

COUNT EVIDENCE, NOT SENTENCES: TEMPERED EVIDENCE FUSION OF LLM JUDGMENTS FOR LONG-TEXT VALUE MEASUREMENT

Yuhe Wu^{1,*}, Rui Qian^{2,*}, Guangyu Wang^{1,*}, Yuran Chen³, Yuanchao Zhu⁴, Junjie Yang⁵
Zhengheng Li⁶, Jiulin Cai⁷, Tianyi Zhang⁸, Zihan Dong⁹, Jiaxin Liu¹, Yujie Chen¹⁰, Guang Zhang^{1,†}

¹HKUST(GZ) ²FDU ³DUFE ⁴UESTC ⁵UMD
⁶SEU ⁷USTC ⁸Independent ⁹Georgia Tech ¹⁰CUHK(SZ)
{yuheewuu, qianruii}@gmail.com, guangzhang@hkust-gz.edu.cn

ABSTRACT

Large language models (LLMs) are increasingly used to measure public value orientations from long social media posts, yet such posts often mix background, quotations, concessions, and only a few stance-bearing sentences. Existing approaches either ask the model to predict a document-level label directly, which can be overconfident, or aggregate sentence-level predictions by majority or soft voting, which treat uncertain and decisive sentences as equally informative. We formulate long-text value measurement as a decision-fusion problem and propose Tempered Evidence Fusion (TEF), a training-free rule that weights each sentence’s log-odds by its normalized information gain, as derived from a generalized Bayesian posterior. This makes the fused score nearly vanish for uncertain sentences while preserving the Bayes-optimal weight of decisive evidence. We further introduce Multi-event Insight Network Dimensions (MIND), a benchmark of 8,358 Chinese and English posts spanning five years of public events and six value dimensions. On MIND, TEF outperforms the strongest baseline among Direct, Majority Vote, and Soft Vote by an average of 4.5 accuracy points and 4.6 macro-F1 points across five LLMs and two languages. MIND dataset and code are available at [GitHub](#).

Index Terms— Decision fusion, large language models, uncertainty, calibration, computational social science

1. INTRODUCTION

Public value orientations are key latent variables in computational social science, capturing how people position themselves on issues such as economic regulation, climate policy, or cultural change. Social media makes these orientations observable through text at scale [1, 2]. Yet this scale, together with the subjectivity of human annotation, has motivated growing interest in using LLMs as measurement instruments [3, 4]. Existing LLM-based pipelines usually either ask the model to read the whole post and return a document-level label [5, 6, 7], or split the post into sentences and aggregate sentence-level predictions by voting [8].

However, we observe that little research has examined how sentence-level LLM judgments should be fused when evidence is unevenly distributed across a long post. In Fig. 1, long social media posts often mix background information, quotations, and concessions, with only a few stance-bearing sentences. In such cases, direct prediction can collapse heterogeneous evidence into one overconfident label, while majority or soft voting can allow uncertain sentences to outvote a few decisive ones. This leads us to ask: *How can sentence-level judgments be combined so that decisive evidence dominates the final decision while uncertain sentences are discounted?*

*Equal contribution †Corresponding author

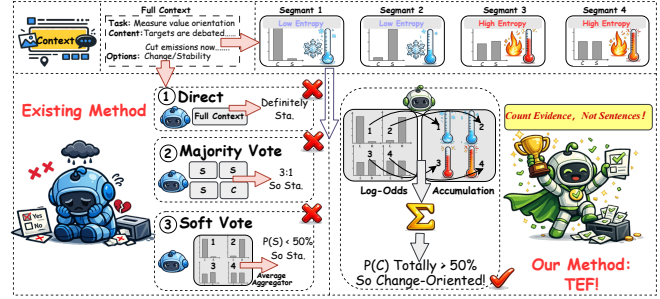


Fig. 1. Long-text value measurement requires counting informative evidence rather than treating all sentences equally. **LEFT:** Direct prediction collapses mixed content into one label, while voting can let many uncertain sentences outweigh a few decisive ones. **RIGHT:** TEF weights sentence log-odds by normalized information gain, discounting uncertain judgments so decisive evidence drives the document-level orientation.

Bearing this in mind, we formulate long-text value measurement as a decision-fusion problem, where each sentence acts as a local detector that emits a soft decision. Under conditional independence and calibrated sentence posteriors, the Bayes-optimal document-level rule sums sentence log-likelihood ratios. This observation places existing voting rules on a common scale: majority voting keeps only the sign of each sentence judgment, and soft voting uses a bounded surrogate that underweights highly decisive sentences. At the same time, directly summing LLM logits is unreliable because next-token probabilities are only approximately calibrated, especially near the point of maximal uncertainty.

To this end, we present **Tempered Evidence Fusion (TEF)**, a training-free fusion rule that tempers each sentence’s log-odds by its normalized information gain, as derived from a generalized Bayesian posterior [9, 10]. The resulting score nearly vanishes for uncertain sentences but preserves the Bayes-optimal weight of decisive evidence, yielding a soft censoring effect similar in spirit to censoring rules in distributed detection [11]. In Fig. 1 (right), TEF counts evidence rather than sentences, allowing a few informative sentences to determine the document-level orientation.

Our contributions are threefold. (i) We cast long-text value measurement with LLMs as a decision-fusion problem and derive TEF from a tempered Bayesian posterior, showing that it suppresses uncertain sentences while preserving the weight of decisive evidence. (ii) We introduce MIND, a five-year, cross-lingual benchmark of 8,358 posts over six value dimensions. (iii) We show that TEF outperforms the strongest of the Direct, Majority Vote, and Soft Vote baselines by 4.5 accuracy points and 4.6 macro-F1 points on average across five LLMs and two languages, yielding better-calibrated estimates.

2. RELATED WORK

LLM-based stance and ideology measurement has expanded from short-form texts such as tweets to longer, multi-issue documents [12, 13, 14]. In parallel, the reliability of LLM confidence has been studied from multiple perspectives, including token-level probabilities [15, 16], verbalized confidence [17, 18], and aggregation methods beyond majority voting [19]. In signal processing, fusing local decisions is a classical problem [20]: log-linear opinion pools combine probabilistic judgments [21], the Chair–Varshney rule weights local decisions by their reliabilities [22], and censoring schemes transmit only observations with informative likelihood ratios [11]. Despite these connections, long-text measurement with LLMs still largely relies on equal-weight voting. In contrast, we formulate long-text value measurement as sentence-level decision fusion and derive an evidence-aware fusion rule, TEF, that discounts uncertain sentences while preserving the weight of decisive ones.

3. PROPOSED TEF

Problem Formulation. We formulate long-text value measurement as a decision-fusion problem. The goal is to infer a latent social variable from a collection of textual segments whose evidential strength may vary substantially within the same document. Let $\mathcal{X}$ denote the input space and let $\mathcal{Y} = \{c_1, \dots, c_K\}$ be the set of target measurement categories. For an observation $X \in \mathcal{X}$, the latent measurement variable Y takes values in $\mathcal{Y}$. We segment each long text into N sentence-level units,

$$X \mapsto \mathcal{D}(X) = \{D_1, D_2, \dots, D_N\}, \quad (1)$$

where D_i denotes the i -th segment. This granularity preserves local semantic coherence while keeping inference computationally manageable. For each segment D_i , a language model $\mathcal{M}$ returns a probability distribution over the measurement categories, denoted by $\mathbf{p}_i = \mathcal{M}(D_i) = (p_i^{(1)}, \dots, p_i^{(K)})$, where each entry $p_i^{(k)} = P_{\mathcal{M}}(Y = c_k \mid D_i)$ and $\mathbf{p}_i \in \Delta^{K-1}$. The distribution encodes both a local prediction $\hat{y}_i = \arg \max_k p_i^{(k)}$ and its uncertainty, measured by entropy $H(\mathbf{p}_i) = -\sum_{k=1}^K p_i^{(k)} \log p_i^{(k)}$. Common aggregation rules discard part of this probabilistic evidence. Majority voting uses only the segment-level hard labels, while soft voting averages probability vectors with equal weight,

$$\hat{Y}_{MV} = \text{mode}(\{\hat{y}_i\}_{i=1}^N), \quad \hat{Y}_{SV} = \arg \max_{c \in \mathcal{Y}} \frac{1}{N} \sum_{i=1}^N p_i^{(c)}. \quad (2)$$

Both rules treat all segments as equally reliable. This is problematic for scientific measurement from long texts, where decisive evidence may appear in only a few segments and many other segments may be ambiguous, rhetorical, or weakly relevant. TEF instead aggregates reliability-weighted evidence. Let $w_i = \omega(\mathbf{p}_i) \in [0, 1]$ be the reliability assigned to segment D_i , and let $g : [0, 1] \rightarrow \mathbb{R}$ be a scoring transformation. The document-level measurement is $\hat{Y} = \arg \max_{c \in \mathcal{Y}} \sum_{i=1}^N w_i g(p_i^{(c)})$. The following paragraphs specify ω and g , which together define the fusion rule of TEF.

Reliability Weighting. TEF instantiates the preceding framework with entropy-based reliability weighting and one-vs-rest log-odds evidence. Algorithm 1 gives the complete procedure. We use normalized information gain as the segment reliability, $w_i = 1 - \frac{H(\mathbf{p}_i)}{\log K} \in [0, 1]$. This weight is close to one for concentrated distributions and close to zero for near-uniform distributions. Thus confident segments contribute strongly, while uncertain segments are softly censored.

Evidence Aggregation. For each category c , TEF transforms the segment posterior into a one-vs-rest log-odds score, $g(p_i^{(c)}) = \log \frac{p_i^{(c)}}{1 - p_i^{(c)}}$. The aggregated evidence for category c is

$$\text{Score}(c) = \sum_{i=1}^N w_i \log \frac{p_i^{(c)}}{1 - p_i^{(c)}}. \quad (3)$$

The final measurement outcome is $\hat{Y} = \arg \max_{c \in \mathcal{Y}} \text{Score}(c)$. In implementation, probabilities are clipped to $[\epsilon, 1 - \epsilon]$ before applying the logit, and log-odds values are clipped to $[-M, M]$. Unless otherwise stated, we set $\epsilon = 10^{-6}$ and $M = 10$. Unlike majority voting, TEF does not reduce $\mathbf{p}_i$ to a hard label before aggregation. Unlike soft voting, TEF does not assume that all segments are equally informative. It therefore preserves distributional information while reducing the impact of noisy or ambiguous segments.

Algorithm 1 Tempered Evidence Fusion

Require: Text X , model $\mathcal{M}$, categories $\mathcal{Y} = \{c_1, \dots, c_K\}$, clip bound M , stability constant ϵ .

Ensure: Measurement outcome $\hat{Y}$

```

1:  $\{D_1, \dots, D_N\} \leftarrow \mathcal{D}(X)$  ▷ Segment into sentences
2: Initialize  $\text{Score}(c) \leftarrow 0$  for all  $c \in \mathcal{Y}$ 
3: for  $i = 1$  to  $N$  do
4:    $\mathbf{p}_i \leftarrow \mathcal{M}(D_i)$  ▷ Segment-level distribution
5:    $H_i \leftarrow -\sum_{k=1}^K p_i^{(k)} \log p_i^{(k)}$  ▷ Entropy
6:    $w_i \leftarrow 1 - H_i / \log K$  ▷ Reliability weight
7:   for  $c \in \mathcal{Y}$  do
8:      $\tilde{p} \leftarrow \text{clip}(p_i^{(c)}, \epsilon, 1 - \epsilon)$ 
9:      $\ell \leftarrow \text{clip}\left(\log \frac{\tilde{p}}{1 - \tilde{p}}, -M, M\right)$ 
10:     $\text{Score}(c) \leftarrow \text{Score}(c) + w_i \ell$ 
11:   end for
12: end for
13:  $\hat{Y} \leftarrow \arg \max_{c \in \mathcal{Y}} \text{Score}(c)$ 
14: return  $\hat{Y}$ 

```

Theoretical Analysis.

Proposition 1 (Information Preservation). *Since the hard label $\hat{y}_i = \arg \max_k p_i^{(k)}$ is a deterministic function of $\mathbf{p}_i$, the data processing inequality [23] gives $I(Y; \mathbf{p}_i) \geq I(Y; \hat{y}_i)$. Thus, hard voting can discard reliability information even when two segments support the same category with different confidence.*

TEF follows this principle by retaining the full distribution $\mathbf{p}_i$ instead of reducing each segment to $\hat{y}_i$. Its entropy-based weight w_i distinguishes decisive evidence from ambiguous evidence, so two segments with the same hard label can contribute different amounts. The method therefore preserves the information discarded by majority voting while producing an additive document-level score.

Proposition 2 (Tempered Bayesian Pooling). *For a binary contrast with equal priors, calibrated conditionally independent segment posteriors satisfy the additive Bayes log-odds identity [24] $\log \frac{P(Y=c|D_{1:N})}{P(Y=\bar{c}|D_{1:N})} = \sum_{i=1}^N \log \frac{P(Y=c|D_i)}{P(Y=\bar{c}|D_i)}$. Therefore, evidence is naturally accumulated on the log-odds scale rather than through hard counts or raw probability averaging.*

TEF uses the one-vs-rest transform $g(p) = \log \frac{p}{1-p}$ to implement this additive evidence scale. Because LLM posteriors are approximate, it tempers each log-odds contribution [9] with $w_i = (\log K - H(\mathbf{p}_i)) / \log K$, the normalized information gain from a uniform prior. Near-uniform segments therefore have little influence, while low-entropy stance-bearing segments retain a contribution close to the Bayesian additive statistic. This gives TEF its soft-censoring behavior, as uncertain sentences do not accumulate by sheer number, while decisive sentences can determine the final document-level measurement outcome even when they are few.

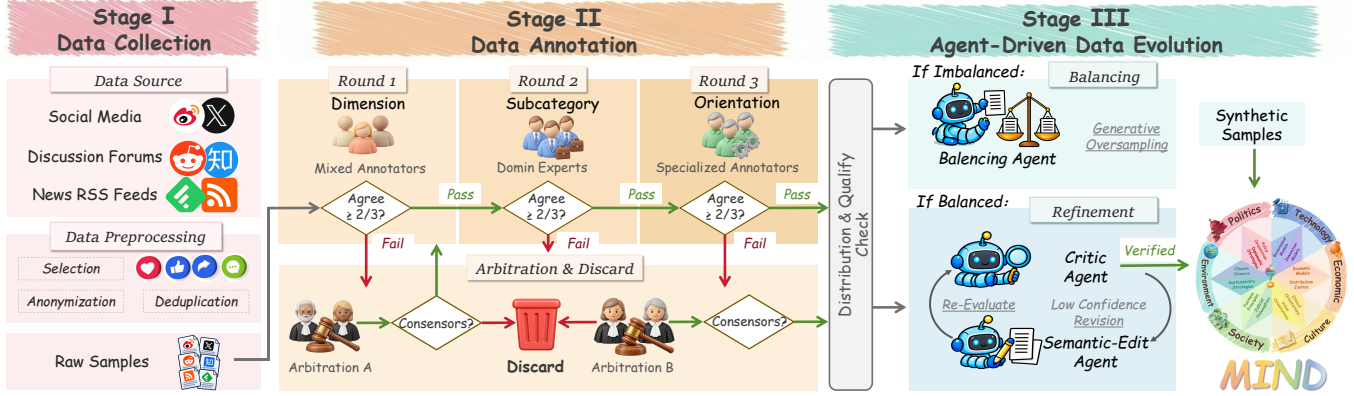


Fig. 2. Construction pipeline of MIND. Posts are collected from public sources, filtered for relevance, anonymized, and deduplicated, then annotated in three rounds for dimension, subcategory, and orientation before being balanced and refined by LLM agents.

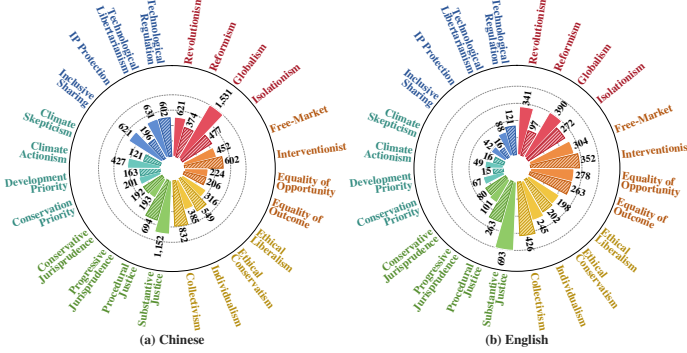


Fig. 3. Statistics of MIND. Radial bars report the number of labels for each value orientation, with colors denoting dimensions and paired bars representing the opposing orientations of each subcategory. Panels (a) and (b) correspond to Chinese and English posts.

4. THE MIND BENCHMARK

We construct the MIND dataset, a longitudinal, cross-lingual benchmark for measuring public value orientations. The final corpus contains 8,358 posts, including 5,474 Chinese posts and 2,884 English posts over six value dimensions in Fig. 3. MIND is built through a three-stage pipeline. **Stage I: Data Collection.** We collect candidate posts from X, Weibo, Reddit, Zhihu, discussion forums, and news feeds from January 2020 to January 2025, covering major public events such as the Russia–Ukraine war and the rise of generative AI. For each value dimension, keyword search is anchored to high-search-volume events. Raw posts are filtered for relevance and opinion content, anonymized, de-duplicated, and screened by an automatic quality score. **Stage II: Data Annotation.** We adopt a three-round annotation framework to label the value dimension, subcategory, and orientation in sequence. Each round uses three annotators and accepts a label only when at least two agree. Failed or disputed cases enter arbitration, samples without a stable majority are discarded, and accepted samples pass distribution and quality checks before the evolution stage. **Stage III: Agent-Driven Data Evolution.** To address class imbalance and semantic ambiguity, we introduce an agent-driven evolution stage. When a category is underrepresented, a balancing agent generates synthetic samples through generative oversampling. For low-confidence or ambiguous samples, a critic agent evaluates the alignment between text and label, and a semantic-edit agent revises the text to make the labeled orientation explicit. Only verified samples are retained in the final benchmark.

5. EXPERIMENTS

Experiment Settings. We evaluate three open-source models, Qwen2.5-7B [25], LLaMA3-8B [26], and Qwen3-14B [27], and two proprietary models, DeepSeek-V3.2 [28] and GPT-4o-mini [29], using one shared prompt per language for TEF, Direct, Majority Vote, and Soft Vote. Each prompt specifies the dimension and its two positions and requests a single answer token. We read the answer-token log-probabilities without sampling. Open-source models are served on RTX 3090 GPUs. We set $\epsilon = 10^{-6}$ and $M = 10$, and report dimension-level accuracy and macro-F1.

Main Results. In Table 3, TEF has the highest accuracy and macro-F1 for every model in both languages, exceeding the strongest baseline by 4.5 accuracy and 4.6 macro-F1 points on average, so the gain comes from the fusion rule rather than from model-specific tuning. At the dimension level, TEF is the best or tied for best in 116 of 120 comparisons and never trails the strongest baseline by more than 2.6 points. Segmentation alone does not ensure improvement, since equal-weight voting, whether hard or soft, does not consistently improve on Direct. The gains of TEF concentrate where the baselines are weakest, on the smaller open-source models, on culture and politics in Chinese, and on environment in English, whereas on dimensions where Direct is already strong TEF matches it.

Ablation Study. Table 4 evaluates TEF by removing one component at a time. On Qwen2.5-7B, both variants retain most of TEF’s average gain over Soft Vote; on DeepSeek-V3.2, however, both fall below Soft Vote in both languages on average. Their complementarity reflects their distinct roles: the log-odds transform emphasizes decisive evidence, while entropy weighting suppresses uncertain sentences. The largest accuracy drop occurs on DeepSeek-V3.2’s English environment dimension, where removing entropy weighting or the log-odds transform reduces accuracy by 22.9 or 23.8 percentage points.

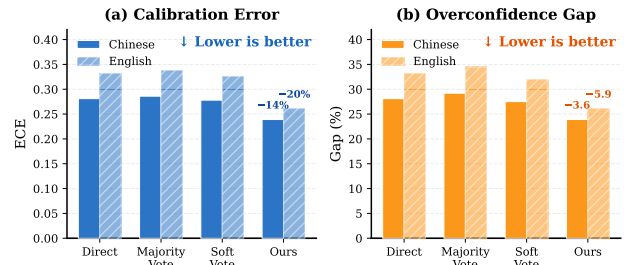


Fig. 4. Calibration on Qwen2.5-7B. TEF achieves the lowest expected calibration error and overconfidence gap across both Chinese and English settings. Compared with Direct, Majority Vote, and Soft Vote, TEF aligns confidence with accuracy by tempering weak evidence.

Table 3. Main results on MIND. Accuracy (Acc) and macro-F1 (F1) in percent for every model, language, and dimension. Cul., Eco., Soc., Pol., Tech., and Env. denote culture, economy, society, politics, technology, and environment. MV and SV denote Majority Vote and Soft Vote, and Ours denotes TEF. For each model, language, dimension, and metric, the best method is in **bold** and the second best is underlined.

Model	Method	Chinese												English											
		Cul.		Eco.		Soc.		Pol.		Tech.		Env.		Cul.		Eco.		Soc.		Pol.		Tech.		Env.	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Qwen2.5-7B	Direct	44.6	44.2	44.6	44.5	74.4	73.2	55.2	54.1	72.1	66.3	82.6	81.8	54.0	53.4	47.2	47.1	73.5	69.7	56.6	53.6	62.9	56.2	70.6	64.6
	MV	44.4	43.7	43.4	41.3	64.0	64.0	50.5	50.3	67.6	64.7	71.6	71.5	53.1	52.2	50.5	43.6	49.7	49.7	60.9	53.3	62.9	58.1	51.4	48.9
	SV	43.5	42.8	43.2	40.6	65.7	65.6	51.8	51.5	69.9	66.3	73.4	73.3	53.2	52.5	49.0	41.8	50.8	50.8	60.8	53.7	63.4	61.3	55.0	52.3
	Ours	72.1	69.9	47.7	45.1	81.5	80.6	68.3	62.3	72.8	66.5	85.1	84.1	62.5	58.2	59.4	59.4	79.6	77.3	67.8	66.4	64.6	61.6	80.7	67.2
LLaMA3-8B	Direct	63.5	58.2	36.4	35.9	73.8	70.7	60.6	56.5	70.9	64.9	80.8	79.5	62.9	53.4	41.1	40.8	74.8	64.1	49.9	48.9	54.9	52.6	59.6	56.6
	MV	59.9	51.8	36.7	36.5	68.9	68.6	55.9	54.1	66.5	61.9	74.4	73.9	54.9	48.9	47.8	35.3	60.2	59.7	56.5	<u>50.4</u>	54.9	54.4	48.6	47.6
	SV	60.8	53.6	<u>37.3</u>	<u>37.0</u>	73.1	<u>72.9</u>	54.3	52.8	<u>70.0</u>	64.1	<u>78.2</u>	<u>77.5</u>	55.4	49.0	<u>48.1</u>	36.5	62.2	61.5	<u>54.6</u>	46.8	<u>52.0</u>	<u>51.7</u>	50.5	48.5
	Ours	65.9	59.9	38.1	38.0	82.7	82.0	64.4	62.4	70.9	65.1	78.2	77.5	63.3	56.2	55.8	53.5	80.5	77.6	67.5	63.4	58.9	58.7	78.0	70.5
Qwen3-14B	Direct	73.7	71.7	41.3	39.9	83.9	82.8	61.3	58.6	74.5	68.9	87.9	86.8	60.4	46.6	<u>56.2</u>	<u>56.2</u>	82.3	76.3	50.1	46.3	60.0	53.6	74.3	60.1
	MV	69.5	65.7	41.3	39.9	81.2	79.5	60.9	58.1	73.1	63.9	84.4	83.6	60.6	54.0	<u>56.2</u>	<u>56.2</u>	82.5	76.7	49.7	46.1	57.1	52.6	75.2	63.4
	SV	73.7	71.8	40.9	38.3	83.7	82.6	61.1	58.4	74.5	68.9	87.9	86.8	60.6	54.0	55.8	55.8	80.2	72.8	49.9	46.5	60.6	54.0	75.2	63.4
	Ours	73.8	71.9	42.1	40.7	83.9	82.8	68.6	64.0	74.8	69.2	88.0	87.0	62.3	54.3	57.9	57.8	82.5	76.8	50.7	47.4	60.6	54.1	76.1	64.2
DeepSeek-V3.2	Direct	68.7	66.0	39.6	39.4	79.8	78.8	62.3	59.8	73.7	67.1	86.4	85.0	60.2	59.9	49.5	49.5	79.2	76.9	55.7	51.2	57.1	52.2	52.3	50.4
	MV	68.5	66.3	38.6	38.5	78.5	77.5	61.9	61.6	71.5	66.6	83.6	82.6	59.0	57.9	57.7	58.5	<u>77.2</u>	74.9	61.8	49.2	57.1	54.5	54.1	52.7
	SV	68.5	65.8	40.0	39.8	80.2	79.2	62.6	62.2	73.3	67.6	85.9	84.9	59.8	59.6	58.1	58.0	79.1	76.8	62.7	51.2	58.3	55.5	51.4	49.7
	Ours	75.6	73.5	42.4	40.5	83.9	82.7	62.8	62.3	74.4	68.0	86.4	85.4	63.7	60.3	58.8	58.8	83.9	80.3	62.7	55.4	60.0	58.7	74.3	66.4
GPT-4o-mini	Direct	63.9	53.8	42.2	39.1	82.0	81.2	67.4	63.7	74.9	61.8	83.6	82.3	60.9	47.1	49.8	42.0	81.7	77.0	54.4	53.3	63.4	55.2	79.8	68.7
	MV	64.0	54.4	<u>42.2</u>	38.9	80.2	78.8	<u>67.7</u>	65.3	73.0	60.4	82.0	80.1	59.8	45.7	49.8	40.3	79.7	75.6	56.1	55.5	61.7	53.9	81.7	68.1
	SV	64.3	54.3	42.0	38.3	81.3	79.9	<u>67.7</u>	65.2	74.2	61.0	85.9	84.3	60.8	46.8	50.3	42.0	81.3	76.8	55.8	54.5	62.3	54.9	80.7	68.5
	Ours	74.0	71.2	42.9	41.1	<u>81.4</u>	<u>79.9</u>	67.8	65.3	75.3	68.9	85.9	84.4	64.2	57.5	50.3	42.8	83.9	79.7	57.5	56.7	64.6	56.6	83.5	72.4

Table 4. Ablation of TEF. Accuracy and macro-F1 (%) on Qwen2.5-7B and DeepSeek-V3.2; abbreviations as in Table 3. w/o Entropy keeps only the log-odds sum; w/o Log-Odds averages entropy-weighted probabilities instead of log-odds. Ours is the full TEF (**bold**).

Model	Language	Method	Cul.		Eco.		Soc.		Pol.		Tech.		Env.		Average	
			Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Qwen2.5-7B	Chinese	Ours	72.1	69.9	47.7	45.1	81.5	80.6	68.3	62.3	72.8	66.5	85.1	84.1	71.3	68.1
		w/o Entropy	71.3	68.8	45.6	42.9	79.9	79.1	67.4	61.4	71.7	65.8	81.6	80.7	69.6	66.5
		w/o Log-Odds	71.8	69.6	46.2	43.5	80.3	79.5	67.2	61.1	72.2	66.2	81.5	80.6	69.9	66.8
	English	Ours	62.5	58.2	59.4	59.4	79.6	77.3	67.8	66.4	64.6	61.6	80.7	67.2	69.1	65.0
		w/o Entropy	61.8	56.6	58.0	58.0	77.2	75.2	66.1	65.1	62.3	56.2	75.2	60.9	66.8	62.0
		w/o Log-Odds	61.8	57.0	58.5	58.4	77.1	75.1	66.1	65.0	63.4	56.7	76.1	61.7	67.2	62.3
DeepSeek-V3.2	Chinese	Ours	75.6	73.5	42.4	40.5	83.9	82.7	62.8	62.3	74.4	68.0	86.4	85.4	70.9	68.7
		w/o Entropy	68.1	65.4	39.2	39.0	79.6	78.7	62.4	61.9	72.7	67.1	85.7	84.8	68.0	66.2
		w/o Log-Odds	68.5	65.7	39.3	39.1	79.5	78.5	62.7	62.2	73.3	67.6	84.9	83.8	68.0	66.2
	English	Ours	63.7	60.3	58.8	58.8	83.9	80.3	62.7	55.4	60.0	58.7	74.3	66.4	67.2	63.3
		w/o Entropy	58.9	58.5	57.1	56.9	78.5	76.3	62.0	50.5	53.7	51.4	51.4	49.7	60.3	57.2
		w/o Log-Odds	59.3	59.1	56.6	56.4	78.8	76.5	62.4	50.9	54.3	51.3	50.5	48.9	60.3	57.2

Table 1. Robustness Analysis. Average accuracy (%) over both languages under the original, verbose, and minimal prompts. Strongest baseline is the best of Direct, MV, and SV per prompt; Ours is TEF.

Model	Strongest baseline			Ours		
	Original	Verbose	Minimal	Original	Verbose	Minimal
Qwen2.5-7B	61.5	60.8	59.2	70.2	69.8	68.6
LLaMA3-8B	60.8	60.0	58.2	67.0	66.4	65.3
Qwen3-14B	67.2	66.5	64.8	68.4	67.9	66.8
DeepSeek-V3.2	65.0	64.3	62.7	69.1	68.5	67.5
GPT-4o-mini	67.2	66.5	65.0	69.3	68.8	67.8

Table 2. Efficiency Analysis. Total tokens and cost over the twelve evaluation subsets and mean time to first token (TTFT) on RTX 3090 GPUs. Voting covers MV and SV, which issue the same queries.

Model	Method	Tokens (M)	TTFT (ms)	Cost (\$)
Qwen2.5-7B	Direct	0.5	23.7	0.09
	Voting	2.7	14.4	0.27
	Ours	2.7	15.8	0.28
LLaMA3-8B	Direct	0.6	28.6	0.10
	Voting	3.1	15.2	0.28
	Ours	3.1	17.3	0.30

Analysis. 1) *Calibration.* Because measurement outputs are thresholded or manually reviewed, document-level confidence should reflect correctness. In Fig. 4, Direct and both voting rules are overconfident on Qwen2.5-7B, whereas TEF achieves the lowest expected calibration error [30, 31] and overconfidence gap. Its confidence is therefore better suited to selective review. 2) *Robustness.* To test whether the gain is due to the fusion rule rather than the shared prompt, we rerun all methods with a verbose prompt that adds role background and annotation guidelines and with a minimal prompt that retains only the classification directive. In Table 1, TEF remains best under every prompt and degrades the least under either perturbation. 3) *Efficiency.* In Table 2, TEF uses the same single-token queries as the voting methods, so its token count is identical. Its time to first token differs by only about two milliseconds. Relative to Direct, sentence-level methods cost roughly three times as much, but TEF is the only one that consistently turns this overhead into a performance gain.

6. CONCLUSION

We cast long-text value measurement as the fusion of soft sentence decisions and derived TEF, which tempers each sentence’s log-odds by its information gain. On MIND, TEF is more accurate than direct prediction and voting for five LLMs in two languages and better calibrated on Qwen2.5-7B, and its gain survives changes of the prompt.

7. REFERENCES

- [1] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini, “Political polarization on Twitter,” in *Proceedings of the International AAAI Conference on Web and Social Media*, Barcelona, Spain, 2011, vol. 5, pp. 89–96.
- [2] Byunghwee Lee, Rachith Aiyappa, Yong-Yeol Ahn, Haewoon Kwak, and Jisun An, “A semantic embedding space based on large language models for modelling human beliefs,” *Nature Human Behaviour*, vol. 9, no. 9, pp. 1928–1940, 2025.
- [3] Nick Pangakis and Sam Wolken, “Keeping humans in the loop: Human-centered automated annotation with generative AI,” in *Proceedings of the International AAAI Conference on Web and Social Media*, 2025, vol. 19, pp. 1471–1492.
- [4] Tunazzina Islam and Dan Goldwasser, “Uncovering latent arguments in social media messaging by employing LLMs-in-the-loop strategy,” in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 7397–7429.
- [5] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli, “ChatGPT outperforms crowd workers for text-annotation tasks,” *Proceedings of the National Academy of Sciences*, vol. 120, no. 30, pp. e2305016120, 2023.
- [6] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang, “Can large language models transform computational social science?,” *Computational Linguistics*, vol. 50, no. 1, pp. 237–291, 2024.
- [7] Petter Törnberg, “Large language models outperform expert coders and supervised classifiers at annotating political social media messages,” *Social Science Computer Review*, vol. 43, no. 6, pp. 1181–1195, 2024.
- [8] Dimitrios Tsirmpas, Ioannis Gkionis, Georgios Th. Papadopoulos, and Ioannis Mademlis, “Neural natural language processing for long texts: A survey on classification and summarization,” *Engineering Applications of Artificial Intelligence*, vol. 133, pp. 108231, 2024.
- [9] Pier Giovanni Bissiri, Chris C. Holmes, and Stephen G. Walker, “A general framework for updating belief distributions,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 78, no. 5, pp. 1103–1130, 2016.
- [10] Peter Grünwald and Thijs van Ommen, “Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it,” *Bayesian Analysis*, vol. 12, no. 4, pp. 1069–1103, 2017.
- [11] Constantino Rago, Peter Willett, and Yaakov Bar-Shalom, “Censoring sensors: A low-communication-rate scheme for distributed detection,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 32, no. 2, pp. 554–568, 1996.
- [12] Rudra Ranajee Saha, Laks V. S. Lakshmanan, and Raymond T. Ng, “Stance detection with explanations,” *Computational Linguistics*, vol. 50, no. 1, pp. 193–235, 2024.
- [13] Abu Ubaida Akash, Ahmed Fahmy, and Amine Trabelsi, “Can large language models address open-target stance detection?,” in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 971–985.
- [14] Prasanta Bhattacharya, Hong Zhang, Yiming Cao, Wei Gao, Brandon Siyuan Loh, Joseph J. P. Simons, and Liang Ze Wong, “Rethinking stance detection: A theoretically-informed research agenda for user-level inference using language models,” *arXiv preprint arXiv:2502.02074*, 2025.
- [15] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig, “How can we know when language models know? On the calibration of language models for question answering,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 962–977, 2021.
- [16] Saurav Kadavath et al., “Language models (mostly) know what they know,” *arXiv preprint arXiv:2207.05221*, 2022.
- [17] Katherine Tian et al., “Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 5433–5442.
- [18] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar, “A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions,” *ACM Computing Surveys*, vol. 58, no. 3, pp. 1–38, 2025.
- [19] Rui Ai, Yuqi Pan, David Simchi-Levi, Milind Tambe, and Haifeng Xu, “Beyond majority voting: LLM aggregation by leveraging higher-order information,” *arXiv preprint arXiv:2510.01499*, 2025.
- [20] Robert R. Tenney and Nils R. Sandell, “Detection with distributed sensors,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-17, no. 4, pp. 501–510, 1981.
- [21] Christian Genest and James V. Zidek, “Combining probability distributions: A critique and an annotated bibliography,” *Statistical Science*, vol. 1, no. 1, pp. 114–135, 1986.
- [22] Z. Chair and Pramod K. Varshney, “Optimal data fusion in multiple sensor detection systems,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-22, no. 1, pp. 98–101, 1986.
- [23] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, Wiley, 2nd edition, 2005.
- [24] Pramod K. Varshney, *Distributed Detection and Data Fusion*, Springer, New York, NY, 1997.
- [25] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al., “Qwen2.5 technical report,” 2024, arXiv:2412.15115.
- [26] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al., “The Llama 3 herd of models,” 2024, arXiv:2407.21783.
- [27] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al., “Qwen3 technical report,” 2025, arXiv:2505.09388.
- [28] DeepSeek-AI, “DeepSeek-V3.2: Pushing the frontier of open large language models,” 2025, arXiv:2512.02556.
- [29] OpenAI, “GPT-4o mini: Advancing cost-efficient intelligence,” OpenAI blog, July 2024.
- [30] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht, “Obtaining well calibrated probabilities using Bayesian binning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015, pp. 2901–2907.
- [31] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger, “On calibration of modern neural networks,” in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 1321–1330.