical realizations in official English translations, driven by local statutory requirements.

open-source, human-in-the-loop workbench. Unlike traditional CAT tools, DeepTrans Studio is fully programmable, allowing researchers to visualize and intervene in the multi-agent decision process (as shown in Figure 6), effectively transforming the translator from a text editor to a "semantic engineer."

System Architecture: Engine vs. Interface The architecture consists of two distinct components to balance scientific rigor with practical usability:

- **LSE-Core (The Engine):** A fully automated, headless Python multi-agent system. This module executes the cascading *Anchoring-Shaping-Polishing* pipeline in batch mode. Crucially, all experimental results reported in Section 5 (Table 1 & 3) were generated using this engine without any human intervention.
- **DeepTrans Studio (The Workbench):** A web-based visualization frontend that wraps the LSE-Core. It provides a "Human-in-the-Loop" interface for legal experts to inspect intermediate agent outputs.

Operational Modes Based on this decoupled architecture, the system supports dual-mode operation:

1. **Automated Batch Mode (Experimental):** Utilized for large-scale evaluation. The Python engine processes the test set autonomously, ensuring fair comparison with baselines.
2. **Interactive Mode (Curation):** Utilized for the dataset construction described in Appendix B. This mode allows experts to intervene in the decision process, ensuring the ground truth data achieves the highest legal standard.

G Human Evaluation Metrics

To ensure rigorous manual assessment, we employed legal experts to evaluate translations based on two critical dimensions: **Fidelity** and **Fluency**. Specifically, the *Fidelity* metric rigorously assesses both terminological precision and deontic equivalence (legal force). Meanwhile, the *Fluency* metric evaluates the linguistic register and discourse smoothness. The detailed scoring guidelines provided to the annotators are presented in Table 9.

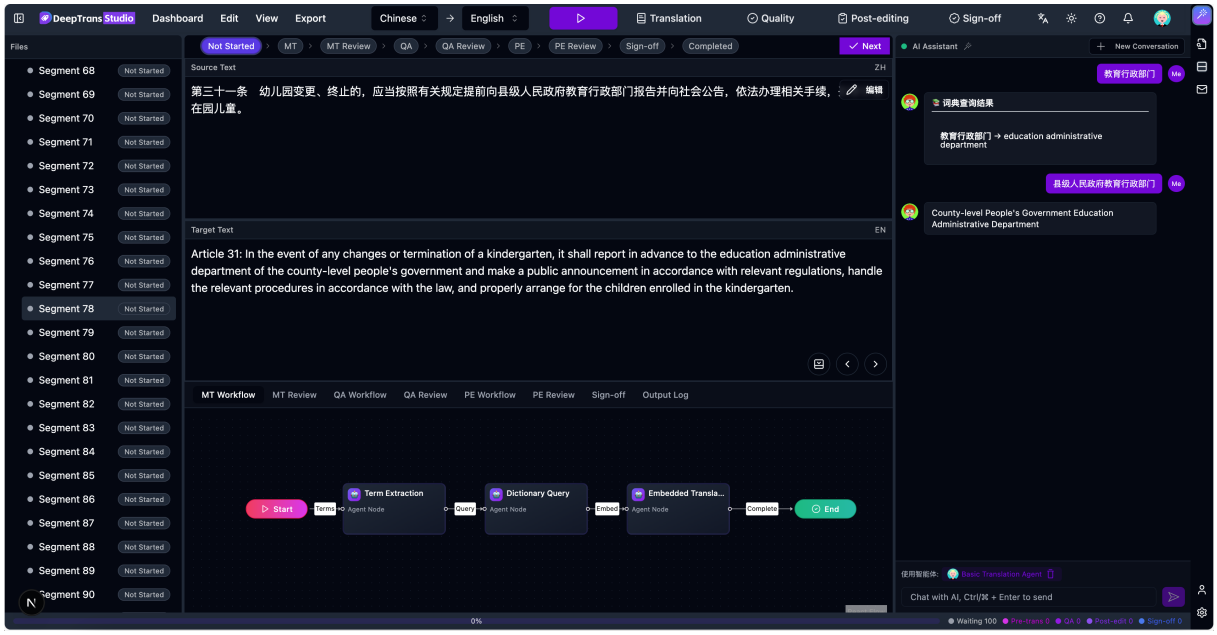


Figure 6: **Interface of DeepTrans Studio (Interactive Mode).** This visualization exposes the internal reasoning of the automated LSE engine. The workflow maps the theoretical layers to the workbench stages: (A) **MT Workflow (≈ Layer 1 Anchoring):** Displays immutable legal entities (e.g., "Education Administrative Department") locked against the external terminology base. (B) **QA Workflow (≈ Layer 2 Shaping):** Visualizes the enforcement of the Deontic Protocol, where the system autonomously rejects probability-based errors (e.g., "should") in favor of normative validity ("shall"). (C) **PE Workflow (≈ Layer 3 Polishing):** Shows the final discourse smoothing. It is worth noting that the entire system was implemented by **a single developer** in three months. *Note: While this interface supports manual intervention, the main experimental results were generated using the underlying Python engine in fully automated mode.*

Dimension	Score	Criteria
1. Fidelity *Core: Accuracy & Legal Force*	5	**Perfect:** Precise legal terminology (e.g., "Force Majeure"). Strict deontic force maintained (e.g., rigid distinction of *shall/may*). Identical normative intent.
	4	**High Quality:** Core terms and legal concepts are accurate. Deontic force is generally consistent, though minor stylistic variations exist. No ambiguity in rights/obligations.
	3	**Passable:** Informational meaning is preserved, but lacks professional rigidity. May use weaker modals (e.g., "should" instead of "shall") or generic vocabulary.
	2	**Poor:** Contains terminological errors or "Deontic Drift" (e.g., confusing obligation with permission). Legal validity is compromised.
	1	**Unusable:** Critical mistranslations of key entities. Normative intent is distorted or reversed.
2. Fluency *Core: Register & Grammar*	5	**Native:** Natural, authoritative legislative style. Grammatically perfect with sophisticated sentence structures.
	4	**Fluent:** Smooth and grammatical. Reads like a professional translation, though isolated phrasings may be slightly stiff.
	3	**Readable:** Understandable but exhibits obvious "translationese". Sentence structures are loose or repetitive.
	2	**Awkward:** Frequent grammatical errors or awkward phrasing that hinders reading flow. Chaotic structure.
	1	**Unreadable:** Disjointed, unintelligible, or nonsensical output.

Table 9: **Human Evaluation Guidelines.** The scoring scale (1-5) covers the two critical dimensions of legislative translation: **Fidelity** (assessing both terminological anchoring and normative equivalence) and **Fluency** (assessing linguistic register).