CLASE: A Hybrid Method for Chinese Legalese Stylistic Evaluation

Yiran Rex Ma[†1,2], **Yuxiao Ye**[†3], **Huiyuan Xie**[3]
[1]School of Foreign Languages, Peking University
[2]Center for Digital Humanities, Peking University
[3]Department of Computer Science and Technology, Tsinghua University
Beijing, China
yiranrexma@outlook.com

Abstract

Legal text generated by large language models (LLMs) can usually achieve reasonable factual accuracy, but it frequently fails to adhere to the specialised stylistic norms and linguistic conventions of legal writing. In order to improve stylistic quality, a crucial first step is to establish a reliable evaluation method. However, having legal experts manually develop such a metric is impractical, as the implicit stylistic requirements in legal writing practice are difficult to formalise into explicit rubrics. Meanwhile, existing automatic evaluation methods also fall short: reference-based metrics conflate semantic accuracy with stylistic fidelity, and LLM-as-a-judge evaluations suffer from opacity and inconsistency. To address these challenges, we introduce ***CLASE*** (**C**hinese **L**eg**A**lese **S**tylistic **E**valuation), a hybrid evaluation method that focuses on the stylistic performance of legal text. The method incorporates a hybrid scoring mechanism that combines 1) linguistic feature-based scores and 2) experience-guided LLM-as-a-judge scores. Both the feature coefficients and the LLM scoring experiences are learned from contrastive pairs of authentic legal documents and their LLM-restored counterparts. This hybrid design captures both surface-level features and implicit stylistic norms in a transparent, reference-free manner. Experiments on 200 Chinese legal documents show that CLASE achieves substantially higher alignment with human judgments than traditional metrics and pure LLM-as-a-judge methods. Beyond improved alignment, CLASE provides interpretable score breakdowns and suggestions for improvements, offering a scalable and practical solution for professional stylistic evaluation in legal text generation*.

Keywords: Legal text generation, Stylistic evaluation, Hybrid evaluation, LLM-as-a-judge

1. Introduction

Large Language Models (LLMs) have significantly advanced legal content generation, with advances on legal reasoning tasks and bar examinations (Katz et al., 2024). These developments have generated considerable interest in automating legal document production, ranging from contract drafting to judicial decision writing. However, while LLMs excel at generating factually accurate and logically coherent legal content, they consistently struggle with the specialized stylistic conventions ("legalese") that characterize professional legal discourse (Tiersma, 1999; Court Writing Committee, 2010).

Legal writing demands adherence to established stylistic norms that extend far beyond semantic correctness. Professional legal documents must conform to specific collocation patterns, formal register requirements, and domain-specific linguistic conventions that signal authority and credibility within the legal community and facing the general public (Foley, 2002; Li and Wang, 2021; Lu and Yuan, 2021; Sun and Cheng, 2017; Li, 2022).

LLM-generated legal documents exhibit two primary categories of stylistic deficiencies: *insufficient sophistication*, failing to meet the expectations through inappropriate colloquialisms or non-standard term choices; and *excessive stylistic elaboration*, where models artificially create a formal impression through verbose, unnecessarily complex, or archaic constructions. This overcompensation often results in hallucinated legal con-

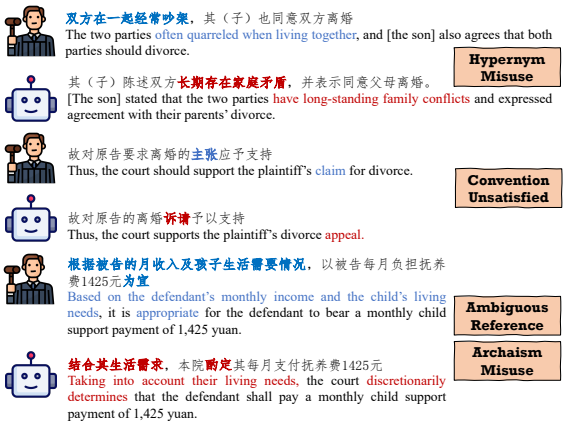

Figure 1: Comparison of Authentic Legal Writing and LLM-Generated Counterpart.

[†]Equal contribution.

*Code and data for CLASE is available at: https://github.com/rexera/CLASE.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 642–653
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

cepts. These dual deficiencies risk undermining professional acceptability (see Figure 1).

Current evaluation for legal text generation focuses on **reasoning capabilities** while neglecting **stylistic quality assessment** (Chalkidis et al., 2022; Xiao et al., 2018; Zhong et al., 2020). This assumes that semantic accuracy, represented by factual correctness, logical argumentation accuracy, and legal knowledge demonstration, constitutes the primary criterion for legal text quality, whereas stylistic appropriateness is just as important in legal competence.

Stylistic quality in legal writing encompasses multiple intricate dimensions that traditional evaluation approaches fail to capture. These include: 1) precise diction and lexical choice; 2) adherence to linguistic/legal conventions; 3) appropriate lexical collocation patterns; and 4) domain-specific common sense and world knowledge (Foley, 2002). These elements collectively contribute to what legal practitioners recognize as "professional" or "sophisticated" writing style, yet they remain largely underexplored in current automated evaluation systems.

Traditional natural language generation (NLG) metrics could be inadequate for capturing stylistic nuances in legal text. Both branches of n-gram and embedding-based methods conflate semantic similarity with stylistic appropriateness (Mellish and Dale, 1998; Papineni et al., 2002; Zhang et al., 2020). These metrics may award high scores to texts that preserve semantic content while exhibiting stylistic violations that would be immediately apparent to legal professionals. Recent advances in LLM-as-a-Judge evaluation (Chan et al., 2023) offer intuitive solutions through instruction understanding, but 1) suffer from limited interpretability (Wang et al., 2023); 2) exhibit biases and consistency issues that undermine their reliability.

The challenge is compounded by the implicit nature of legal stylistic expertise. Writing appropriateness relies on tacit professional knowledge that resists explicit formalization. Legal professionals develop stylistic intuition through years of practice and exposure to exemplary texts, making it difficult to translate this expertise into comprehensive evaluation rubrics. Manual annotation by legal experts is prohibitively time-consuming for large-scale evaluation needs.

To address these gaps, we introduce CLASE[1] (**C**hinese **L**eg**A**lese **S**tylistic **E**valuation), a hybrid evaluation framework designed specifically for assessing stylistic fidelity in legal text generation. Our approach combines objective linguistic feature analysis with experience-guided LLM evaluation, addressing the limitations of existing methods while maintaining interpretability and reference-free operation. The framework employs contrastive learning using authentic legal documents and their stylistically-restored counterparts, enabling automatic acquisition of evaluation criteria that reflect actual professional expectations without requiring manual annotation. CLASE operates in three phases: 1) automated synthesization of contrastive training pairs; 2) training-free contrastive learning that builds experience pools; and 3) hybrid scoring combining linguistically-grounded objective measures with subjectively-informed LLM assessment.

Our primary contributions include:

- **A novel hybrid evaluation architecture** that addresses the neglected dimension of stylistic quality in legal text generation, combining objective linguistic feature analysis with experience-guided LLM evaluation to provide comprehensive stylistic assessment while maintaining interpretability and reference-free operation.
- **A contrastive learning framework** that eliminates expensive manual annotation requirements while ensuring alignment with professional standards, employing authentic legal documents and their deliberately stylistically-degraded counterparts to automatically acquire evaluation criteria that capture actual model defects and reflect actual professional expectations.
- **Empirical validation** demonstrating improved correlation with human expert judgments compared to traditional metrics and pure LLM-based approaches, with interpretable, actionable natural language feedback with improvement strategies.

CLASE addresses a gap in domain-specific, stylistics-oriented evaluation, providing a scalable and transparent solution for professional stylistic assessment. The core principles can extend to other domains requiring specialized stylistic conformity, offering a general approach to stylistic evaluation in professional text generation.

2. Related Work

2.1. NLG Evaluation

NLG evaluation operates across three primary paradigms: *n-gram based lexical metrics*, *neural embedding approaches*, and *LLM-as-a-Judge systems*. 1) Reference-based, n-gram lexical metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie,

[1]With inspiration from Spanish "*con clase*" (having class).

2005) measure lexical similarity between generated text and human-written references, but these approaches face challenges in capturing semantic adequacy and stylistic appropriateness (Reiter and Belz, 2009; Novikova et al., 2017). 2) Embedding-based neural metrics like BERTScore (Zhang et al., 2020) and MoverScore (Zhao et al., 2019) leverage contextual representations to improve correlation, though they measure semantic distance outside of formal dimensions (Li et al., 2024). 3) LLM-as-a-Judge paradigm (Liu et al., 2023; Chan et al., 2023) offers reference-free assessment capabilities but introduces challenges including self-enhancement bias, positional bias, and interpretability concerns (Zheng et al., 2023; Jiang et al., 2024). Other reference-free evaluation methods have emerged to address reference scarcity (Ito et al., 2025), employing techniques such as learning from human judgments through regression models (Rei et al., 2021), similarity-based approaches, and pseudo-rating methods. However, they often require substantial training and face challenges in ensuring consistent evaluation criteria.

2.2. Legal Text Generation and Evaluation

Early research focused on adapting general-purpose models through fine-tuning on legal corpora (Chalkidis et al., 2020), leading to dedicated legal LLMs including LawGPT (Zhou et al., 2024), ChatLaw (Cui et al., 2023), and Lawyer-LLaMA (Huang et al., 2023). These models demonstrate improvements on legal tasks compared to general models, with applications spanning document summarization (Deroy et al., 2023), case analysis (Savelka et al., 2023), and legal question answering (Zhong et al., 2020). As for benchmarking, LawBench (Fei et al., 2024) provides multi-dimensional assessment across memorization, understanding, and application levels. LegalEval-Q (yunhan and gengshen, 2025) focuses on clarity, coherence, and terminology. Chinese legal benchmarks including LAiW (Dai et al., 2025), UCL-Bench (Gan et al., 2025), JuDGE (Su et al., 2025), and CaseGen (Li et al., 2025) offer evaluation covering various capability levels and practical applications. Gap analysis research (Hou et al., 2024; Ma, 2025) has identified issues in LLM-generated reasoning/analysis, highlighting the need for fine-grained evaluation, such as hybrid approaches combining automated metrics with expert human judgment (Guha et al., 2023; Shao et al., 2025).

2.3. Style and Stylistic Evaluation

Prevalent perspective in computational linguistics outlines "style" broadly as extents of formality, politeness, simplicity, personality, emotion, etc. (Jin et al., 2022), focusing on style transfer (periods, genre, authors...), authorship attribution/stylistic fingerprints of LLMs (Bitton et al., 2025), and stylometric analysis (Juola et al., 2008; Argamon et al., 2007). Recent advances in content-independent style embeddings address content leakage challenges. StyleDistance (Patel et al., 2025) employs LLM-synthetic contrastive parallel text to create controlled stylistic variations to train separate embeddings, which gives us crucial insights into addressing stylistic quality assessment. Our work focuses on stylistic *quality* assessment, which differs from classical style transfer/analysis. We assess adherence to domain-specific conventions and professional standards in Chinese legal contexts.

3. Method

CLASE adopts a three-stage approach (see Figure 2): *contrastive pair synthesization*, *training-free contrastive learning*, and *hybrid scoring*. The framework requires no manual annotations while capturing both surface-level linguistic patterns and implicit stylistic norms.

3.1. Contrastive Pair Synthesization

We construct training exemplars from authentic Chinese judgment documents[2], **focusing on the reasoning sections** where stylistic quality most critically impacts professional acceptability. For each original text segment t_{gold}, we generate contrastive learning pairs through a two-stage, prompt-guided transformation: $t_{\text{reverse}} = \pi_1(t_{\text{gold}})$ and $t_{\text{restored}} = \pi_2(t_{\text{reverse}})$, where π_1 performs stylistic degradation by converting legalese to colloquial expression while preserving semantics, named entities, and topic chains, and π_2 attempts restoration from the degraded text back to legalese.

3.2. Training-Free Contrastive Learning

This stage operates through structured steps to accumulate stylistic knowledge without manual annotations. Each learning step $\tau^{(i)}$ processes a contrastive pair $(t^{(i)}_{\text{gold}}, t^{(i)}_{\text{restored}})$ to extract labeled exemplars for regression.

[2]Following professional practice, we list five sections in first-instance civil judgments: 1) header (case and party information), 2) facts (claims and findings), 3) reasoning (dispute analysis and rationale), 4) judgment (legal basis and outcome), and 5) footer (appeal information and signatures). This constitutes "split" in Figure 2.

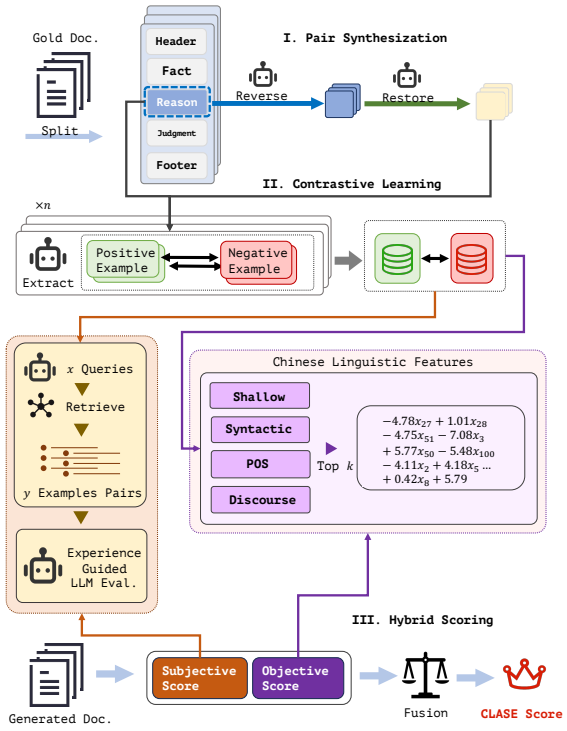

Figure 2: CLASE Overview: (I) contrastive pair synthesization from authentic legal documents; (II) training-free contrastive learning to build positive/negative example pools; (III) hybrid scoring combining objective linguistic features with experience-guided LLM evaluation.

For the i-th learning step $\tau^{(i)}$, we define the input as a single contrastive pair $(t^{(i)}_{\text{gold}}, t^{(i)}_{\text{restored}})$. The goal of each learning step is to identify stylistic issues and extract corresponding positive-negative exemplar pairs. Within each step, we perform guided comparison to identify a set of stylistic problems $\mathcal{I}^{(i)} = \{I^{(i)}_1, I^{(i)}_2, ..., I^{(i)}_{m_i}\}$, where each $I^{(i)}_j$ represents a specific stylistic issue discovered by comparing the gold and restored texts:

$$\mathcal{I}^{(i)} = \pi_{\text{identify}}(t^{(i)}_{\text{gold}}, t^{(i)}_{\text{restored}})$$

where π_{identify} outputs structured issue descriptions. For each identified problem $I^{(i)}_j \in \mathcal{I}^{(i)}$, we extract a positive-negative exemplar pair $(e^{(i,j)}_{\text{pos}}, e^{(i,j)}_{\text{neg}})$ through a prompt-constrained **one-to-one correspondence** $\Theta : e^{(i,j)}_{\text{pos}} \leftrightarrow e^{(i,j)}_{\text{neg}}$, ensuring paired positions in both pools address the same stylistic aspect. The extracted exemplars are accumulated into two separate pools: positive experience pool $\mathcal{P}_{\text{pos}}$ and negative experience pool $\mathcal{P}_{\text{neg}}$.

After completing N learning steps, we perform logistic regression on the accumulated experience pools. Inspired by the method in Qiu et al. (2018)'s Chinese readability assessment, we extract linguistic features $F(e) = \{f_1, f_2, ..., f_{100}\}$ from each exemplar e in both pools, encompassing surface-level characteristics including character complexity, part-of-speech distributions, syntactic patterns, and discourse markers, among others[3]. Features are z-score normalized to ensure commensurate scales before coefficient comparison. The regression model learns to distinguish positive from negative exemplars:

$$P(\text{positive}|F(e)) = \sigma(\mathbf{w}^T F(e) + b)$$

where $\mathbf{w}$ represents learned feature weights and σ is the sigmoid function. This provides a feature-based scoring mechanism that generalizes from the accumulated exemplars to evaluate new texts.

Algorithm 1 Training-Free Contrastive Learning

Input: Document pairs $\{(t^{(i)}_{\text{gold}}, t^{(i)}_{\text{restored}})\}_{i=1}^{N}$
Output: Experience pools $\mathcal{P}_{\text{pos}}$, $\mathcal{P}_{\text{neg}}$, coefficients $\mathbf{w}$

$\mathcal{P}_{\text{pos}} \leftarrow \emptyset$, $\mathcal{P}_{\text{neg}} \leftarrow \emptyset$
for $i = 1$ to N **do**
  $\mathcal{I}^{(i)} \leftarrow \pi_{\text{identify}}(t^{(i)}_{\text{gold}}, t^{(i)}_{\text{restored}})$
  for $j = 1$ to $|\mathcal{I}^{(i)}|$ **do**
    $(e^{(i,j)}_{\text{pos}}, e^{(i,j)}_{\text{neg}}) \leftarrow \Phi(I^{(i)}_j)$
    $\mathcal{P}_{\text{pos}} \leftarrow \mathcal{P}_{\text{pos}} \cup \{e^{(i,j)}_{\text{pos}}\}$, $\mathcal{P}_{\text{neg}} \leftarrow \mathcal{P}_{\text{neg}} \cup \{e^{(i,j)}_{\text{neg}}\}$
    {Θ correspondence}
  end for
end for
$X \leftarrow$ [FeatureExtract(e) for $e \in \mathcal{P}_{\text{pos}} \cup \mathcal{P}_{\text{neg}}$]
$y \leftarrow$ [1 for $e \in \mathcal{P}_{\text{pos}}$] + [0 for $e \in \mathcal{P}_{\text{neg}}$]
$\mathbf{w} \leftarrow$ LogisticRegression(X, y)
return $\mathcal{P}_{\text{pos}}$, $\mathcal{P}_{\text{neg}}$, $\mathbf{w}$

3.3. Hybrid Scoring

CLASE produces a final score $\Psi(t)$ through hybrid combination of objective (*CLASE Obj*) and subjective (*CLASE Subj*) assessments. Without reference, for each generated text t, it eventually offers a sigmoid-fused, $[0, 1]$ score from both linguistic features and experience-guided LLM judging.

Given input text t, *CLASE Obj* extracts linguistic features $F(t) = \{f_1, f_2, ..., f_{100}\}$ and select the top-k features based on logistic regression coefficient magnitudes. The objective score is computed as:

$$\Psi_{obj}(t) = 10 \times \sigma(\mathbf{w}^T F_k(t))$$

[3]Refer to Qiu et al. (2018) or CLASE repository (https://github.com/rexera/CLASE) for a comprehensive list of features.

where $F_k(t)$ represents the selected k features, **w** contains learned regression weights, and the output is normalized to $[0, 10]$.

CLASE Subj evaluates seven dimensions based on legal writing practice through retrieval-augmented, experience-guided assessment.

We define the dimension set $\mathcal{D} = \{d_1, d_2, ..., d_7\}$ with respective weights: noun usage (30%), verb usage (30%), adjective usage (20%), function words (5%), sentence coherence (5%), sentence structure (5%), and collocations (5%).

For each dimension $d \in \mathcal{D}$, 1) LLM judge π analyzes the input text t and generates x queries focusing on potential stylistic issues within dimension d. 2) for each query, we retrieve top-y negative exemplars from $\mathcal{P}_{\text{neg}}$ and obtain their corresponding positive exemplars through Θ. 3) π score the text in $[0, 10]$ using these contrastive exemplar pairs as contextual guidance.

The final CLASE score combines both components through empirically-calibrated hybrid fusion. We use equal weighting (0.5 each) based on pilot studies showing optimal performance at this balance. The sigmoid transformation ensures output normalization while preserving relative rankings:

$$\Psi(t)' = 0.5 \times \Psi_{obj}(t) + 0.5 \times \Psi_{subj}(t)$$

$$\Psi(t) = \frac{1}{1 + \exp(-10 \times (\Psi(t)'/10 - 0.5))}$$

Algorithm 2 Hybrid Scoring

Input: Text t, experience pools $\mathcal{P}_{\text{pos}}, \mathcal{P}_{\text{neg}}$, weights **w**, dimension set $\mathcal{D}$, k, query count x, retrieval count y
Output: Final CLASE score $\Psi(t)$
// Objective Scoring
$F_k(t) \leftarrow$ ExtractTopKFeatures(t, k)
$\Psi_{obj}(t) \leftarrow 10 \times \sigma(\mathbf{w}^T F_k(t))$
// Subjective Scoring
$\Psi_{subj}(t) \leftarrow 0$
for $d \in \mathcal{D}$ **do**
  $\mathcal{A}_d \leftarrow \pi$.AnalyzeText($t$, d) {π identifies potential issues in dimension d}
  $Q \leftarrow \pi$.GenerateQueries($\mathcal{A}_d$, x)
  $\mathcal{E} \leftarrow \emptyset$
  for $q_i \in Q$ **do**
    $N_{q_i} \leftarrow$ TopKSimilar(q_i, $\mathcal{P}_{\text{neg}}$, y) {Retrieve y similar negatives}
    $P_{q_i} \leftarrow$ GetCorresponding(N_{q_i}, $\mathcal{P}_{\text{pos}}$) {Get paired positives via Θ}
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{(P_{q_i}, N_{q_i})\}$
  end for
  $\Psi_{subj}(t, d) \leftarrow \pi$.Evaluate($t$, $\mathcal{E}$, d)
  $\Psi_{subj}(t) \leftarrow \Psi_{subj}(t) + \beta_d \cdot \Psi_{subj}(t, d)$
end for
$\Psi(t) \leftarrow$ SigmoidFusion($\Psi_{obj}(t)$, $\Psi_{subj}(t)$)
return $\Psi(t)$

4. Experiments

4.1. Experimental Setup

We build, train, and test CLASE on Qwen-2.5 model family (Qwen et al., 2025). **At learning time**, we conduct experiments using 4000 Chinese civil judgment documents (learning step $N = 4000$)[4]. For the contrastive pair synthesization phase, we employ the 7B model for stylistic degradation (π_1), 32B for restoration (π_2), and 72B for experience pool construction.

At test time, we sample 200 additional documents from the same data source, outside the training scope. We follow the same pipeline to generate colloquially reverse versions, then employ GPT-4o (OpenAI, 2024) to create restored versions with controlled variations to simulate different legalese proficiency levels and varied emphasis on legal writing requirements: efficiency, thoroughness, structure, formality, educational[5].

The top-k feature selection uses absolute coefficient values from L2-regularized logistic regression. Text segmentation follows jieba tokenization with character-level boundary detection for span alignment. For the subjective scoring component, we use the 72B model with a naive Chain-of-Thought (CoT) prompting (Wei et al., 2022) as the judge model. For retrievals, we use embeddings of the 7B model with cosine similarity as the distance metric. In the main experiments, we configure the retrieval parameters as query count $x = 10$ and retrieval count $y = 10$, with the objective component using top $k = 25$ features. All experiments are conducted on 8 NVIDIA A800-SXM4-80GB GPUs with vLLM (Kwon et al., 2023) for model inference.

4.2. Evaluation

We recruit two legal domain experts to conduct evaluation for 200 restored documents across the seven aforementioned stylistic dimensions and respective weights. Each expert independently assigns scores from 0-10 for each dimension comparing gold and restored documents. Inter-annotator agreement achieves Krippendorff's alpha of 0.72 (Krippendorff, 2011), indicating reliability. We evaluate system performance using Pearson correlation coefficient r (Pearson, 1895), Spearman rank correlation ρ (Spearman, 1904),

[4] https://wenshu.court.gov.cn

[5] Legal writing serves as not only an instrument for the rule of law, judicial practices, and law enforcement, but also a vital medium for shaping public legal awareness. Detailed implementation is in our repository.

and Kendall's τ (Kendall, 1938).

4.3. Baselines

1) Traditional reference-based metrics include standard n-gram methods (character-level F1, BLEU, ROUGE, METEOR) and embedding-based semantic similarity (BERTScore). The F1 baseline uses character-level exact matching with precision-recall harmonic mean. 2) LLM-as-a-Judge methods employ GPT-4o-mini and Qwen-2.5-72B in both reference-based ("-ref") and reference-free configurations, evaluating across the same seven stylistic dimensions with 0-10 scoring. 3) CLASE variants include subjective-only (CLASE-Subj), objective-only (CLASE-Obj), and hybrid fusion (CLASE-Mix).

5. Results and Analysis

Table 1: Main Correlation Results

Method	r	ρ	τ
ROUGE	0.4250	0.4807	0.3355
BLEU	0.4160	0.3383	0.2306
F1	0.3907	0.3214	0.2213
BERTScore	0.3145	0.2010	0.1378
METEOR	0.3586	0.2949	0.2020
GPT-4o-mini-ref	0.3206	0.3061	0.2083
Qwen2.5-72B-ref	0.3049	0.2482	0.1808
GPT-4o-mini	0.0689	0.0747	0.0528
Qwen2.5-72B	0.2416	0.2336	0.1855
CLASE-Subj	0.3165	0.3063	0.2148
CLASE-Obj	0.7923	0.7568	0.5692
CLASE-Mix	**0.8271**	**0.8109**	**0.6180**

Table 2: Variance and Dispersion Analysis

Method	Std	Variance	CV
CLASE-Mix	**2.021**	**4.086**	**0.365**
Human	1.722	2.965	0.383
Qwen2.5-72B-ref	0.639	0.408	0.085
GPT-4o-mini-ref	0.574	0.330	0.082
CLASE-Subj	0.473	0.223	0.079
GPT-4o-mini	0.473	0.223	0.062
ROUGE	0.363	0.132	0.669
Qwen2.5-72B	0.176	0.031	0.022
CLASE-Obj	0.118	0.014	2.695
METEOR	0.103	0.011	0.231
BLEU	0.088	0.008	0.385
F1	0.072	0.005	0.120
BERTScore	0.035	0.001	0.043

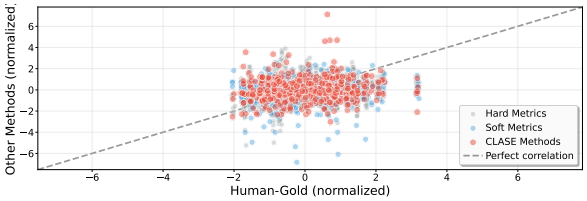

Figure 3: Correlation analysis between evaluation methods and human judgments. Points closer to the diagonal line indicate better alignment with human evaluation.

5.1. Main Results

Correlation As shown in Table 1, 1) hard metrics fail to capture stylistic variation. These metrics focus on surface-level lexical matching or semantic similarity, which is inadequate for capturing stylistic features or differentiating documents based on stylistic variation alone. 2) LLM-as-a-judge methods (soft metrics) have inherent limitations in providing fine-grained numerical scores for style. It is better suited for generating qualitative, natural language feedback and could benefit from the grounding provided by objective measures. 3) Notably, CLASE-Obj achieves strong correlation. This suggests that the contrastive learning approach successfully identifies useful linguistic features that correlate with professional legalese. When fused with CLASE-Subj, it gains incremental improvement for introducing flexible nature of LLMs.

Figure 3 visualizes this alignment through normalized scores, where the x-axis represents standardized human judgments (z-scores) and the y-axis shows standardized scores from each evaluation method. The diagonal line $y = x$ indicates perfect correlation. CLASE effectively cluster closer to the diagonal line compared to hard and soft metrics, indicating superior capability in capturing human-perceived stylistic quality.

Dispersion Table 2 reveals critical evaluation characteristics through the coefficient of variation (CV), which normalizes variability by mean scores to enable fair comparison. Traditional metrics exhibit extremely low CV values, whereas LLM-as-a-Judge methods show relatively higher values yet still more "conservative" than human experts, indicating insufficient sensitivity to stylistic differences. CLASE-Subj performs comparably to or slightly better than LLM-as-a-judge baselines. CLASE-Mix closely matches human expert dispersion, suggesting appropriate discriminative power for stylistic assessment.

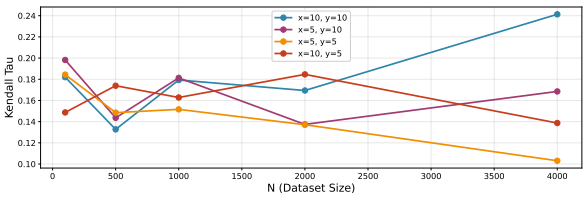

Figure 4: Ablation study on training set size and retrieval parameters for subjective scoring component, measured by Kendall's τ.

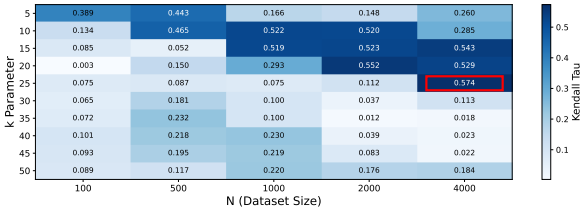

Figure 5: Ablation study on training set size and significant feature count for objective scoring component, measured by Kendall's τ.

Figure 6: Example of CLASE's natural language feedback identifying specific stylistic issues and improvement suggestions.

5.2. Ablation Study

For the subjective score (Figure 4), performance generally improves with a larger training set size (N), more queries per document (x), and more retrieved example pairs per query (y). This shows that a richer context of positive and negative examples helps guide the LLM toward more accurate evaluations. For the objective score (Figure 5), we analyzed the impact of training set size (N) and the number of significant features (k). Correlation peaks at $N = 4000, k = 25$, indicating that a focused set of salient linguistic markers is optimal.

5.3. Case Study: Stylistic Feedback

Due to preceding CoT reasoning, CLASE-Subj provides interpretable natural language reason/feedback that identifies stylistic deficiencies and improvement directions. Figure 6 demonstrates how CLASE offers actionable comments about legal writing quality. The feedback examples represent issues that practitioners consistently identify when reviewing LLM-generated legal documents in actual judicial settings. Through field observations in Chinese courts, we documented frequent stylistic complaints from legal professionals, including:

1) inappropriate lexical choices for legal basis versus factual reasoning (different formal terms must be used when citing statutes versus referencing evidentiary materials); 2) omission of conventional emphatic phrases that signal judicial authority and legal compliance ("reasonable and lawful", "in accordance with the law"); and 3) subtle distinctions between copular constructions that carry different degrees of formality and certainty in legal contexts.

CLASE detects all these precise issues without any annotation of such domain-specific knowledge in the example in Figure 6. The system learns to distinguish nuanced professional conventions purely through contrastive analysis, without explicit knowledge injection. It demonstrates that sophisticated domain-specific stylistic expertise, which traditionally requires years of training, could be **potentially** acquired through automated, contrastive learning, bringing out exciting "intelligence" that pure numerical scores cannot offer.

5.4. Discussion

Generalizability Current validation focuses on Chinese civil judgment reasoning sections. However, CLASE's core contrastive learning principles offer promising extensibility. The framework's reliance on natural language knowledge representation suggests potential generalization advantages—stylistic expertise encoded in natural language exhibits transferability compared to numerical features or domain-specific rules. Extension to other legal domains, broader professional writing contexts, or other languages would require domain-specific experience pool construction while preserving the underlying methodology.

Computational considerations While contrastive learning requires substantial resources for pair generation and experience pool construction, the resulting system operates efficiently in deployment. Once trained, experience pools remain static[6] and retrieval can scale well with pre-computed embeddings and indexing. The framework's computational profile is front-loaded during development rather than inference, making it suitable for production environments where training costs amortize over extensive usage.

Scaling characteristics Our ablation studies reveal complex scaling patterns that challenge simple, "brute force" "scaling law" assumptions. While performance generally improves with training set size up to $N = 4000$, marginal gains diminish beyond certain thresholds. Similarly, optimal feature selection ($k = 25$) suggests that excessive linguistic features introduce noise rather than improving discrimination. These findings indicate that effective deployment requires careful hyperparameter optimization rather than naive scaling strategies.

Interpretability and Hybrid Necessity While CLASE-Obj demonstrates strong performance, reliance on objective features alone is insufficient, especially when we consider evaluation as the reward signal of LLM reinforcement loop. Guiding generation policy with only objective metrics is prone to Goodhart's Law; once specific stylistic markers become explicit optimization targets, models might game the metric without genuine stylistic improvement. Furthermore, numerical feature weights offer limited explanatory power. CLASE-Subj is imperative as it provides: 1) semantic grounding to prevent adversarial overfitting to surface features; and 2) actionable natural language feedback (as shown in Section 5.3), which is essential for human-in-the-loop workflows.

6. Conclusion

We introduce CLASE, a hybrid evaluation framework that addresses the under-explored challenge of stylistic quality assessment in Chinese legal text generation. By combining objective linguistic feature analysis with experience-guided LLM evaluation, our approach provides a reference-free solution that captures both surface-level patterns and implicit professional conventions in Chinese legal writing.

The key insight underlying CLASE is that professional stylistic expertise can be acquired through contrastive analysis rather than explicit annotation. Our training-free contrastive learning framework successfully extracts meaningful stylistic knowledge from authentic legal documents and their deliberately degraded counterparts, eliminating the need for expensive manual score annotations while maintaining alignment with professional standards.

Experimental results on 200 Chinese legal documents demonstrate that CLASE achieves higher correlation with human expert judgments compared to traditional metrics and pure LLM-based approaches. The framework not only provides quantitative scores but also generates interpretable natural language feedback that identifies specific stylistic deficiencies and improvement directions—capabilities that numerical metrics cannot offer.

The work makes three primary contributions to the intersection of natural language evaluation and legal AI: establishing a hybrid architecture that balances objectivity with nuanced judgment, demonstrating that sophisticated domain-specific conventions can emerge from automated contrastive analysis, and providing a scalable framework that extends beyond Chinese legal text to other domains requiring specialized stylistic conformity.

While current validation focuses on civil judgment documents within a specific model family, the core principles of contrastive stylistic learning offer a general approach to professional text evaluation. Future work should explore cross-domain generalization, computational efficiency optimization, and deeper integration of legal writing principles to enhance both performance and interpretability. As legal AI systems become increasingly sophisticated in generating factually accurate content, frameworks like CLASE become essential for ensuring that automated text generation meets the professional standards expected in legal practice.

7. Ethical Considerations

This research primarily involves publicly available legal documents (Chinese civil judgments) and standard large language models. The datasets do not contain private or sensitive individual information that is not already part of the public record. The proposed evaluation framework aims to improve the professional quality of automated legal writing. We anticipate no direct negative societal impacts. However, users should be aware that automated evaluation tools, including CLASE, should serve as assistants to, rather than replacements for, human legal professionals.

[6] The experience pool can be updated periodically to incorporate new domain knowledge and improve performance over time.

8. Limitations

We acknowledge several limitations in the current work. First, our experiments are restricted to Chinese civil judgments. While we believe the contrastive learning methodology is transferrable, the specific implementation of the objective component (using Chinese linguistic features) is language-dependent and would require adaptation for other languages. Second, the contrastive pair synthesis relies on the assumption that the degradation model alters style without corrupting semantics. While our strong correlation results suggest the validity of the generated data for training, we did not perform a large-scale manual audit of the synthetic pairs. Third, compared to lightweight metrics like BLEU, CLASE involves a multi-stage pipeline which demands higher computational resources, potentially limiting immediate adoption in low-latency applications.

While CLASE-Obj demonstrates strong empirical performance, the underlying mechanisms by which logistic regression weights capture professional stylistic preferences remain opaque. CLASE-Subj exhibits sophisticated pattern recognition capabilities but lacks deep causal understanding. It identifies specific stylistic deficiencies without comprehending the underlying professional rationales or historical evolution of legal writing conventions. This "knowing-what but not knowing-why" limitation restricts the framework's ability to provide educational insights or adapt to evolving professional standards. Future work should explore incorporating explicit legal writing principles to achieve more comprehensive stylistic expertise.

9. Acknowledgements

We thank the reviewers for their human touch, dedication, and insightful feedback. This work was supported by the Natural Language Processing Lab at Tsinghua University (TsinghuaNLP).

References

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yehonatan Bitton, Elad Bitton, and Shai Nisan. 2025. Detecting stylistic fingerprints of large language models. *arXiv preprint arXiv:2503.01659*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.

Chi-Min Chan, Weize Chen, Yujia Su, Jiahui Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Court Writing Committee. 2010. Court writing guide. Legal writing guidelines for professional practice.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2025. LAiW: A Chinese legal large language models benchmark. In *Proceedings of the 31st International conference on computational linguistics*, pages 10738–10766.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How ready are pre-trained abstractive models and llms for legal case judgement summarization? *arXiv preprint arXiv:2306.01248*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, et al. 2024. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7933–7962.

Richard Foley. 2002. Language, law and legal writing: An introduction to legal discourse. *Legal Writing: The Journal of the Legal Writing Institute*, 8:1–35.

Ruoli Gan, Duanyu Feng, Chen Zhang, Zhihang Lin, Haochen Jia, Hao Wang, Zhenyang Cai, Lei Cui, Qianqian Xie, Jimin Huang, et al. 2025. Uclbench: A chinese user-centric legal benchmark for large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7945–7988.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.

Abe Bohan Hou, William Jurayj, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2024. Gaps or hallucinations? gazing into machine-generated legal analysis for fine-grained text evaluations.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report.

Takumi Ito, Kees van Deemter, and Jun Suzuki. 2025. Reference-free evaluation metrics for text generation: A survey.

Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex Pentland, Yoon Kim, Deb Roy, et al. 2024. Leveraging large language models for learning complex legal concepts through storytelling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7194–7219.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Patrick Juola et al. 2008. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1(3):233–334.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Klaus Krippendorff. 2011. *Computing Krippendorff's Alpha Reliability*. University of Pennsylvania.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*.

Haitao Li, Jiaying Ye, Yiran Hu, Jia Chen, Qingyao Ai, Yueyue Wu, Junjie Chen, Yifan Chen, Cheng Luo, Quan Zhou, and Yiqun Liu. 2025. Casegen: A benchmark for multi-stage legal case documents generation. *arXiv preprint arXiv:2502.17943*.

Huixian Li. 2022. Lexical, syntactic and textual features of shipping legal documents in chinese and english. *SCIREA Journal of Sociology*, 6(2):83–103.

Shifang Li and Yifan Wang. 2021. A study of cohesion in the chinese legal text: Based on criminal procedure law of the people's republic of china. *Theory and Practice in Language Studies*, 11(12):1709–1716.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. Leveraging large language models for nlg evaluation: A survey. *CoRR*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Nan Lu and Chuanyou Yuan. 2021. Legal reasoning: A textual perspective on common law judicial opinions and chinese judgments. *Text & Talk*, 41(1):71–93.

Yiran Rex Ma. 2025. Do androids question electric sheep? a multi-agent cognitive simulation of philosophical reflection on hybrid table reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 143–164, Vienna, Austria. Association for Computational Linguistics.

Chris Mellish and Robert Dale. 1998. Evaluation in the context of natural language generation. *Computer Speech & Language*, 12(4):349–373.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. 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 8662–8685, Albuquerque, New Mexico. Association for Computational Linguistics.

Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the royal society of London*, 58(347-352):240–242.

Xinying Qiu, Kebin Deng, Likun Qiu, and Xin Wang. 2018. Exploring the Impact of Linguistic Features for Chinese Readability Assessment. In Xuanjing Huang, Jing Jiang, Dongyan Zhao, Yansong Feng, and Yu Hong, editors, *Natural Language Processing and Chinese Computing*, volume 10619, pages 771–783. Springer International Publishing, Cham.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan Van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André FT Martins, and Alon Lavie. 2021. Are references really needed? unbabel-ist 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Jaromir Savelka, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4).

Peizhang Shao, Linrui Xu, Jinxi Wang, Wei Zhou, and Xingyu Wu. 2025. When large language models meet law: Dual-lens taxonomy, technical advances, and ethical governance.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.

Weihang Su, Baoqing Yue, Qingyao Ai, Yiran Hu, Jiaqi Li, Changyue Wang, Kaiyuan Zhang, Yueyue Wu, and Yiqun Liu. 2025. JuDGE: Benchmarking judgment document generation for chinese legal system.

Yuxiu Sun and Le Cheng. 2017. Linguistic variation and legal representation in legislative discourse: A corpus-based multi-dimensional study. *International Journal of Legal Discourse*, 2(2):315–339.

Peter Meijes Tiersma. 1999. *Legal Language*. University of Chicago Press.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. Cail2018: A large-scale legal dataset for judgment prediction.

Li yunhan and Wu gengshen. 2025. LegalEval-Q: A new benchmark for the quality evaluation of llm-generated legal text.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9701–9708.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. Lawgpt: A chinese legal knowledge-enhanced large language model.