CORE: Discovering Intrinsic Ranking Preferences in LLMs via Consistent Ego-Correction

Anonymous ACL submission

Abstract

Large language models are powerful listwise rerankers, yet their performance remains highly sensitive to prompt variations, undermining their reliability for real-world applications. To address this, we propose CORE, a new fine-tuning framework that mitigates this instability by learning a model's intrinsic, prompt-invariant ranking preferences. CORE integrates two complementary mechanisms: a guidance strategy adapted from Classifier-Free Guidance to calibrate the generative process against stylistic variations, and a consistency loss based on differentiable Kendall's Tau to regularize the model's internal ordinal judgments. On standard TREC Deep Learning and BEIR benchmarks, CORE establishes new state-of-the-art ranking performance. Crucially, CORE demonstrates superior robustness, reducing performance variance across diverse prompts by over 80% compared to standard fine-tuning. Our work presents a principled and effective method for building powerful and trustworthy LLM-based reranking systems.

1 Introduction

Large Language Models (LLMs) have recently emerged as powerful components in information retrieval (IR) and document ranking systems (Sun et al., 2023a; Long et al., 2025; Gao et al., 2025; Zhuang et al., 2024b). Due to their strong semantic understanding, reasoning, and generation capabilities, LLMs can be adapted to ranking tasks via flexible prompting paradigms (zero-shot, few-shot) or fine-tuning (supervised tuning, direct preference optimization) (Sun et al., 2025), often surpassing the capabilities of traditional neural rankers. LLMs can assess relevance with a nuance that traditional, sparse-vector or dense-vector models often miss (Ma et al., 2023; Sun et al., 2023b). This allows them to capture subtle semantic relationships between a query and a document, making them an invaluable component in modern search pipelines.

LLM ranking approaches can be broadly categorized into pointwise, pairwise, listwise, and setwise paradigms (Sun et al., 2025; Long et al., 2025; Zhuang et al., 2024b). These differ in how the LLM processes relevance signals. In a pointwise approach (Sachan et al., 2022; Zhuang et al., 2024a; Fan et al., 2025), the LLM evaluates each document's relevance to the query independently. For instance, an LLM may be prompted to output a binary "Yes/No" or a score indicating whether a given document is relevant to the query, and each document is ranked by this score. This paradigm is straightforward and allows parallel scoring of documents, but it ignores inter-document comparisons. In a pairwise approach (Qin et al., 2024; Chen et al., 2025), the LLM compares two documents at a time to decide which is more relevant (e.g. prompting "Which document, A or B, is more relevant to the query?"). Repeating such comparisons across document pairs can yield a preference-based ranking. Pairwise LLM rankers such as PRP (Qin et al., 2024) tend to achieve high accuracy by directly modeling comparative relevance, at the cost of requiring many LLM inference calls (quadratic in the number of documents). In a listwise approach (Ren et al., 2025; Liu et al., 2025), the LLM considers the entire set of candidate documents and produces a sorted list in one go. Listwise prompting can capture complex dependencies between documents (such as diversity or redundancy) and fully leverage the LLM's generative ability to output an ordered list. Early works like RankT5 (Zhuang et al., 2023) showed the feasibility of sequence-to-sequence ranking, and more recent zero-shot systems like RankGPT (Sun et al., 2023a) and RankVicuna (Pradeep et al., 2023a) have demonstrated strong listwise re-ranking performance using GPT-4 and Vicuna-13B respectively.

Despite the superior performance of LLM-based rankers, the reliability of LLM-based rerankers is severely undermined by a critical weakness:

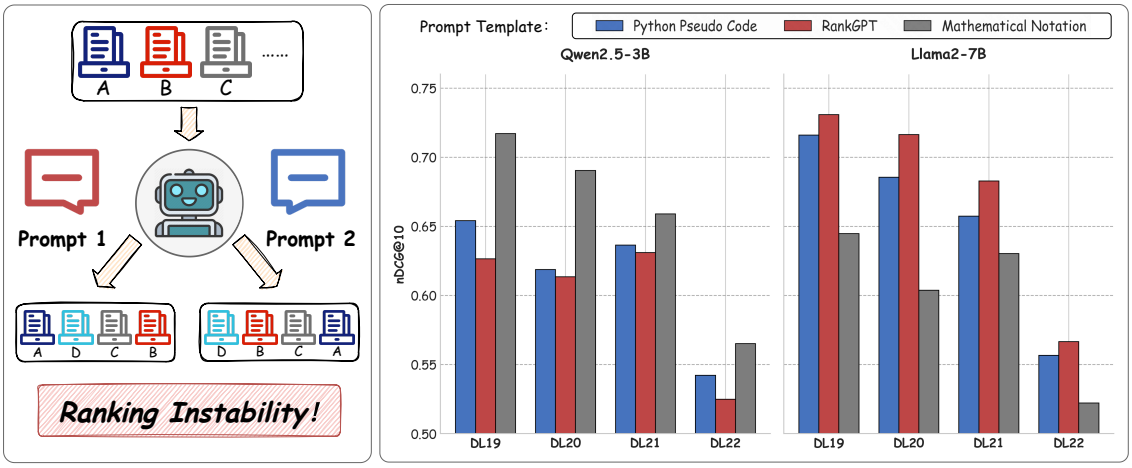

Figure 1: The problem of prompt sensitivity in LLM rerankers. The left panel provides a conceptual illustration: for the same set of documents, two semantically equivalent but stylistically different prompts can lead to disparate ranking outcomes. The right panel presents empirical evidence, showing that the nDCG@10 performance for both Qwen2.5-3B and Llama2-7B in the zero-shot setting varies significantly across three different prompt templates on the TREC DL datasets, highlighting the severity of this issue.

prompt sensitivity. LLMs are notorious for their susceptibility to prompt wording and format – seemingly minor differences in how the query and instructions are phrased can lead to significant changes in the output ranking (Chatterjee et al., 2024; Sclar et al., 2024a; Arabzadeh and Clarke, 2025). For an identical query and document set, subtle, semantically irrelevant variations to the prompt—such as changes in wording, output format, or even the initial order of documents—can lead to dramatically different ranked lists. This instability manifests as significant positional biases and a stark lack of invariance to input permutations (Tang et al., 2024; Sun et al., 2025). We visually demonstrate this problem in Figure 1.

Furthermore, when controlling for the same LLM, the effectiveness gaps between pointwise, pairwise, listwise, and setwise methods become much smaller once prompt variations are taken into account (Sun et al., 2025). It implies that the seemingly better prior studies may stem from superior prompt engineering. This prompt sensitivity undermines the reliability of LLM rankers: it becomes unclear whether a performance gain is due to a truly better ranking technique or just a better prompt.

To move beyond merely observing this instability and toward a principled solution, we argue that the goal is to uncover the model's intrinsic preference—a stable, core ranking capability shielded from superficial prompt variations. Just as a master chef aims to replicate a signature dish's taste consistently regardless of the kitchen's conditions, we seek to distill the LLM's core ranking capability. We find a powerful analogy for this challenge in Sigmund Freud's structural model of the psyche (Freud, 1923). We metaphorically define the model's latent, intrinsic ranking capability as its **"Id"** (R^*)—its intrinsic preference. In contrast, the actual, observable ranked list produced under a specific prompt p is its **"Ego"** (R_p)—the Prompt-specific realization. From this perspective, the observed ranking Instability is not just a surface-level inconsistency; it signifies a fundamental "cognitive gap" between the model's singular "Id" and its multiple, prompt-dependent "Egos". Therefore, the central goal of this work is to resolve this cognitive gap: we aim to learn the robust, prompt-invariant "Id" by treating the inconsistent "Egos" as signals for regularization. In this framework, the ground-truth relevance labels serve as the "Superego"—the external ideal that guides and corrects the Ego's alignment with the Id. We term this process "Consistent Ego-Correction".

To resolve this gap, we propose a novel fine-tuning framework, CORE (COnsistent Reranking via Ego-correction), to discover and stabilize the intrinsic preferences of LLM rankers. The framework integrates two key components: **1) External Behavior Calibration via Inverse CFG:** Adapting Classifier-Free Guidance (CFG) (Ho and Salimans, 2022), this mechanism uses a generic, task-defining prompt to anchor the model's response, mitigating biases from prompt-specific realizations. **2) Internal Judgment Consistency via Differentiable**

Kendall's Tau: A novel loss function based on a differentiable Kendall's Tau (Kendall, 1938; Guan et al., 2024) is used to enforce consistent relative rankings of documents across multiple prompt variations. These components compel the model to develop a robust ranking function that is insensitive to superficial prompt phrasing.

The main contributions of this work are summarized as follows:

- We propose a novel cognitive framework (Id-Ego) and provide a formal technical interpretation for it, offering a new perspective for understanding and addressing the prompt sensitivity problem in LLM rankers.
- We design the CORE methodology, which uniquely combines a CFG-based guidance mechanism with an ordinal consistency regularizer, specifically for enhancing the robustness of listwise reranking.
- We conduct extensive empirical evaluations on multiple standard IR benchmarks, showing that CORE not only achieves state-of-the-art ranking effectiveness but, more importantly, exhibits significantly enhanced robustness against prompt variations.

2 Problem Formulation

We focus on the task of listwise generative reranking. Given a query q and a candidate document set $D = \{d_1, \ldots, d_N\}$, the goal is to generate a permutation π that orders documents by relevance. In the LLM-based paradigm, this function is parameterized by a prompt P constructed from instructions I, query q, and documents D: $P = f_{\text{prompt}}(I, q, D)$. The model autoregressively generates the ranked list $\pi_P = \text{LLM}_\theta(P)$.

The Id-Ego Framework. A critical challenge in this paradigm is ranking instability: semantically equivalent but stylistically different prompts $\mathcal{P} = \{P_{prm}^{(1)}, \ldots, P_{prm}^{(K)}\}$ often yield divergent rankings. To formalize this, we propose the **Id-Ego** cognitive framework. We posit the existence of an intrinsic, prompt-invariant ranking preference R^*, metaphorically termed the model's **"Id"**. However, in practice, we only observe prompt-conditioned realizations R_p, or the **"Ego"**, which are contaminated by superficial stylistic variations. Ranking instability essentially represents the cognitive gap between the latent R^* and the observed $\{R_p\}$. Therefore, our objective is to recover the robust R^* by treating the inconsistent R_p as noisy signals for regularization, using ground-truth labels as the "Superego" to guide this alignment.

3 Related Work

LLM Ranking and Prompt Sensitivity. LLM-based reranking has evolved across pointwise (Sachan et al., 2022; Zhuang et al., 2023), pairwise (Qin et al., 2024), and listwise paradigms (Sun et al., 2023a). While listwise models like RankGPT and RankVicuna leverage full context for superior global sorting (Sun et al., 2025, 2023a; Pradeep et al., 2023a; Waldo and Boussard, 2024), they expose a critical vulnerability: extreme prompt sensitivity. Minor variations in wording or initial document order can drastically alter ranking outcomes (Chatterjee et al., 2024; Sclar et al., 2024b; Arabzadeh and Clarke, 2025; Hu et al., 2024). This instability often manifests as positional bias, violating the principle of permutation invariance (Tang et al., 2024) and complicating scientific evaluation.

Mitigating Instability. Current mitigation strategies fall into two categories. **Inference-time** methods, such as permutation self-consistency (Tang et al., 2024), aggregate outputs from multiple inputs to neutralize bias (Wang et al., 2023; Zhou et al., 2024). While effective, they incur high computational costs.**Training-time** approaches are more fundamental, employing data augmentation (Ngweta et al., 2025; Wei et al., 2025) or contrastive objectives (Qiang et al., 2024) to instill robustness. Distinct from these, our framework CORE adopts a Bayesian perspective (Zhao et al., 2021; Fortuin, 2021; Sam et al., 2024) to learn an intrinsic ranking preference ("Id") robust to superficial prompt noise ("Ego"). We operationalize this via inverse Classifier-Free Guidance (Ho and Salimans, 2022) and a differentiable Kendall's Tau consistency loss (Kendall, 1938; Guan et al., 2024; Zheng et al., 2023), achieving robustness in a single forward pass without inference overhead.

4 Method

In this section, we present our methodology, CORE, a fine-tuning framework designed to instill prompt-invariance ability in listwise rerankers. The overall architecture of our framework is illustrated in Figure 2. It resolves the "cognitive gap" between a model's intrinsic ranking capability ("Id") and its prompt-dependent outputs ("Egos") through a dual

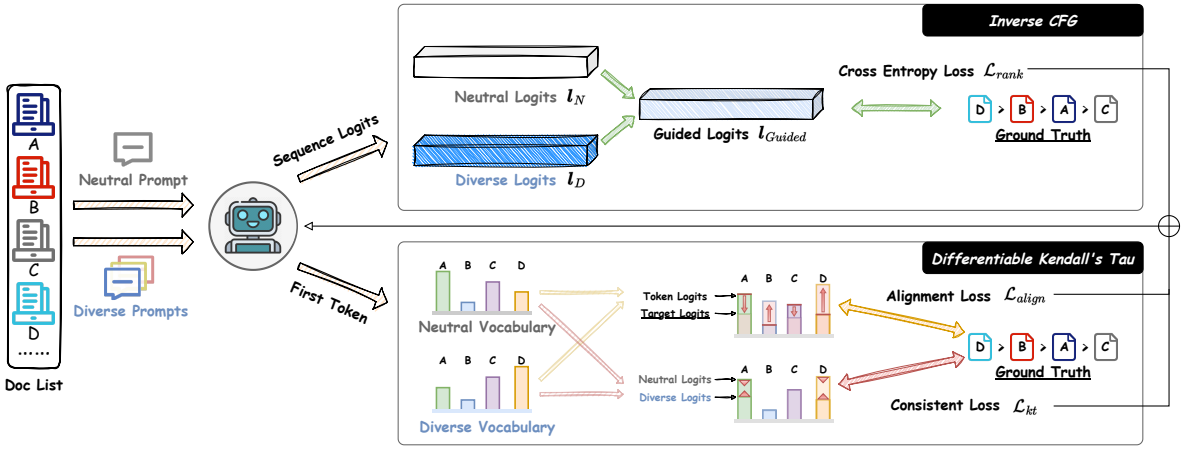

Figure 2: Overview of the proposed CORE framework. It fine-tunes the model through two complementary mechanisms. **Top:** This module calibrates the model's external generative process. It interpolates the logits produced under a diverse prompt with those from a canonical neutral prompt to compute guided logits for the main ranking loss, L_{rank}. **Bottom:** This module regularizes the model's internal ranking judgment. It uses the first-token logits as a differentiable proxy for the overall preference, supervising it for accuracy with an alignment loss, L_{align}, and for stability across prompts with a consistency loss, L_{kt}.

strategy that we term "external behavior and calibration internal judgment consistency." Externally, we calibrate its step-by-step generative behavior to be less susceptible to stylistic prompt variations. Internally, we enforce consistency on the model's holistic ranking judgment before generation begins.

4.1 External Calibration via Inverse CFG

Our first component targets the model's generative process. The goal is to make the model's final output robust to prompt variations. To achieve this during training, we use two types of prompts: a single, canonical **neutral prompt** (p_{neu}) that represents the pure ranking task, and a set of **diverse prompts** (p_{prm}) that contain stylistic variations.

This approach draws a conceptual parallel to Classifier-Free Guidance (CFG) in diffusion models (Ho and Salimans, 2022; Chung et al., 2025). While standard CFG was originally designed to amplify the influence of a condition to improve stylistic adherence, our objective is the **inverse**: to mitigate the influence of the prompt's stylistic "noise" and steer the model back toward its intrinsic ranking preference. Inspired by this principle, we treat the output from the noisy p_{prm} as the "conditional output" and the output from p_{neu} as the stable "unconditional" baseline.

Formally, let $\boldsymbol{l}_{neu}$ and $\boldsymbol{l}_{prm}$ denote the logit vectors produced using the neutral and diverse prompts. We compute the calibrated logit vector, $\boldsymbol{l}_{guided}$, by interpolating between them to dampen the noise:

$$\boldsymbol{l}_{guided} = \boldsymbol{l}_{neu} + w(\boldsymbol{l}_{prm} - \boldsymbol{l}_{neu}) \tag{1}$$

where $w \in [0, 1]$ is the calibration weight. This formulation aligns the potentially biased output from a diverse prompt with the stable baseline. We theoretically verify that this interpolation strictly tightens the upper bound of the mis-ranking probability:

Theorem 4.1 (Calibration Guarantee)**.** *Assume intrinsic preference l^*, and observed logits l_{prm}, l_{neu} contain independent sub-Gaussian noise with variances σ_p^2, σ_n^2. Let m^* be the intrinsic margin. The mis-ranking probability is bounded by:*

$$\mathbb{P}[(l_g)_y \leq (l_g)_b] \leq \exp\left(-\frac{(m^*)^2}{2\sigma_S^2(w)}\right)$$

where $\sigma_S^2(w) = (1-w)^2\sigma_n^2 + w^2\sigma_p^2$ is the combined variance. For any $w < 1$, this bound is strictly lower than using l_{prm} alone.

Proof. Detailed derivation is provided in Appendix A.3.

Finally, we apply a standard listwise ranking loss on these robust guided logits:

$$\mathcal{L}_{rank} = -\sum_{j=1}^{N} y_j \log \sigma(\boldsymbol{l}_{guided,j}) \tag{2}$$

This anchors the model's output to the core ranking task, reducing sensitivity to prompt forms.

4.2 Internal Consistency via Differentiable Kendall's Tau

Complementing external calibration, we further enforce internal consistency on the model's ranking preferences. Our goal is to minimize the ordinal distance between rankings generated by neutral and diverse prompts. Since standard rank correlation metrics are non-differentiable, we propose a fully differentiable approximation to enable end-to-end regularization.

Conceptually, Kendall's Tau measures ordinal correlation by comparing the number of concordant pairs (P_c), where items are in the same relative order, against discordant pairs (P_d). The formula for its simplest form, τ_a, is:

$$\tau_a = \frac{P_c - P_d}{\frac{1}{2}N(N-1)} \tag{3}$$

This metric directly aligns with our objective of maximizing ordinal agreement.

However, applying Equation 3 directly to optimize a generative LLM encounters two fundamental challenges. First, the discrete nature of autoregressive generation disrupts gradient flow. The transition from model weights to a final, sorted textual sequence (e.g., "[3] >[1] > [2]") involves non-differentiable argmax operations, rendering the direct optimization of the final ranking order infeasible. This necessitates a continuous proxy for the model's ranking preference. Second, the standard Kendall's Tau coefficient is non-differentiable. It relies on the discrete counting of concordant and discordant pairs, which prevents its direct application as a loss function in gradient-based training. Our method systematically addresses these challenges by introducing a differentiable ranking proxy and a smoothed loss function.

Differentiable Ranking Proxy. To solve the first challenge, we need a differentiable signal that represents the model's ranking judgment. Inspired by the insights from FIRST (Reddy et al., 2024), which showed that the logit distribution of the first generated token can serve as a powerful proxy for the model's preference over the entire list, we adopt this technique.

We extract the logits corresponding to each document identifier (e.g., "[doc_1]', "[doc_2]') from this initial token's distribution, forming a vector of preference scores $Z = \{z_1, \ldots, z_N\}$. To ensure these scores are meaningful, we first supervise them to be accurate using a pairwise **Judgment Alignment Loss**, L_{align}:

$$\mathcal{L}_{align} = \sum_{r_i < r_j} \frac{1}{i+j} \log(1 + \exp(z_i - z_j)) \tag{4}$$

where the sum is over all ground-truth pairs where document i is more relevant than j. This loss effectively teaches the model to use the first-token logits to express a correct internal assessment of the document list.

Consistent Rank Correlation. Having obtained a differentiable ranking signal Z, we now address the second challenge: the non-differentiability of the Kendall's Tau metric itself. We formulate a differentiable variant, τ_d, that replaces the implicit, non-differentiable sign function used for comparing pairs with the smooth, differentiable hyperbolic tangent ("tanh") function.

This allows us to create a **Judgment Consistency Loss**, $L_{consist}$, that maximizes the correlation between the judgment from a neutral prompt ($Z_{neutral}$) and a diverse prompt (Z_{prompt}):

$$L_{kt} = -\tau_d(Z_{neutral}, Z_{prompt}) \tag{5}$$

where $\tau_d(Z_1, Z_2) = \frac{1}{C_N^2} \sum_{i<j} \tanh\big(k(z_{i,1} - z_{j,1})\big) \cdot \tanh\big(k(z_{i,2} - z_{j,2})\big)$. k is a scaling factor controlling the sharpness of the approximation.

4.2.1 Combined Internal Judgment Loss

Finally, we combine the alignment and consistency losses into a single loss term, $L_{consist}$, that holistically supervises the model's internal judgment for both accuracy and consistency:

$$L_{consist} = \alpha L_{align} + \beta L_{kt} \tag{6}$$

where α and β are balancing hyperparameters.

4.3 Final Training Objective

The complete CORE framework is then trained end-to-end by uniting the external behavior calibration loss (L_{rank}) and the internal judgment consistency loss ($L_{consist}$). The hyperparameters α and β effectively control the balance between all three underlying loss components. The final objective is a straightforward sum:

$$L_{CORE} = L_{rank} + L_{consist} \tag{7}$$

Both mechanisms are active only during training. At inference, the model uses a standard single forward pass, without additional computational cost.

5 Experiments

5.1 Experiment Setup

Datasets. For fine-tuning, we use ∼40k GPT-4 labeled instances created from 5k queries sampled from MS MARCO (Nguyen et al., 2016), following the setup of (Pradeep et al., 2023b). Each training instance consists of a query and a variable number ($\leq$ 20) of candidate passages that need to be reranked. These automatically labeled pairs serve as supervision to align the model's internal preference signal with ground-truth relevance.

For evaluation, we adopt two categories of benchmarks. First, the TREC Deep Learning tracks (MS MARCO passages), including DL19 (Craswell et al., 2020), DL20 (Craswell et al., 2021) (MS MARCO v1), and DL21 (Craswell et al., 2025a), DL22 (Craswell et al., 2025b) (MS MARCO v2), which are widely used for listwise reranking and allow direct comparison with prior work. Second, we consider a diverse subset of BEIR (Thakur et al., 2021) tasks to assess cross-domain robustness and prompt sensitivity, covering `climate-fever`, `dbpedia-entity`, `fever`, `fiqa`, `hotpotqa`, `nfcorpus`, `scidocs`, `scifact`, and `trec-covid`. Unless otherwise specified, we rerank the top-100 documents retrieved by a first-stage retriever for each query.

Evaluation Metrics. We report **nDCG@10** (Järvelin and Kekäläinen, 2002) as the primary evaluation metric, following common practice in listwise reranking. Since our focus is on the second-stage reranking setting, we always rerank the top-100 documents retrieved by a first-stage retriever and thus do not report retrieval-oriented metrics such as MAP@100. Because large language models exhibit inherent stochasticity and instability, we evaluate each model across multiple runs with temperature fixed at zero, and report the average performance. This procedure ensures fair and stable comparison.

Implementation. We instantiate CORE on a decoder-only LLM and fine-tune it with our CORE method. Training uses mixed precision and gradient accumulation. Our baseline models, denoted as "RankX" (e.g., RankZephyr, RankQwen), are standard supervised fine-tuning (SFT) implementations that follow the methodology of RankZephyr (Pradeep et al., 2023b). Models fine-tuned with our approach are denoted as "CORE_X" (e.g., CORE_Qwen). Unless otherwise specified, the base model for both baseline and CORE-finetuned variants is Qwen2.5-3B. For sliding-window listwise decoding, we adopt a window size of 20 and a stride of 10, a setup comparable to prior work (Sun et al., 2023a; Pradeep et al., 2023a,b). At inference time, all models, including CORE, follow the standard **autoregressive decoding process** to ensure a fair comparison. To minimize instability, we unify all evaluations under a single **neutral prompt** ($p_{neutral}$), rather than varying prompt templates. We set the maximum context length to 8192 tokens; when the combined input exceeds this limit, we truncate the input to fit within the window. More specific training details and prompt templates can be seen in the appendix A.1 and appendix A.2.

Baselines. We compare CORE against a comprehensive set of baselines to validate its effectiveness. The comparison includes standard retrievers (**BM25** and **SPLADE++ ED**) to establish a performance floor. Our primary competitors are state-of-the-art open-source listwise rerankers, which represent the direct supervised fine-tuning (SFT) counterparts to our method: **RankVicuna** (Pradeep et al., 2023a), **RankZephyr** (Pradeep et al., 2023b), and **RankQwen**, which we implemented by applying the RankZephyr methodology to the Qwen base model. To situate our work in the broader landscape, we also include the powerful proprietary model **RankGPT**$_4$ (Sun et al., 2023a).

5.2 Overall Performance

To assess the overall effectiveness of our proposed CORE framework, we first evaluate its performance on the widely-used TREC Deep Learning tracks (DL19–DL22) and a diverse set of BEIR tasks for cross-domain generalization.

Results on TREC Deep Learning Tracks. As shown in Table 2, our CORE-finetuned models demonstrate superior performance over existing state-of-the-art open-source listwise rerankers. Specifically, **CORE_Zephyr** achieves an average nDCG@10 of **0.752**, surpassing its SFT counterpart RankZephyr (0.738). Our best-performing model, **CORE_Qwen**, further improves the average performance to **0.765**, outperforming the highly competitive RankQwen baseline (0.755). These consistent improvements across most individual tracks highlight CORE's ability to enhance the core ranking effectiveness of LLMs.

Dataset	Contriever	RankVicuna	RankZephyr	RankQwen	CORE_Qwen
Climate-FEVER	0.237	**0.282**	0.256	0.234	0.223
DBPedia	0.413	0.500	0.500	0.508	**0.512**
FEVER	0.758	0.810	0.801	**0.831**	0.830
FiQA	0.329	0.359	0.422	0.462	**0.478**
HotpotQA	0.638	0.735	0.716	0.740	**0.761**
NFCorpus	0.328	0.331	**0.427**	0.384	0.390
SciDocs	0.165	0.184	**0.377**	0.192	0.208
SciFact	0.677	0.705	0.656	0.768	**0.783**
TREC-COVID	0.596	0.713	0.784	0.879	**0.885**
Average	0.460	0.513	0.549	0.555	**0.563**

Table 1: Overall Results on BEIR tasks. We report nDCG@10 on the top-100 documents retrieved by Contriever. CORE_Qwen achieves the highest average score, demonstrating strong cross-domain generalization. Best scores are in **bold**, second-best are underlined.

Method	DL19	DL20	DL21	DL22	Average
BM25	0.506	0.480	0.446	0.269	0.425
SPLADE++ED	0.731	0.720	0.685	0.571	0.676
RankGPT_4	0.746	0.708	0.772	0.718	0.736
RankVicuna	0.746	0.747	0.701	0.582	0.694
RankZephyr	0.744	0.762	0.750	0.696	0.738
RankQwen	0.765	0.774	0.753	0.710	0.755
CORE_Zephyr	**0.774**	0.781	0.740	0.712	0.752
CORE_Qwen	0.764	**0.805**	**0.770**	**0.723**	**0.765**

Table 2: Overall Results on TREC DL Tracks (DL19–DL22). Metric is nDCG@10 on top-100 candidates. CORE-finetuned models consistently outperform their SFT counterparts. Best scores are in **bold**, second-best are underlined.

Backbone	**Method**	**Mean**	**Std Dev**	**Spread**
LLaMA2-7B	SFT	0.691	0.068	0.155
	CORE	**0.741**	**0.012**	**0.028**
Qwen2.5-3B	SFT	0.739	0.018	0.039
	CORE	**0.761**	**0.005**	**0.007**

Table 3: Summary of prompt robustness on **LLaMA2-7B** and **Qwen2.5-3B**. We report the mean, standard deviation (Std Dev), and performance spread of nDCG@10 scores across four prompts. CORE significantly reduces variance and improves the average score.

Results on BEIR Cross-Domain Tasks. To evaluate generalization capabilities, we test our models on nine diverse tasks from the BEIR benchmark. The results in Table 1 show that **CORE_Qwen** achieves the highest average nDCG@10 score of **0.562**, outperforming strong baselines like RankQwen (0.555) and RankZephyr (0.549). The performance gains are particularly significant on challenging domains such as FiQA and SciFact. This demonstrates that the robust ranking preferences learned through CORE translate well to a wide variety of domains, showcasing its strong generalization ability.

5.3 Robustness to Prompt Variations

The central claim of our work is that CORE can mitigate prompt sensitivity. To verify this, we compare models trained with CORE against standard SFT across four semantically equivalent but stylistically different prompts.

The results, presented visually in Figure 3 and summarized in Table 3, provide strong evidence for our claim. On both LLaMA2-7B and Qwen2.5-3B, the standard SFT model exhibits high performance variance, with scores fluctuating dramatically depending on the prompt. In stark contrast, the CORE-trained models show remarkable stability. For instance, on LLaMA2-7B, CORE reduces the performance spread (max-min difference) from a substantial 0.155 to just 0.028, while also improving the average score. This strongly demonstrates that CORE successfully learns a more robust, prompt-invariant ranking function.

5.4 Effect of CORE Components

To understand the individual contributions of CORE's key components, we conduct a thorough ablation study across three distinct model backbones of varying sizes: Qwen2.5-0.5B, Qwen2.5-1.5B, and Qwen2.5-3B. The consolidated results are presented in Table 4.

A consistent trend emerges from the results: the full CORE framework consistently yields the best average performance across all model sizes. Removing the internal consistency loss (*w/o* $L_{consist}$) generally leads to a drop in performance, demonstrating the benefit of regularizing the model's internal judgment. A further degradation typically occurs when both the consistency loss and the in-

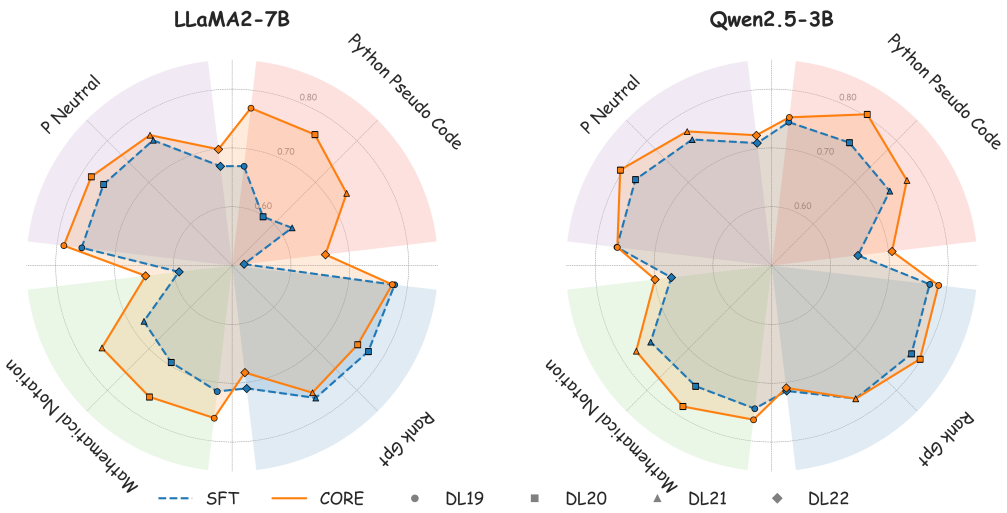

Figure 3: Performance of SFT vs. CORE on LLaMA2-7B and Qwen2.5-3B across four different prompt templates on TREC DL datasets. The radar plots visually demonstrate CORE's key advantage: its performance (solid orange line) is consistently high across all prompts, forming a larger and more regular shape compared to the erratic performance of standard SFT (dashed blue line).

Backbone	Method	DL19	DL20	DL21	DL22	Avg
Qwen2.5-0.5B	CORE	0.738	0.758	0.722	0.610	**0.707**
	w/o $L_{consist}$	0.731	0.731	0.700	0.609	0.693
	w/o CFG & $L_{consist}$ (SFT)	0.731	0.721	0.685	0.571	0.677
Qwen2.5-1.5B	CORE	0.759	0.786	0.746	0.708	**0.750**
	w/o $L_{consist}$	0.760	0.786	0.747	0.695	0.747
	w/o CFG & $L_{consist}$ (SFT)	0.768	0.764	0.750	0.700	0.746
Qwen2.5-3B	CORE	0.771	0.762	0.771	0.734	**0.759**
	w/o $L_{consist}$	0.769	0.771	0.760	0.728	0.757
	w/o CFG & $L_{consist}$ (SFT)	0.765	0.774	0.753	0.710	0.751

Table 4: Ablation study of CORE components across three different model backbones. We report nDCG@10 on TREC DL datasets. The results show that the full CORE framework consistently achieves the best performance. Best average scores for each backbone are in **bold**.

verse CFG mechanism are removed, which reduces the model to a standard Supervised Fine-Tuning (SFT) baseline (*w/o CFG & $L_{consist}$*).

Interestingly, the magnitude of the improvement varies with model scale. The impact of the CORE components is most pronounced on the 0.5B model, while the performance differences are more subtle on the 1.5B model. Nevertheless, the full CORE configuration remains the most effective or tied for the best across all tested backbones. These comprehensive results confirm that both the external calibration and the internal consistency regularizer are complementary components for enhancing ranking performance. CORE proves especially beneficial for smaller, more volatile models, underscoring its potential for building robust and efficient rerankers in resource-constrained scenarios.

6 Conclusion

In this work, we addressed the critical challenge of prompt sensitivity in LLM-based rerankers. We introduced CORE, a novel fine-tuning framework that stabilizes a model's intrinsic ranking preferences. CORE employs a dual strategy, combining an inverse CFG mechanism for external behavior calibration with a differentiable Kendall's Tau loss for internal judgment consistency. Experiments on TREC DL and BEIR benchmarks confirm that CORE achieves state-of-the-art performance and yields significantly more robust and stable rankings across diverse prompts. Our work represents a key step towards more reliable LLM systems, and the proposed framework offers a promising direction for mitigating input sensitivity in other text generation tasks.

7 Limitations

While CORE enhances robustness, it has limitations. First, training requires dual-path processing (neutral and diverse prompts), increasing computational overhead compared to standard SFT, though inference cost remains unchanged. Second, our internal consistency loss relies on the "first-token proxy" assumption, which may be less effective for tasks requiring complex multi-step reasoning. Finally, like many listwise rerankers, we currently rely on sliding windows for long candidate lists, warranting future exploration into long-context capabilities.

References

Negar Arabzadeh and Charles L. A. Clarke. 2025. A human-ai comparative analysis of prompt sensitivity in llm-based relevance judgment. *CoRR*, abs/2504.12408.

Anwoy Chatterjee, H. S. V. N. S. Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A prompt sensitivity index for large language models. In *EMNLP Findings*, pages 14550–14565.

Yiqun Chen, Qi Liu, Yi Zhang, Weiwei Sun, Xinyu Ma, Wei Yang, Daiting Shi, Jiaxin Mao, and Dawei Yin. 2025. Tourrank: Utilizing large language models for documents ranking with a tournament-inspired strategy. In *WWW*.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. 2025. CFG++: manifold-constrained classifier free guidance for diffusion models. In *ICLR*.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. *CoRR*, abs/2102.07662.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2025a. Overview of the TREC 2021 deep learning track. *CoRR*, abs/2507.08191.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2025b. Overview of the TREC 2022 deep learning track. *CoRR*, abs/2507.10865.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. *CoRR*, abs/2003.07820.

Yongqi Fan, Xiaoyang Chen, Dezhi Ye, Jie Liu, Haijin Liang, Jin Ma, Ben He, Yingfei Sun, and Tong Ruan. 2025. Tfrank: Think-free reasoning enables practical pointwise LLM ranking. *CoRR*, abs/2508.09539.

Vincent Fortuin. 2021. Priors in bayesian deep learning: A review. *CoRR*, abs/2105.06868.

Sigmund Freud. 1923. *The Ego and the Id*. Internationaler Psychoanalytischer Verlag.

Jingtong Gao, Bo Chen, Xiangyu Zhao, Weiwen Liu, Xiangyang Li, Yichao Wang, Wanyu Wang, Huifeng Guo, and Ruiming Tang. 2025. Llm4rerank: Llm-based auto-reranking framework for recommendations. In *WWW*, pages 228–239.

Yuchen Guan, Runxi Cheng, Kang Liu, and Chun Yuan. 2024. Kendall's τ coefficient for logits distillation. *CoRR*, abs/2409.17823.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598.

Chi Hu, Yuan Ge, Xiangnan Ma, Hang Cao, Qiang Li, Yonghua Yang, Tong Xiao, and Jingbo Zhu. 2024. Rankprompt: Step-by-step comparisons make language models better reasoners. In *COLING*, pages 13524–13536.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

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

Qi Liu, Bo Wang, Nan Wang, and Jiaxin Mao. 2025. Leveraging passage embeddings for efficient listwise reranking with large language models. In *WWW*, pages 4274–4283.

Kehan Long, Shasha Li, Chen Xu, Jintao Tang, and Ting Wang. 2025. Precise zero-shot pointwise ranking with llms through post-aggregated global context information. *CoRR*, abs/2506.10859.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *NIPS*.

Lilian Ngweta, Kiran Kate, Jason Tsay, and Yara Rizk. 2025. Towards llms robustness to changes in prompt format styles. In *NAACL*.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023a. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *CoRR*, abs/2309.15088.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023b. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *CoRR*, abs/2312.02724.

Yao Qiang, Subhrangshu Nandi, Ninareh Mehrabi, Greg Ver Steeg, Anoop Kumar, Anna Rumshisky, and Aram Galstyan. 2024. Prompt perturbation consistency learning for robust language models. In *EACL Findings*, pages 1357–1370.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *NAACL Findings*, pages 1504–1518.

Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md. Arafat Sultan, Deevya Swain, Avirup Sil, and Heng Ji. 2024. FIRST: faster improved listwise reranking with single token decoding. In *EMNLP*.

Ruiyang Ren, Yuhao Wang, Kun Zhou, Wayne Xin Zhao, Wenjie Wang, Jing Liu, Ji-Rong Wen, and Tat-Seng Chua. 2025. Self-calibrated listwise reranking with large language models. In *WWW*, pages 3692–3701.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *EMNLP*.

Dylan Sam, Rattana Pukdee, Daniel P. Jeong, Yewon Byun, and J. Zico Kolter. 2024. Bayesian neural networks with domain knowledge priors. *CoRR*, abs/2402.13410.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024a. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *ICLR*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024b. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *ICLR*.

Shuoqi Sun, Shengyao Zhuang, Shuai Wang, and Guido Zuccon. 2025. An investigation of prompt variations for zero-shot llm-based rankers. In *ECIR*, volume 15573, pages 185–201.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023a. Is chatgpt good at search? investigating large language models as re-ranking agents. In *EMNLP*, pages 14918–14937.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023b. Is chatgpt good at search? investigating large language models as re-ranking agents. In *EMNLP*, pages 14918–14937.

Raphael Tang, Xinyu Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2024. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *NAACL*, pages 2327–2340.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS*.

Jim Waldo and Soline Boussard. 2024. Gpts and hallucination: Why do large language models hallucinate? *ACM Queue*, 22(4):10.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *ICLR*.

Chenxing Wei, Yao Shu, Mingwen Ou, Ying Tiffany He, and Fei Richard Yu. 2025. PAFT: prompt-agnostic fine-tuning. *CoRR*, abs/2502.12859.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *ICML*, pages 12697–12706.

Kaipeng Zheng, Huishuai Zhang, and Weiran Huang. 2023. Diffkendall: A novel approach for few-shot learning with differentiable kendall's rank correlation. In *NeurIPS*.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine A. Heller, and Subhrajit Roy. 2024. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *ICLR*.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2024a. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. In *NAACL*.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. Rankt5: Fine-tuning T5 for text ranking with ranking losses. In *SIGIR*, pages 2308–2313.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024b. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *SIGIR*.

A Appendix

A.1 Implementation Details

Training. We fine-tuned on the RankZephyr dataset (39k instances) using a single NVIDIA A40 (48GB) with DeepSpeed ZeRO-3. We utilized AdamW (lr=**5e-6**, cosine decay, **50** warmup steps), a global batch size of 32 (per-device 2, grad accumulation 16), and a fixed seed **42**.

Inference. We follow a standard two-stage pipeline. Candidates (top-100) are retrieved via SPLADE++ (EnsembleDistil). The LLM reranker processes these using a sliding window (size 20, stride 10) to produce a final top-20 list (Pradeep et al., 2023b).

A.2 Prompt Templates

Our robustness experiments utilized three **diverse prompts** ($\mathcal{P}_{train}$), representing standard instructions versus conversational, code-based, and mathematical styles, respectively. The specific templates are illustrated in Figure 4.

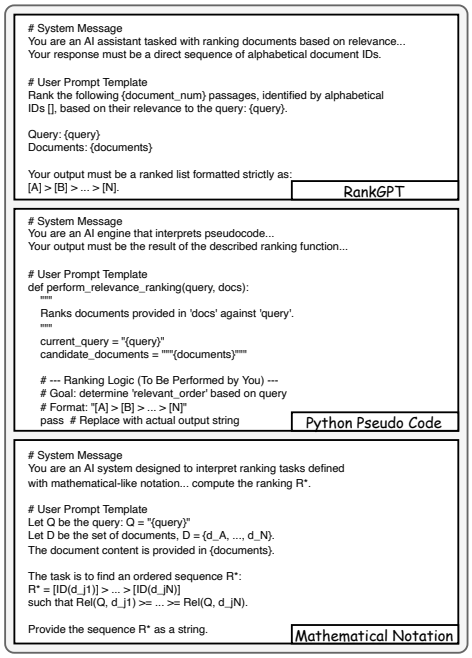

Figure 4: System messages and user prompts. Placeholders are filled at runtime.

A.3 Theoretical Analysis of Inverse CFG

We theoretically show that interpolating diverse logits with a stable neutral prompt strictly decreases the mis-selection probability bound.

Theorem A.1 (Shrinkage Calibration Reduces Mis--selection Probability)**.** *Assume there exists an intrinsic (prompt-invariant) logit vector $l^\star$. Let the observed logits from the diverse prompt and the neutral prompt be:*

$$l_{\text{prm}} = l^\star + \varepsilon, \qquad l_{\text{neu}} = l^\star + \delta,$$

where ε, δ are independent noise vectors with $\mathbb{E}[\varepsilon] = \mathbb{E}[\delta] = 0$. Let y be the index of the correct document and b be the index of the strongest competitor (incorrect document). Define the intrinsic margin $m^\star := l^\star_y - l^\star_b > 0$.

Assume the margin noises $X := \varepsilon_y - \varepsilon_b$ and $Y := \delta_y - \delta_b$ are zero-mean sub-Gaussian random variables with variance parameters σ_p^2 and σ_n^2, respectively.

Consider the calibrated logits l_g defined by the Inverse CFG mechanism:

$$\begin{aligned} l_g &= l_{\text{neu}} + w(l_{\text{prm}} - l_{\text{neu}}) \\ &= (1-w)l_{\text{neu}} + w l_{\text{prm}}, \quad w \in [0,1]. \end{aligned}$$

Let the combined variance be $\sigma_S^2(w) = (1-w)^2\sigma_n^2 + w^2\sigma_p^2$. The mis-selection probability is bounded by:

$$\mathbb{P}[(l_g)_y \le (l_g)_b] \le \exp\left(-\frac{(m^\star)^2}{2\sigma_S^2(w)}\right).$$

Proof. Let $m_g := (l_g)_y - (l_g)_b$ be the calibrated margin. The mis-selection event occurs when $m_g \le 0$. Substituting the expression for l_g, we derive the margin as a convex combination of the noisy margins:

$$m_g = (1-w)(m^\star + Y) + w(m^\star + X) = m^\star + S,$$

where $S := (1-w)Y + wX$ represents the combined noise. Since X and Y are independent sub-Gaussian variables, their linear combination S is also sub-Gaussian with parameter $\sigma_S^2 = (1-w)^2\sigma_n^2 + w^2\sigma_p^2$.

Applying the Chernoff bound:

$$\mathbb{P}[m_g \le 0] = \mathbb{P}[S \le -m^\star] \le \exp\left(-\frac{(m^\star)^2}{2\sigma_S^2}\right).$$

Substituting σ_S^2 yields the stated theorem. □

Remark. The bound is minimized when the combined variance $\sigma_S^2(w)$ is minimized. As long as the neutral prompt provides a non-trivial signal, mixing it with the diverse prompt ($w < 1$) yields a lower error bound than using the diverse prompt alone ($w = 1$), confirming the effectiveness of our calibration strategy.

A.4 Statement on AI Usage

We utilized LLMs solely for copy-editing, code debugging, and formatting assistance. All conceptual contributions, experimental designs, and analyses are original to the authors.