What Makes an Ideal Quote? Recommending "Unexpected yet Rational" Quotations via Novelty

Powei Chang[1], Jin Xiao[1], Guanglei Yue[1]
Qianyu He[2], Yanghua Xiao[2], Deqing Yang[1], Jiaqing Liang[1]*
[1]School of Data Science, Fudan University,
[2]Shanghai Key Laboratory of Data Science,
College of Computer Science and Artificial Intelligence, Fudan University
{bwzhang24, jinxiao23, glle24, qyhe21}@m.fudan.edu.cn,
{shawyh, yangdeqing, liangjiaqing}@fudan.edu.cn

Abstract

Quotation recommendation enriches writing by suggesting quotations that fit a given context, but prior systems largely focus on topical relevance and overlook what makes quotes memorable. Based on a user study, we find that preferred quotations are often *unexpected yet rational*, motivating the goal of selecting quotes that are contextually novel while semantically coherent. We propose NOVELQR, which (1) uses a generative label agent to map quotations and contexts into multi-dimensional deep-meaning labels for label-enhanced retrieval, and (2) reranks candidates with a token-level novelty estimator that mitigates auto-regressive continuation bias. Experiments on bilingual datasets across diverse domains show that NOVELQR is preferred by human judges and improves overall recommendation quality over strong baselines, while achieving competitive novelty estimation. (Code: Github link)

Figure 1: An ideal quote should not only fit the context, but also be novel, adding aesthetic value to writing. As shown in the third example, the best quote often feels unexpected at first, but makes perfect sense in context.

1 Introduction

> *"Poetic language must appear strange and wonderful."*
>
> ——Aristotle

Famous quotations (Tan et al., 2015) play an important role in academic writing and daily communication, as they can "provide authority for arguments and enhance persuasiveness" and "add color and aesthetics to articles" (MacLaughlin et al., 2020). An appropriate quotation not only helps readers understand complex ideas more accurately, but also adds aesthetic feeling. Therefore, recommendation of high-quality quotations has become an important task in natural language generation.

This raises a fundamental question: **what makes an ideal quote?** Building on Shklovsky's *Defamiliarization theory* (Crawford, 1984), "*Art aims to renew perception by making the familiar unfamiliar, slowing down understanding and provoking reflection.*" In this sense, an ideal quotation should not merely restate a point, but challenge habitual thinking and invite deeper interpretation. Related theories in communication and linguistics, such as *Closure theory* (Kruglanski and Webster, 1996), suggest that a writing technique prompting deeper thinking and enhancing aesthetic appeal is the *unfamiliar, complex*, and *profound* content.

To examine whether users truly prefer "unfamiliar" quotations, we conduct a large-scale user study and controlled behavioral experiments. The results show that, among rationally appropriate options, participants systematically favor more novel quotations and treat novelty as a complementary dimension of quotation quality. We therefore define an ideal quote as "**unexpected yet rational**" (Figure 1): readers may feel briefly puzzled when first encountering the third recommended quote, but then experience a sudden sense of insight once

*Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7985–8017
July 2-7, 2026

they relate it to the context. Such quotations deepen the expressive power of the context while avoiding clichés and mediocrity.

With this defamiliarization- and user-study–driven view of what constitutes a high-quality quote, we revisit quotation recommendation (Tan et al., 2015, 2016, 2018; Ahn et al., 2016; Wang et al., 2021, 2022). Prior systems mostly reduce the task to semantic matching over quote text (e.g., QuoteR (Qi et al., 2022), QUILL (Xiao et al., 2025)), emphasizing surface-level rationality while overlooking deeper meanings and the "unexpected" dimension. Our analysis shows that even strong LLMs struggle to infer deep meanings from quotations in isolation, and that naive logit-based novelty metrics suffer from an *auto-regressive continuation bias*, such as surprisal (Futrell et al., 2019) and KL-divergence (Gamon, 2006). These observations motivate a formulation that (1) retrieves quotations in **a deep semantic space** reflecting their underlying intents, and (2) measures contextual novelty at the **token level** while mitigating *continuation bias*.

In summary, achieving high-quality quotation recommendation requires addressing two key challenges: (1) capturing the deep meanings and intents behind quotations, and (2) measuring novelty while mitigating *continuation bias*.

To address these challenges, we propose NOVELQR, a novelty-driven, retrieval-augmented framework for quotation recommendation. A label enhancement module first builds a deep-meaning quotation knowledge base using a generative label agent that interprets each quote into multi-dimensional labels. These labels are used to derive deep-meaning embeddings and support fine-grained hard filtering to ensure semantic rationality. Given a user context, we retrieve candidate quotations by deep-meaning similarity, then apply a token-level novelty estimator that focuses on "novelty tokens" to mitigate *continuation bias*. Finally, we integrate novelty, popularity, and matching signals into a unified scoring function to re-rank candidates. We evaluate performance on bilingual datasets spanning diverse real-world domains by combining our test sets with existing benchmarks, collecting human ratings of rationality, novelty, and engagement, and calibrating an LLM-as-judge against these ratings to enable detailed evaluation and ablations. Contributions are as follows:

- We formalize ideal recommendation as selecting quotes that are unexpected yet rational, grounded in *defamiliarization* and user studies.
- We develop NOVELQR, an end-to-end novelty-driven system with a generative label agent that constructs a deep-meaning knowledge base and enables semantic similarity retrieval with fine-grained hard filtering for rationality.
- We identify an *auto-regressive continuation bias* in logit-based novelty estimation and propose a token-level method that focuses on "novelty tokens" to substantially mitigate this bias.

2 Related Work

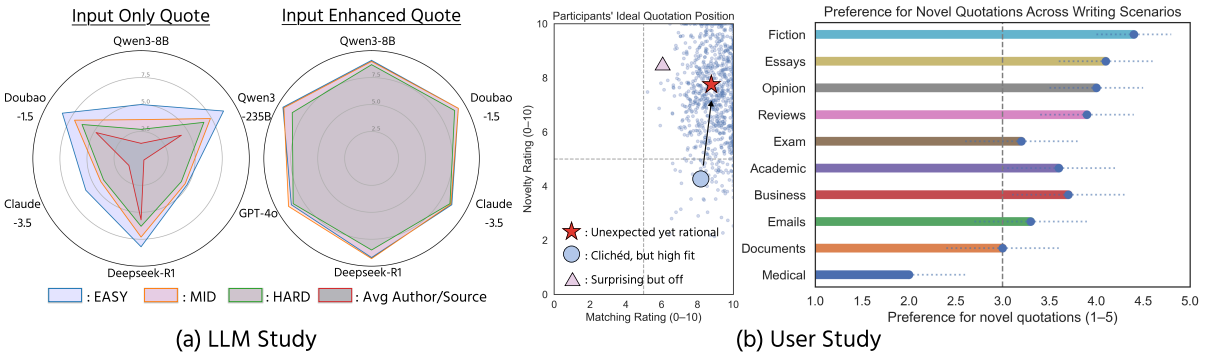
Quotation Recommendation. Work on quotation (quote) recommendation has mainly targeted semantic relevance. Early methods framed it as learning to rank with handcrafted features (Tan et al., 2015; Lee et al., 2016; Tan et al., 2016, 2018; Ahn et al., 2016; Wang et al., 2021, 2022), later replaced by neural models based on CNN/LSTM, Transformers, GRUs, and BERT. More recently, QUILL (Xiao et al., 2025) adopts a RAG-style framework and offers a comprehensive benchmark. QuoteR (Qi et al., 2022) and QUILL provide bilingual test sets, which we use together with our NOVELQR-BENCH benchmark. However, existing systems **largely optimize relevance** and **do not explicitly model the aesthetic value** or novelty of quotations.

Novelty Estimation. Textual novelty has been studied mainly from two angles: Zhang et al. (2025b) introduce NoveltyBench and view novelty as answer diversity, while Li et al. (2022); McCoy et al. (2023) describe that transformers prefer high-frequency words and reduce output diversity. For operationalizing novelty or surprise, prior work uses bayesian surprise (Pimentel et al., 2014; Futrell et al., 2019), KL divergence (Gamon, 2006), and metrics such as embedding distance (Shibayama et al., 2021). These logit-based approaches perform poorly on quote novelty from ***auto-regressive continuation bias***.

3 Empirical Study

3.1 Do LLMs truly understand quotations?

Most quotation systems either generate quotations with LLMs or retrieve them via embedding-based search, typically operating on the quotation in isolation. This leaves a central question underexplored: **to what extent do models actually grasp the deep meanings of quotations, and how can this understanding be improved?**



(a) LLM Study (b) User Study

Figure 2: **Empirical result.** (a) The evaluation results of the only-quote (left) and enhanced-quote (right) scene. All models perform significantly better with enhanced inputs, demonstrating the effectiveness of guided prompt in deep meaning understanding. (b) In user studies, (left) participants perceive ideal quotations as *"unexpected yet rational"*(★), while current models tend to produce clichéd-but-high-fit ones (●); (right) across various writing scenarios, *novelty* consistently emerges as a key dimension of quotation quality.

Setup. We construct a diagnostic evaluation over quotations from three genres (classical Chinese, modern Chinese, and modern English), each paired with expert-written interpretations of their underlying semantics. Quotations are bucketed into three difficulty bands (EASY, MID, HARD). We evaluate several closed- and open-source LLMs on two tasks: (1) explaining the deep meaning of a quotation and (2) identifying its author or source, under two prompting conditions: *quote only* versus an *enhanced quote* that includes auxiliary contextual information. Details are given in Appendix H, and evaluation results appear in Figure 2(a).

Findings. With only the quote as input, all models perform poorly at capturing deep meanings, regardless of size: even on the EASY subset, GPT-4o's average score remains below the threshold for high-quality semantic understanding, and author/source identification is similarly weak, indicating **difficulty understanding deep meanings of quotations**. By contrast, enhanced-quote prompts yield substantial gains, where average scores approach 9.0 even on HARD items, and a smaller Qwen3-8B model matches GPT-4o. These results suggest that LLMs can effectively **grasp deep meanings when supplemented with auxiliary information**. This motivates enriching the quotation knowledge base with labels before retrieval.

3.2 Do users actually want "unexpected yet rational" quotations?

To ensure that our objective is aligned with user needs, we conduct four complementary user studies (details in Appendix E).

Questionnaire. We first ran an online questionnaire with $N = 964$ respondents across diverse ages and work fields. On 0–10 scales, an "ideal" quotation is rated as almost obligatorily appropriate (9.1) and also novel (7.4), and users see these two dimensions as complementary rather than conflicting: most place their ideal quotation in the high-match, non-trivially-novel region and many are willing to trade a small amount of fit for extra novelty. Scenario questions further show that novelty is strongly valued in everyday expressive writing, and open-ended answers repeatedly describe good quotations as those that *"fit the context but still feel fresh"*.

Controlled experiments. We then ran small controlled studies with 100 participants to test these preferences in behavior. In rating, pairwise-choice, and cloze-style fill-in tasks that explicitly control contextual fit, participants consistently prefer quotations that are *novel-but-rational* over clichéd ones. After reading a short description of a defamiliarization-like effect, they describe it as exactly the kind of impact they want quotations to have in expressive writing, supporting our choice of *unexpected yet rational* as the target objective. As summarized in Figure 2 (b), users perceive an ideal quotation as **unexpected yet rational**, which emerges as one important dimension of quotation quality alongside basic appropriateness.

4 Methodology

To address the two challenges of "difficulty understanding deep meanings of quotations" and "semantically rational but lacking novelty", we propose a quotation recommendation system (Figure 3),

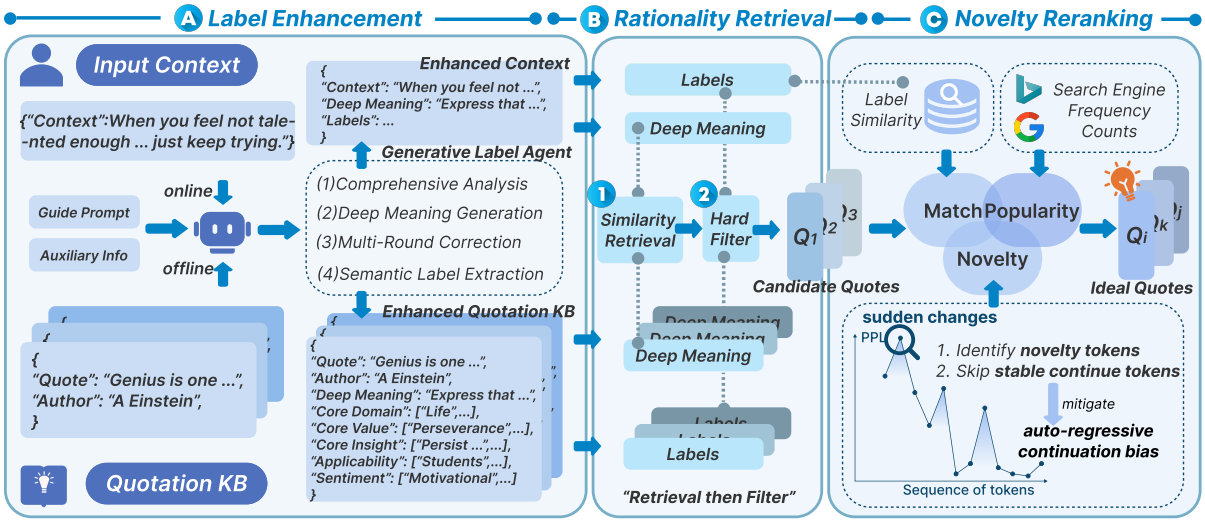


Figure 3: Overview of our novelty-driven quotation recommendation framework: (1) **Label Enhancement**, where the generative label agent enhances understanding of the quotation knowledge base (KB) and user-given context; (2) **Rationality Retrieval**, which "retrieves then filters" quotations using deep meanings and labels; and (3) **Novelty Reranking**, which highlights the continuation bias, and introduces the method to mitigate it and estimate novelty.

which consists of three steps:

4.1 Step 1: Label Enhancement

Existing quotation recommendation systems typically retrieve candidates by embedding the *raw quotation text*. However, our empirical analysis shows that even strong LLMs often fail to capture the *deep semantic meanings* of quotations from their surface strings alone, making direct retrieval over raw quotes unreliable. We therefore preprocess our quotation knowledge base (KB) with **Label Enhancement** before retrieval: **a generative label agent** produces both deep semantic interpretations and multi-dimensional labels, and performs the same procedure online for user-provided contexts. Following Section 3.1, we adopt *Qwen3-8B* (Team, 2025b) as the backbone of this agent.

As illustrated in Figure 3 (details in Appendix J), the label agent executes four steps:
(1) **Comprehensive Analysis:** Given auxiliary information (author, source, and related context), the LLM analyzes the quotation from multiple perspectives, including author background, historical–cultural context, and emotional connotations.
(2) **Deep Meaning Generation:** Based on the comprehensive analysis beyond the quotation, we extract and concisely summarize the deep semantic meanings within 50 words ("*Express that ...*").
(3) **Multi-round Correction:** The agent self-critiques and refines its explanations for up to $R = 3$ rounds, checking for superficiality, over-interpretation, and logical gaps; around 4.6% of outputs are rejected (protocol in Appendix J.2).
(4) **Semantic Label Extraction:** Finally, structured semantic labels are extracted from Core Domains, Insights, Values, Applicability, and Sentiment Tone (Prompts in Appendix N.2).

Through label enhancement, quotations in the KB are mapped into an interpretable deep-meaning space equipped with rich labels, forming the **basis** for our label-enhanced retrieval module. Retrieving over these interpretations, rather than raw quotation embeddings, yields **more rational and controllable recommendations**.

4.2 Step 2: Rationality Retrieval

Traditional systems usually retrieve quotations by measuring similarity over the *raw quotation text*, which we refer to as **Quote-based Retrieval (QR)**. While embedding-based similarity over surface strings can work for simple, explicit quotations, it often returns candidates that are only superficially related but misaligned with the deeper intent of the context. Moreover, QR skips the interpretive step and does not mimic how humans first analyze a context before choosing an appropriate quotation.

Building on label enhancement, we instead perform retrieval over *deep semantic meanings*, which act as a bridge for semantic retrieval. We term this module **Label-enhanced Retrieval (LR)**. LR follows a "**retrieve–then–filter**" pipeline.

In the retrieval step, an embedding model encodes the deep meanings of all quotations, as well as the input context, and we retrieve $TopN$ can-

didates with the highest similarity in this deep-meaning space. We then apply a hard filter based on label similarity in the "Core Domain/Value/Insight" dimensions with a threshold T. Human verification shows that our generated labels have less than 3% distortion (Appendix J.3), so this signal allows the filter to reliably remove semantically implausible candidates while rarely discarding valid ones.

We tune on a held-out validation set and use $TopN = 50, T = 0.7$ in all experiments (Appendix B). However, the goal of this stage is not to produce the final recommendation, but to **construct a pool of semantically rational candidates** for the subsequent novelty-aware reranker.

4.3 Step 3: Novelty Reranking

Given a candidate pool that is largely rational with respect to the context, the final stage focuses on ranking quotations by their degree of *"unexpectedness"* while mitigating *auto-regressive continuation bias* in standard surprisal-style scores. We combine three factors: **novelty** S_N, **semantic match** S_M, and **popularity** S_P.

Novelty. We first define quotation novelty. Intuitively, a novel quotation is unfamiliar and difficult for the model to predict under the given context (Futrell et al., 2019). We therefore measure novelty via differences in the model's own logits. For a candidate quotation $q = \{x_1, \ldots, x_T\}$, let

$$p_{\text{prior}}(x_t) = p(x_t|x_1, \cdots, x_{t-1}) = p(x_t|X_{<t}), \quad (1)$$

$$p_{\text{cond}}(x_t) = p(x_t|C, x_1, \cdots, x_{t-1}) = p(x_t|C, X_{<t}), \quad (2)$$

be the token distributions without and with the external context c, respectively. We define the log-probability difference R_t and compute **online**,

$$R_t = \log p_{\text{prior}}(x_t) - \log p_{\text{cond}}(x_t). \quad (3)$$

If $R_t > 0$, the token becomes harder to predict under the context, reflecting the kind of *"sudden turn"* or *"surprise"* that we seek.

However, standard logit-based novelty estimators can exhibit errors that stem from what we term **auto-regressive continuation bias**[1]. In other words, since the model performs inference through next word prediction, some common expressions exhibit continuity problems. For example, after given context "*When you feel not talented enough to finish this project. Don't worry about that, just keep trying*", it is difficult to predict "*Genius is one percent*" in the beginning, whereas predicting the subsequent phrase "*inspiration and ninety-nine percent perspiration*" becomes inevitable. If we **predict at the word-level or quote-level**, it will **cause bias** in the final average calculation. To mitigate this, we model quotation novelty at the token level and emphasize **novelty tokens** rather than treating all tokens uniformly (the bias is illustrated in Figure 14).

Concretely, we first run the quotation through the language model without context and compute a token-level self-perplexity sequence $\text{PPL}_t = \exp\big(-\log p(x_t \mid x_{<t})\big)$ **offline**. We then examine how this sequence evolves by taking first- and second-order differences:

$$\delta_1(t) = \text{PPL}_t - \text{PPL}_{t-1}, \; |\delta_2(t)| = |\delta_1(t) - \delta_1(t-1)|. \quad (4)$$

Large $|\delta_2(t)|$ indicates a sudden change in the local Self-PPL pattern (Xie et al., 2024; Shin et al., 2024). To obtain a smooth, non-negative signal, we define

$$\Delta_2(t) = \log\big(1 + |\delta_2^{\text{pad}}(t)|\big), \quad (5)$$

where $\delta_2^{\text{pad}}(t)$ denotes $\delta_2(t)$ with appropriate padding at boundaries. We then normalize $\Delta_2(t)$ within each quotation to obtain weights in $[0, 1]$:

$$w_t = \frac{\Delta_2(t) - \min_t \Delta_2(t)}{\max_t \Delta_2(t) - \min_t \Delta_2(t) + \epsilon} \in [0, 1], \quad (6)$$

where ϵ is a small constant to avoid division by zero and convert $\{w_t\}$ into a distribution over tokens,

$$\tilde{w}_t = \frac{w_t}{\sum_{j=1}^{T} w_j}, \quad (7)$$

so that $\sum_t \tilde{w}_t = 1$. Tokens with large $\tilde{w}_t$ are treated as novelty tokens, while smooth continuation regions receive little weight.

Finally, we define the token-level novelty score as a weighted average of log-probability differences:

$$S_N = \sum_{t=1}^{T} \tilde{w}_t \Big[\log p(x_t \mid x_{<t}) - \log p(x_t \mid C, x_{<t})\Big]. \quad (8)$$

Positive contributions to S_N come mainly from novelty tokens whose predictability drops under the context, while continuation-like segments contribute little, thereby reducing bias.

Popularity. To avoid spuriously treating extremely rare quotations as "novel", we add a web-based popularity signal. For each quote q we query

[1] See Appendix L for a detailed discussion of continuation bias, our novelty-token design, and the bias analysis from other novelty estimators.

Bing[2] with the exact-phrase query under a depersonalized, region-neutral profile and record the count c. Counts are collected at fixed UTC snapshots (2025.02-04) and reported in KB for reproducibility. We then map c to a bounded score

$$S_P = \frac{1}{1+e^{-z}}, \text{where } z = \frac{log(1+c)-\mu}{\sigma}, \quad (9)$$

where $\mu = 10.53$ and $\sigma = 2.21$ are estimated from $log(1+c)$ over all quotations. It is used as a regularizer to **downweight overly obscure candidates**. Appendix C shows that removing popularity yields a consistent drop, and further analyzes **alignment with human-perceived familiarity** (Spearman $\rho \approx 0.73$, Fleiss' $\kappa = 0.68$) as well as sensitivity to alternative engines (e.g. **Google**).

Semantic Match. Although LR already enforces contextual rationality, we still include a semantic matching term to favor quotations that are more coherent and emotionally consistent with the context. We compute cosine similarity between deep-meaning embeddings of the quotation and context, and rescale it to $[0, 1]$:

$$S_M = \frac{1}{2}\left(\frac{\mathbf{h}_q \cdot \mathbf{h}_c}{\|\mathbf{h}_q\| \, \|\mathbf{h}_c\|} + 1\right), \quad (10)$$

$\mathbf{h}_q, \mathbf{h}_c$ denote the semantic embeddings of the quotation and context, respectively.

Finally, the reranking score is as follows:

$$S_{final} = \lambda_1 \cdot S_N + \lambda_2 \cdot S_P + \lambda_3 \cdot S_M. \quad (11)$$

By adjusting λ_i, we balance novelty against rationality factors to ensure that the recommended quotations are both surprising and acceptable.

Computational cost. The heavy components are run offline, so that online inference only requires embedding similarity searches and logit-difference, with an average end-to-end latency of about $772.2^{+431.3}_{-30.5}$ ms per query (Appendix I).

5 Experiments

In this section, we aim to answer three questions: (1) whether NOVELQR **improves quote recommendation** over strong baselines, (2) whether label-enhanced retrieval yields a **more semantically coherent** candidate set than text-based retrieval, and (3) whether token-level novelty estimation **better captures contextual novelty** by mitigating continuation bias. We further examine the **consistency** between our evaluation and human judgments.

[2]https://www.bing.com/

5.1 Setup

Datasets. We evaluate on three high-quality bilingual test sets of 100 instances each: QuoteR, QUILL, and our proposed test set NOVELQR-BENCH. Together, these sets cover literary, conversational, and expository writing across diverse real-world domains (e.g. literature, science, philosophy, law, etc.). Construction details and statistics are given in Appendix D.

Knowledge Base. The quotations in the knowledge base are from QUILL (Xiao et al., 2025), which we richly label and embed using the *ACGE text embedding* model (Kusupati et al., 2022).

Metrics. Our primary retrieval metrics (HR@5, nDCG@5, MRR@5; †Statistical significance paired bootstrap testing details in Appendix K) are computed from human-annotated labels (Appendix F.1). In addition, to obtain 1–5 auxiliary scores for Match and Novelty at scale, which are averaged over three random seeds for stability, we use an LLM-as-judge (GPT-4o (OpenAI, 2024)) calibrating against expert ratings (Section 5.4).

Settings. All methods share the same hyperparameters. Label-enhanced retrieval uses $TopN = 50$ and $T = 0.7$ (Appendix B), and the reranking weights are fixed to $\{\lambda_1 = 0.70, \lambda_2 = 0.20, \lambda_3 = 0.10\}$ which tuned on a held-out set over different weight combinations (see Appendix A, Table 4).

5.2 Main Result

As shown in Table 1, *Model-based Quotation Generation* baselines perform worst, mainly due to hallucinations (Xiao et al., 2025) and low appropriateness. *Retrieval-augmented Quotation Recommendation* methods perform substantially better, especially those based on semantic matching. Within this family, moving from *Quote-based Retrieval* (QR+w/oReranker) to our *Label-enhanced Retrieval* (LR+w/oReranker) yields a large gain in Match (from 3.99 to 4.55), indicating that the first-stage rationality retrieval provides a much **stronger candidate set** for subsequent reranking.

Fixing LR as the retriever, we then compare different rerankers. Across all test sets, our novelty-aware reranker (LR+Ours) **achieves the best overall performance**, substantially boosting novelty while maintaining high match and strong ranking metrics. Additionally, in a human multiple-choice study (Appendix E.2), **78%** of selections favor our system. Improvements hold for classical literary quotations, modern conversational contexts, and

Method	QuoteR			QUILL			NovelQR-Bench					
	Novelty	Match	Avg	Novelty	Match	Avg	Novelty	Match	Avg	HR†	nDCG†	MRR†
Model-based Quotation Generation												
LLM (GPT-based)	2.85	3.00	2.93	2.76	3.10	2.93	2.85	2.99	2.92	∼	∼	∼
QuoteR (Bert-based)	3.55	3.77	3.66	3.55	4.08	3.82	3.21	3.88	3.54	∼	∼	∼
Retrieval-augmented Quotation Recommendation												
QR + w/o Reranker	3.59	3.93	3.76	3.46	4.04	3.80	3.14	3.99	3.57	0.35	0.26	0.24
QUILL	3.42	3.90	3.66	3.32	4.11	3.72	3.08	4.15	3.62	0.15	0.12	0.11
LR + w/o Reranker	3.78	3.96	3.87	3.63	4.26	3.95	3.40	4.55	3.98	0.55	0.44	0.40
LR + bm25	3.64	3.95	3.80	3.60	4.30	3.98	3.40	4.52	3.96	0.40	0.30	0.23
LR + Bge-large	3.75	**4.00**	3.88	3.60	4.33	3.97	3.61	4.54	4.08	0.56	0.39	0.33
LR + Qwen3-Re	3.85	3.90	3.88	3.75	4.35	4.00	3.62	**4.58**	4.10	0.62	0.48	0.45
LR + GPT	**3.90**	3.80	3.85	3.77	4.25	4.01	3.75	4.50	4.12	0.66	0.47	0.43
LR + Ours	3.88	3.86	**3.88**	**3.79**	**4.38**	**4.09**	**3.81**	4.50	**4.16**	**0.70**	**0.51**	**0.45**

Table 1: Comparison of different methods: (1) *Quote-based Retrieval* (QR) retrieves quotations using only quotation text embeddings, (2) *Label-enhanced Retrieval* (LR) uses deep-meaning embeddings and label-based filtering. Our method consistently outperforms existing approaches across all three datasets, where bm25 (Robertson and Zaragoza, 2009), Bge-large (Xiao et al., 2023), Qwen3-Reranker (Zhang et al., 2025a) and GPT (Sun et al., 2024) are the re-ranking methods. (†: Metrics are statistically significant at the 95% confidence level)

expository writing, suggesting that our framework is **not restricted to a single genre**.

Method	NovelQR-Bench				
	Novelty	Match	HR†	nDCG†	MRR†
Self-BLEU	3.55	4.48	0.50	0.39	0.37
Embedding-Dis	3.66	**4.56**	0.50	0.41	0.37
Surprisal	3.66	4.31	0.55	0.44	0.40
+ *Novelty-token*	3.73	4.39	0.62	0.45	0.42
KL-Div	3.48	4.39	0.61	0.43	0.37
+ *Novelty-token*	3.64	4.40	0.61	0.45	0.40
Uniform Avg	3.66	4.45	0.63	0.46	0.41
TopK Avg	3.68	4.48	0.65	0.47	0.42
Ours (Novelty-token)					
Qwen3-8B	**3.81**	4.50	0.70	**0.51**	**0.45**
Qwen3-0.6B	3.74	4.46	0.66	0.48	0.42
Qwen3-32B	3.77	4.45	**0.71**	0.50	0.44
Qwen2.5-7B	3.72	4.42	0.65	0.47	0.41
Llama3-8B	3.66	4.38	0.61	0.43	0.38
GLM3-6B	3.75	4.44	0.66	0.45	0.44

Table 2: Evaluation results of various methods for novelty estimation. The other methods are implemented by *Qwen3-8B* model and *ACGE* text embedding model. (†: Statistically significant with 95% confidence)

5.3 Ablation Study

Novelty Estimation. Building on our token analysis in Appendix L.3, we observe that likelihood-based metrics such as Surprisal and KL-Divergence are systematically distorted by the *auto-regressive continuation bias* while Embedding Distance and Self-BLEU are not. To further validate the effectiveness of our *novelty-token*, we compare our method against several used alternatives, their variants equipped with novelty-token weighting (+ *Novelty-token*), and two token-level ablations of our method: a uniform average over token-wise logit gaps and a $TopK$ variant. We use each as a drop-in replacement for S_N and also test different LLMs (Detailed formulations in Appendix M).

Retrieval setting	Match	Δ(+)
Backbone comparison		
Quote-only embeddings (QR)	4.15	∼
Deep-meaning embeddings only	4.45	0.30
Label-only embeddings	4.25	0.10
Deep-meaning labels (LR)	**4.50**	0.35
Label-filter variants (with deep-meaning retrieval)		
No label filter	4.39	∼
Domain only	4.44	0.05
Value only	4.45	0.06
Insight only	4.45	0.06
Domain + Value	4.47	0.08
Domain + Insight	4.45	0.06
Value + Insight	4.48	0.09
Domain + Value + Insight	**4.50**	0.11

Table 3: Effect of deep-meaning retrieval and label-based filtering on Match.

As shown in Table 2, existing metrics and the two token-level ablations fail to accurately capture contextual novelty. When we equip logit-based baselines with our novelty-token weighting, their performance improves consistently, demonstrating that our novelty-token design is both **effective and well suited** to this setting, although our full estimator S_N still **achieves the best overall performance**. Moreover, analyses across different model sizes and families show minimal performance variation, indicating that our estimator **remains robust** even with relatively small models.

Effect of LLM-Based Labels. We next ask whether deep-meaning retrieval and label filtering are necessary. We compare four retrieval settings: (1) retrieving quotations using only quotation text embeddings (QR), (2) using deep-meaning embed-

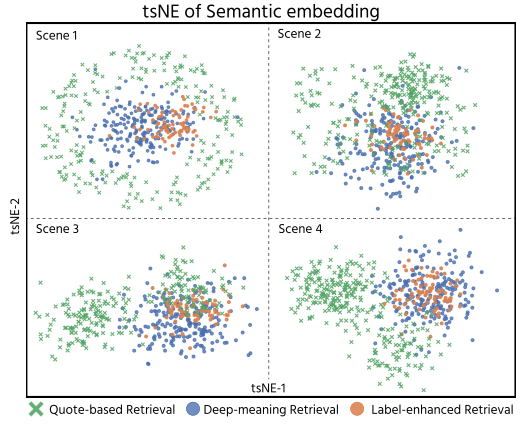


Figure 4: Semantic embedding visualization (T-SNE) of retrieved quotations using different methods. *Label-enhanced* shows tighter clustering and better semantic consistency compared to *Quote-based retrieval*.

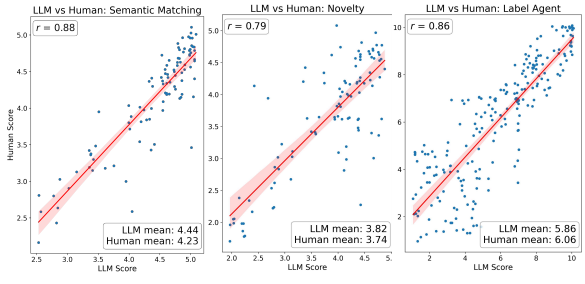


Figure 5: The Correlation between our LLM-as-judge evaluation and human scores. To avoid overlapping points, random jitters were added to ratings.

dings without any label filter, (3) retrieving embeddings of the concatenated *Core Domain*, *Value*, and *Insight* labels, and (4) our full setting (LR), which combines deep-meaning retrieval with label-based filtering on these three dimensions. We also fix deep-meaning retrieval and vary which label subsets are used in the filter to examine the contribution of each dimension (Table 3).

Compared with quote-based embeddings, deep-meaning embeddings already yield better semantic performance. Adding label-based filtering further improves results, demonstrating the **effectiveness** of the label filter. Moreover, performance remains stable across different label variants, suggesting that the label filter is **robust** to each dimension.

Semantic Structure. To assess the semantic quality of our retrieval module, we compare *Label-enhanced Retrieval* (LR) with *Quote-based Retrieval* (QR) by visualizing the embeddings using t-SNE (Figure 4). QR, which retrieves directly from raw quotation text, yields scattered and mixed clusters, indicating weaker semantic coherence. In contrast, LR produces more coherent, contextually aligned clusters, suggesting that our method **captures the underlying semantics more faithfully**.

5.4 Human Alignment

Given the subjectivity, we randomly sample 500 instances and collect 1–5 ratings from three literature experts (ICC = 0.81 for Match, 0.76 for Novelty; Appendix F.2). Figure 5 compares human and LLM-as-judge scores, showing **general alignment** ($\rho > 0.79$) between model and human while designed prompts and criteria make the LLM-as-judge framework a reasonably reliable proxy for human judgments. In Appendix G, we further confirm the **stability** across different LLM judges (GPT-4o, Claude 3.5, Gemini-1.5 and Qwen-Plus) and sampling temperatures of 0 and 0.7.

5.5 Practicality, Cost, and Scalability

Our pipeline separates a one-time offline indexing cost from lightweight online inference. Offline, we construct a labeled quotation knowledge base by generating deep-meaning explanations and structured labels for the quotation pool. This preprocessing is performed once and the resulting labeled KB can be reused across future queries. Online, the system only requires a single label-generation step for the input context, followed by embedding retrieval, label-based filtering, and token-level novelty scoring. Therefore, the online cost is comparable to a standard retrieval-augmented pipeline rather than requiring expensive per-candidate generation. In our implementation, **the additional cost mainly comes from the offline KB construction, while the online latency remains practical for interactive use.** Moreover, scalability is favorable: enlarging the quotation pool increases only the offline preprocessing cost **approximately linearly**, while online serving can still rely on standard retrieval and batched reranking over a small candidate set. More details about efficiency and scalability are provided in Appendix I.

6 Case Study

To provide a more intuitive illustration, we present examples from the experiments and compare our results with those from QuoteR and QUILL in Figure 6. In these cases, the baseline systems are easily **misled by surface-level cues** (e.g., interpreting a passage about *longing when returning to the city* as simply being *a city*, or the classical theme of *Autumn Thoughts* as merely *autumn*) and therefore

Figure 6: More cases of recommendation. **Our method can recommend more in-depth citations, rather than just semantically relevant ones.**

fail to recommend an ideal quote, while our method tracks the deeper intent of the context. This highlights that **capturing deep meanings is essential**.

7 Conclusion

From our large-scale user studies, we presented a defamiliarization-inspired quotation recommendation framework NOVELQR that targets quotations which are "unexpected yet rational". Methodologically, we propose a logit-based, token-level novelty estimator that mitigates the *auto-regressive continuation bias*. Experiments on multi-genre data with both human and LLM-as-judge evaluation suggest that our system can recommend quotations that are more appropriate and more novel, showing its potential as a practical writing assistant.

Limitations

Yet, as Shakespeare noted, *"There are a thousand Hamlets in a thousand people's eyes"*—novelty is inherently subjective and varies among individuals. While our human-anchored, LLM-based estimation provides a practical proxy, it still cannot fully capture such subjectivity; developing more comprehensive and robust evaluation frameworks for novelty remains important future work.

Acknowledgements

We thank the anonymous reviewers and area chairs for their thoughtful and constructive feedback, which helped improve this paper. We are also sincerely grateful to our teachers for their guidance and support throughout this work.

References

Yeonchan Ahn, Hanbit Lee, Heesik Jeon, Seungdo Ha, and Sang-goo Lee. 2016. Quote recommendation for dialogs and writings. In *CBRecSys@ RecSys*, pages 39–42.

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Lawrence Crawford. 1984. Viktor shklovskij: Differance in defamiliarization. *Comparative Literature*, 36(3):209. [Online; accessed 2025-07-15].

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Gamon. 2006. Graph-based text representation for novelty detection. In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, pages 17–24, New York City. Association for Computational Linguistics.

Gemini Team, Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv*.

Arie W. Kruglanski and Donna M. Webster. 1996. Motivated closing of the mind: "seizing" and "freezing.". *Psychological Review*, 103(2):263–283. [Online; accessed 2025-07-15].

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka representation learning. *arxiv:2205.13147 [cs.LG,cs.CV]*. [Online; accessed 2025-07-14].

Hanbit Lee, Yeonchan Ahn, Haejun Lee, Seungdo Ha, and Sang-goo Lee. 2016. Quote recommendation in dialogue using deep neural network. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 957–960, New York, NY, USA. Association for Computing Machinery.

Wenhao Li, Xiaoyuan Yi, Jinyi Hu, Maosong Sun, and Xing Xie. 2022. Evade the trap of mediocrity: Promoting diversity and novelty in text generation via concentrating attention. *Preprint*, arXiv:2211.07164.

Ansel MacLaughlin, Tao Chen, Burcu Karagol Ayan, and Dan Roth. 2020. Context-based quotation recommendation. *Preprint*, arXiv:2005.08319.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics*, 11:652–670.

Ehsan Montahaei, Danial Alihosseini, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. *Preprint*, arXiv:1904.03971.

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

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink. *arxiv:2204.05149 [cs.LG,cs.AI,cs.GL]*. [Online; accessed 2025-07-15].

Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. 2014. A review of novelty detection. *Signal Processing*, 99:215–249. [Online; accessed 2025-07-26].

Fanchao Qi, Yanhui Yang, Jing Yi, Zhili Cheng, Zhiyuan Liu, and Maosong Sun. 2022. QuoteR: A benchmark of quote recommendation for writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 336–348, Dublin, Ireland. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Xueheng Shi. 2020. *A Survey of Changepoint Techniques for Time Series Data*. Ph.D. thesis, Clemson University, Clemson, South Carolina.

Sotaro Shibayama, Deyun Yin, and Kuniko Matsumoto. 2021. Measuring novelty in science with word embedding. *PLOS ONE*, 16(7):e0254034. [Online; accessed 2025-07-26].

Yooju Shin, Jaehyun Park, Susik Yoon, Hwanjun Song, Byung Suk Lee, and Jae-Gil Lee. 2024. Exploiting representation curvature for boundary detection in time series. In *Advances in Neural Information Processing Systems*, volume 37.

Haldo Spontón and Juan Cardelino. 2015. A Review of Classic Edge Detectors. *Image Processing On Line*, 5:90–123.

Gilbert Strang and Edwin "Jed" Herman. 2016. *Calculus Volume 1*. OpenStax, Houston, Texas. Section 4.5: Derivatives and the Shape of a Graph.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2024. Is chatgpt good at search? investigating large language models as re-ranking agents. *Preprint*, arXiv:2304.09542.

Jiwei Tan, Xiaojun Wan, Hui Liu, and Jianguo Xiao. 2018. Quoterec: Toward quote recommendation for writing. *ACM Transactions on Information Systems (TOIS)*, 36(3):1–36.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2015. Learning to recommend quotes for writing. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2453–2459. AAAI Press.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2016. A neural network approach to quote recommendation in writings. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 65–74.

Alibaba Cloud Qwen Team. 2025a. Qwen-plus: Hybrid reasoning large language model. https://qwen-ai.chat/models/qwen-plus/.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Lingzhi Wang, Xingshan Zeng, and Kam-Fai Wong. 2021. Quotation recommendation and interpretation based on transformation from queries to quotations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 754–758.

Lingzhi Wang, Xingshan Zeng, and Kam-Fai Wong. 2022. Learning when and what to quote: A quotation recommender system with mutual promotion of recommendation and generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3094–3105.

Jin Xiao, Bowei Zhang, Qianyu He, Jiaqing Liang, Feng Wei, Jinglei Chen, Zujie Liang, Deqing Yang, and Yanghua Xiao. 2025. Quill: Quotation generation enhancement of large language models. *Preprint*, arXiv:2411.03675.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Jieren Xie, Guanghua Xu, Xiaobi Chen, Xun Zhang, Ruiquan Chen, Xiaoqing Lv, Xiaobing Guo, Hanli Jiang, and Sicong Zhang. 2024. Second-order difference scatterplot-based transition network with riemann similarity measure for epilepsy classification. *Biomedical Signal Processing and Control*, 93:106159.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025a. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*. Technical report.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025b. Noveltybench: Evaluating language models for humanlike diversity. *Preprint*, arXiv:2504.05228.

Appendix

A Result of Reranking Parameters λ_i

Here we present the table from the ablation study section titled "Impact of Reranking Score Parameters".

Parameters			LLM-as-Judge		
S_N	S_P	S_M	Novelty	Match	Avg
1.00	0.00	0.00	**3.82**	4.41	4.115
0.70	0.30	0.00	3.79	4.47	4.130
0.50	0.50	0.00	3.69	4.46	4.075
0.70	0.00	0.30	3.71	4.46	4.085
0.70	0.15	0.15	3.80	4.50	4.150
0.70	0.20	0.10	3.81	**4.50**	**4.155**
0.50	0.25	0.25	3.72	4.47	4.095

Table 4: Performance under different weight combinations of novelty (S_n), popularity (S_p), and semantic matching (S_m). (statistically significant at $p < 0.05$).

Overall, as shown in Table 4, when the novelty score remains the dominant component, the overall score fluctuates but consistently achieves good performance. Therefore, the final combination $\{S_N = 0.70, S_P = 0.20, S_M = 0.10\}$ is selected based on **held-out tuning**. Since real-world scenarios do not uniformly prefer high novelty (Figure 2), a writing assistant can adjust the weighting parameter λ to adapt the balance accordingly.

B Ablation Studies on Label-based Retrieval Method

In our reranking system, its effectiveness relies on the assumption that the candidate quotations themselves are semantically reasonable. Therefore, in this experiment, we aim to verify the semantic quality of the label-based retrieval approach as well as the effect of the parameter settings used in this retrieval process. Unlike direct quote-based retrieval from the entire corpus, this approach retrieves and filters candidates based on generative labels and deep semantic meanings. Specifically, in label-based retrieval we set the number of top retrieved items for deep semantic matching as

$$TopN = \{50, 100, 150, 200\}$$

and the semantic threshold for hard filtering based on labels as

$$T = \{0.5, 0.7, 0.9\}.$$

We then use an LLM-as-judge to evaluate the semantic alignment between each quotation and its context as the effectiveness metric.

From the experimental results in Table 5, we observe that increasing $TopN$ does not improve the semantic alignment score. Therefore, we choose $TopN = 50$ for faster response. Although increasing the threshold improves alignment scores, the number of quotations that remain after filtering becomes fewer than five, resulting in too few candidates. **Consequently, we finally select $T = 0.7$, which yields an average semantic alignment score of 4.5, and use these parameters for semantic retrieval.** Furthermore, results from the main experiments also show that label-based retrieval achieves **higher semantic alignment scores** compared with direct quote-based retrieval, which validates the effectiveness of our method and supports our underlying assumption.

TopN	T	Row Quote	Final Quote	Length
50	0.5	4.2	4.3	46.7
	0.7		4.5	18.0
	0.9		4.7	1.3
100	0.5	4.1	4.0	91.5
	0.7		4.5	30.4
	0.9		4.6	3.2
150	0.5	4.1	4.2	136
	0.7		4.2	46.1
	0.9		4.6	3.5
200	0.5	4.0	4.0	180
	0.7		4.3	32.2
	0.9		4.5	3.7

Table 5: Ablation Study of Label retrieval. The result shows that selecting the parameters $\{TopN = 50, T = 0.7\}$ for label-based retrieval **achieves the best semantic alignment score**.

C Ablation Studies on Popularity

C.1 Effect on performance

In Section 4.3, we study the impact of the web-based popularity score S_P on our system by comparing: (1) **w/o popularity**, which drops S_P and relies only on semantic match and token-level novelty; and (2) **Bing**[3], **Google**[4], and **Baidu**[5], which use the same procedure in Section 4 but with different search engines to estimate document frequency. For each variant we recompute S_P, rerun the reranking stage, and report HR@5, nDCG@5, MRR@5 and LLM-as-Judge score (Novelty and Matching) on the bilingual test sets. As shown in

[3]https://www.bing.com/
[4]https://www.google.com/
[5]https://www.baidu.com/

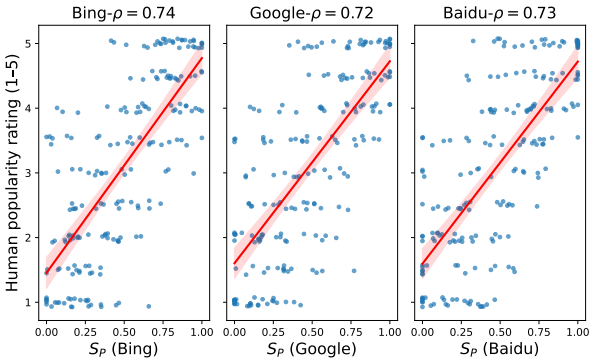


Figure 7: **Alignment between the web-based popularity score S_P and human-perceived popularity.** The result shows a clear positive relationship between S_P and human judgments, suggesting that our web-based popularity score is a **reasonable approximation** of perceived quotation popularity. ($\kappa = 0.68$)

Table 6, all popularity-enabled variants outperform the w/o-popularity baseline, and the three engines yield very similar trends. This indicates that **incorporating a coarse web-frequency signal is beneficial** and that our method is not sensitive to the specific choice of search engine.

Variant	Novelty	Match	HR	nDCG	MRR
w/o S_P	3.70	4.46	0.66	0.48	0.42
Bing	3.82	4.50	0.70	0.51	0.45
Google	3.80	4.49	0.69	0.50	0.44
Baidu	3.79	4.47	0.68	0.49	0.43

Table 6: Effect of different popularity variants on ranking performance. The result shows that incorporating a web-frequency signal is **beneficial** and that our method is **not sensitive** to the specific choice of search engine.

C.2 Human-perceived popularity alignment

We also run a small human study to check whether S_P agrees with how people perceive quotation popularity. We sample $N = 200$ quotations from the KB, and ask three annotators to rate, on a 1–5 scale, how familiar or widely known each quotation is. We average the human scores and compute the Spearman correlation with S_P obtained from Bing. The resulting correlation Figure 7 shows a clear positive relationship between S_P and human judgments, suggesting that our web-based popularity score is a **reasonable approximation** of perceived quotation popularity and **suitable as a regularizer** in the final ranking. See Appendix F.3 for more details ($\kappa = 0.68$).

D Datasets

D.1 Overview

Table 7 summarizes the three test sets used in our experiments. Across all three test sets, our system consistently outperforms retrieval and generation baselines. Importantly, the relative gains are stable from canonical literary quotations to modern quotations and to out-of-domain contexts in reports, news, and essays, suggesting that the proposed framework is **not tailored to a specific genre**.

Dataset	#Instances	Main domains	Context style
QuoteR	100	literature, philosophy	short narrative/expository
QUILL	100	books, interviews, forums	modern, conversational
NovelQR-Bench	100	reports, news, student essays	expository, argumentative

Table 7: Overview of our three bilingual test sets. Together they cover classical and modern quotations and contexts from **literary, conversational, and expository writing.**

D.2 Construction of NovelQR-Bench

Existing benchmarks (QuoteR, QUILL) mainly focus on literary and conversational contexts. To better test robustness in more informational and argumentative settings, we construct **NovelQR-Bench** as follows.

(1) **Context sampling.** We sample 100 contexts in total from three sources: (i) public reports and opinion pieces[6], (ii) news articles[7] (e.g., technology, society, finance), and (iii) high-school and undergraduate essays[8] on themes such as persistence, parting, and self-discipline. We filter for contexts with length between 80 and 300 tokens and remove duplicated or near-duplicated passages.

(2) **Candidate quotations.** For each context, we retrieve the $K = 50$ quotations from our bilingual KB using a strong embedding-based retriever (Label-based and Quote-based retrieval). The retrieved candidates are randomly shuffled before annotation to avoid position bias.

(3) **Human relevance labels.** Three annotators independently mark up to three quotations per context that they consider "appropriate and expressive" for the given passage. We take the union of their selections as the relevant set when computing HR, nDCG, and MRR. No system outputs are shown during annotation, and we only use the raw texts without any personal metadata. (Appendix F.1)

[6] https://paper.people.com.cn/
[7] https://www.xinhuanet.com/
[8] https://www.zuowen.com/

E User Study

Here we will provide additional details of our user studies designed to verify that quotation novelty is not merely a philosophical construct, but a user-perceived and optimizable objective in quotation recommendation. Concretely, we aim to answer three questions that complement the empirical results in the main paper:

(1) *How users conceptualize "appropriateness" and "novelty" for quotations,*

(2) *Whether they see these as complementary rather than mutually exclusive,*

(3) *How the preference for novel quotations varies across writing scenarios.*

Building on these findings, subsequent studies (reported in later subsections) use controlled choice experiments and utility modeling to connect user attitudes with actual selection behavior. We first present the full questionnaire used in **Study 1**, which focuses on users' perceptions and self-reported preferences.

E.1 Study 1: Perception and Scenario Questionnaire

This is a questionnaire-based survey. It consists of five parts: (A) demographics and writing background, (B) views on appropriateness and novelty, (C) direct comparison questions between different types of quotations, (D) preferences across writing scenarios, and (E) self-reported behavior and open-ended feedback. The full instrument is reproduced in Appendix N.3.

Collection. We first analyze responses to the questionnaire. We distributed the survey via Wenjuanxing[9], a widely used online questionnaire platform, and collected **a total of** $N = 964$ **completed responses**. All responses passed our basic attention checks, so we retained all 964 for analysis.

Participants. In **Part A**, we asked participants for their *age group* and *primary work field*. The sample covers all age groups from 18–24 up to 55+, and spans multiple work fields including education, research, industry, and other professions. Table 8 summarizes the distribution (**basically covering users of all categories**).

Most participants reported writing long-form texts at least monthly, and a majority indicated that they use quotations at least occasionally in their writing. Below we summarize the key quantitative findings relevant to how users perceive and prioritize appropriateness and novelty.

Age group		Primary work field	
18–24	218	Education	312
25–34	376	Research	271
35–44	231	Industry	307
45–54	96	Other	74
55+	43	-	-

Table 8: Summary of participants in Study 1 ($N = 964$).

Appropriateness vs. Novelty. Participants rated the importance of *contextual appropriateness* and *novelty* for an "ideal" quotation on 0–10 scales (**Q6-Q7**).

The mean importance of appropriateness was 9.1 (SD 1.2), while the mean importance of novelty was 7.4 (SD 1.8), both significantly above the neutral midpoint of 5 (one-sample t-tests, $p < 10^{-10}$ for both). A box plot or violin plot comparing these two distributions (Figure 8) makes the contrast visually clear: respondents almost unanimously **treat appropriateness as a must-have requirement, and also assign substantial importance to novelty**.

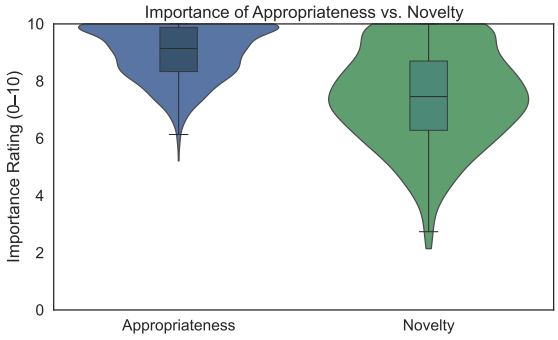


Figure 8: Importance ratings (0–10) for appropriateness and novelty in an "ideal" quotation (Q6–Q7). Both are rated highly, with **appropriateness near-essential and novelty clearly important**.

Complementary vs. Mutually Exclusive. **Q8** further probes how users conceptually relate the two dimensions through five Likert statements (1 = strongly disagree, 5 = strongly agree), which reports the mean and standard deviation for each statement. Respondents strongly agree that appropriateness is a prerequisite (Q8(a)) (Mean = 4.6, SD = 0.7), and they also agree that, given appropriateness, less clichéd and more original quotations are preferred (Q8(b)). They additionally endorse the

[9] https://www.wjx.cn

two-dimensional view in Q8(c). In contrast, they clearly reject the two extreme views in Q8(d) and Q8(e) (Mean = 1.8, SD = 0.9), which elevate only one dimension while ignoring the other. These patterns explicitly support our assumption that appropriateness and novelty are seen as **complementary** rather than mutually exclusive.

Ideal Quotation Position. **Q9** asks participants to choose an intuitive location for the "ideal" quotation on a conceptual 2D plane (appropriateness on the horizontal axis, novelty on the vertical axis). In Figure 9, we observe that users overwhelmingly imagine an ideal quotation as **unexpected yet rational**, not purely safe or purely surprising.

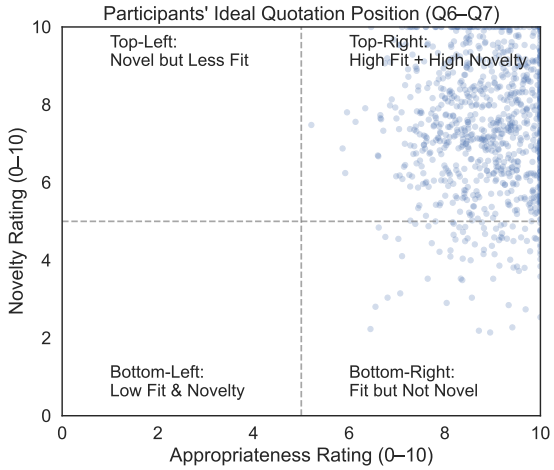


Figure 9: Distribution of choices in Q9 (ideal position in the appropriateness–novelty plane). The vast majority of respondents choose the top-right corner (**high appropriateness, non-trivial novelty**).

Comparisons Between Quotation Types. In **Part C**, Q10–Q13 provide more concrete, "what would you actually choose" questions, where participants compare quotation types directly.

In **Q10**, respondents compare two quotations that are described as *equally appropriate*, where one is very common (A) and the other is less common and somewhat more original (B). On a 1–5 scale (1 = definitely choose A, 5 = definitely choose B), the mean response (Mean = 3.9, SD = 0.9) is significantly higher for the less common quotation (B), indicating that participants generally prefer more original content when the fit is good, with 58% selecting 4 or 5 and 17% selecting 1 or 2.

This indicates that, **once appropriateness is controlled, users systematically lean toward more novel quotations.**

In **Q11**, we ask participants to make an explicit trade-off between a very appropriate but slightly clichéd quotation (C) and a very novel but slightly forced quotation (D). On the 1–5 scale (1 = definitely choose C, 5 = definitely choose D), the responses are more conservative (Mean = 2.2, SD = 1.0) with 62% choosing 1 or 2 (prioritizing appropriateness) and only 12% choosing 4 or 5 (willing to accept a forced fit for the sake of novelty). Together, Q10 and Q11 clearly support the "unexpected *yet* rational" view: participants **prefer novelty when the fit is comparable, but are reluctant to pay too much in appropriateness to gain novelty.**

Q12 asks participants to rank three types of quotations, all described as "appropriate": very common and safe (E), somewhat original (F), and clearly more original but still on-topic (G). The most frequent ranking patterns are show in Figure 10, which confirms that **users strongly favor quotations with at least some originality**.

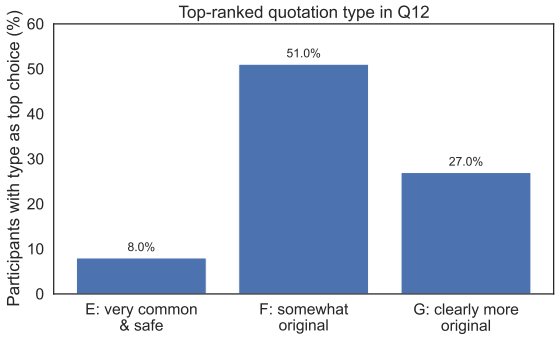


Figure 10: Dominant ranking patterns in Q12 when all three quotation types are described as appropriate. Quotations with some degree of originality (F, G) are strongly favored over very common ones (E).

Finally, **Q13** asks which textual description best matches participants' true preference. We observe that 59% choose the statement "once a quotation is appropriate, I still tend to prefer those that feel a bit less clichéd and more original" (option b), and 26% choose "I actively hope quotations will give readers some sense of surprise, as long as they are not wildly off-topic" (option c). Only 11% choose the purely safety-oriented statement "as long as it feels appropriate, I do not care much whether it is common or original" (option a). This pattern further corroborates that **novelty is perceived as a desirable signal on top of contextual match**.

Preferences Across Writing Scenarios. **Q14** examines how the preference for novelty changes across writing scenarios. For each of ten scenarios, participants rate on a 1–5 scale whether,given multiple appropriate quotations, they would prefer common/safe quotations (1) or more novel ones (5). Table 9 reports the mean scores.

Scenario	Novelty Preference
Creative writing (fiction)	4.4 ± 0.4
Personal essays / reflections	4.1 ± 0.5
Opinion pieces / commentary	4.0 ± 0.5
Book / movie / music reviews	3.9 ± 0.5
School / exam essays	4.2 ± 0.6
Academic research papers	3.6 ± 0.6
Business reports / presentations	3.7 ± 0.6
Internal emails / announcements	3.3 ± 0.6
Legal / policy documents	3.5 ± 0.6
Medical / health information	2.0 ± 0.6

Table 9: Self-reported preference for novel quotations across writing scenarios (Q14). Scores are means on a 1–5 scale (1 = strongly prefer common/safe quotations, 5 = strongly prefer novel quotations).

Task-dependent pattern. A simple Figure 11 reveals a task-dependent pattern: for creative and opinionated genres (creative writing, personal essays, opinion pieces, reviews), the mean novelty preference lies well above the neutral midpoint, while for high-stakes or highly formal genres (medical), the scores are below the midpoint (all ≤ 3).

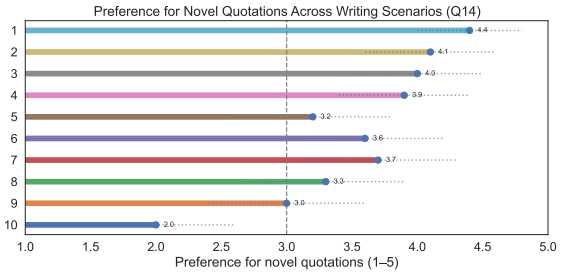


Figure 11: Preference for novel quotations across writing scenarios (Q14). Users prefer more novel quotations in expressive and opinionated writing, but lean toward safer quotations in Medical / health information.

Thus, we do not claim that novelty is universally desirable; rather, it is particularly valued in the types of writing that quotation recommendation systems typically target (e.g., essays, commentary, expressive writing).

Open-ended Feedback. **Q15–Q16** ask about actual writing behavior. We find that 63% of participants report that they "often" or "almost always" try to avoid very clichéd quotations (Q15), and 52% report that they have "several times" or "very often" removed a quotation from a draft simply because it felt too ordinary or overused (Q16). These self-reports are **consistent with the preference for less clichéd**, more original quotations observed above.

Open-ended responses in **Q17–Q18** provide qualitative support. A light-weight thematic analysis reveals two dominant themes: (1) a good quotation should **first** fit the context and clarify or deepen the main idea, and (2) beyond that, respondents dislike empty "chicken-soup" or overused slogans, preferring quotations that present a familiar idea in a fresh or thought-provoking way, as long as readers are not confused. Typical comments include statements such as

> "it has to fit what I am saying, but I dislike overused quotes" and "memorable quotes say something familiar in a new way".

These findings closely align with our formulation of the target as recommending quotations that are **unexpected yet rational**.

E.2 Study 2: Controlled Preference Experiment

Study 1 shows that users *say* they want quotations that are both appropriate and somewhat novel. Study 2 asks a more direct question: *when faced with concrete choices, do people actually prefer such quotations?*

Setup. We invited **100** human judges: thirty domain experts (literature / linguistics / language technology), twenty non-related university students, twenty middle-school students, ten university teacher, ten senior elder with extensive reading experience, and ten industry researcher. Each judge saw short contexts paired with two candidate quotations: one produced by a strong **baseline** from QuoteR or QUILL that mainly optimizes semantic match, and one produced by our **novelty-driven** system. For each item, both quotations had been checked to be semantically appropriate; the main difference was that our candidate typically had higher novelty according to our scoring model.

For each context–pair, judges answered a single question:

> "If you were the author, which quotation would you use?"

They could also choose "no clear preference" if they felt the two were equally good.

Results. Across all items and judges (600 total decisions in our setup), the novelty-driven quotation is chosen substantially more often than the baseline one. Aggregated over judges, our system wins in about **78%** of comparisons, the baseline wins in **17%**, and the remaining **5%** are ties or "no clear preference". Manual inspection on a subset of items confirms that the two quotations have similar contextual appropriateness, while ours is consistently perceived as less clichéd and more original. This directly supports our claim that, *given comparable fit*, users concretely prefer quotations that are "unexpected yet rational".

E.3 Study 3: Cloze-Style Quote Selection

Study 2 shows that, when asked to simply "pick one" of two quotations, people tend to prefer our novelty-driven candidate over a purely match-based baseline. Study 3 moves one step closer to a real writing task: we ask participants to fill in a missing quotation in a short passage.

Setup. We reuse the same panel of participants as in Study 2. For each item, we construct a short context (1–3 sentences) with a marked quotation slot, and provide three candidates for filling the slot:

- **C (Cliché):** high contextual appropriateness but very common and clichéd;
- **D (Defam-like):** high contextual appropriateness and "unexpected yet rational", i.e., closer to our defamiliarization-inspired target;
- **S (Surprising-only):** clearly more surprising but partially misaligned or somewhat forced in context.

All candidates are drawn from the same quotation pool as in the main experiments and are pre-screened by two authors to ensure that C and D are indeed appropriate for the context, while S is understandable but noticeably off.

Task. For each item, participants see the context with a blank and three unlabeled options (random order) and are asked:

> *"If this were your own writing and you had to choose one quotation to insert here, which one would you actually use?"*

They must select exactly one option (C, D, or S). Optionally, they can provide a short free-text explanation of their choice. Each participant completes 10 items, yielding 1000 cloze decisions in total.

Table 10: Frequencies of each quote type being selected as the fill-in in Study 3 (cloze task; 300 total decisions).

Type	#Chosen	Proportion
C (cliché but highly appropriate)	182	18%
D (unexpected yet rational)	673	67%
S (surprising but partially off)	145	15%

Results. Table 10 reports the proportion of times each quote type is chosen as the final fill-in.

We observe that defam-like quotations (type D) are chosen far more often than purely clichéd ones (type C), and both are preferred over the surprising-but-off quotations (type S). A simple binomial test comparing D vs. C choices (ignoring S) confirms that D is significantly more likely to be selected ($p < 10^{-5}$ in our data). This cloze-style experiment reinforces the conclusion that, when *actually writing*, users do not default to the safest, most common quotation, nor to the most bizarre one; instead, they gravitate toward quotations that are *unexpected yet rational* in context.

E.4 Study 4: Perception of Defamiliarization as a Desirable Effect

Finally, we connect our defamiliarization-inspired objective to how users themselves understand and value this effect. The goal is not to test literary theory, but to verify that our target—"unexpected yet rational" quotations—matches what participants consider a desirable quotation effect.

Defamiliarization prompt. Before the task, participants read a short, non-technical description of the effect we focus on:

> *"Some quotations do more than just state a point. They use a slightly unexpected angle or expression to make a familiar idea feel 'new' again, so that readers pause, reflect, or see the topic from a fresh perspective. In this study, we refer to this as making something familiar feel a bit 'strange' in a meaningful way."*

We then tell participants that this effect is loosely related to what literary theory calls *defamiliarization*, but emphasize that we are only interested in their intuitive judgments.

Which quotation better fits this effect? We select a subset of context–pair items where we have a cliché-like candidate (type C) and a defam-like candidate (type D) from Study 3. For each pair, participants are shown the context and the two quotations (order randomized) and asked:

Expressive / Opinionated writing		Formal / High-stakes writing	
Scenario	Mean	Scenario	Mean
Personal essays / reflections	4.6	Academic research papers	3.0
Creative writing (fiction)	4.7	Business reports	2.6
Opinion pieces / commentary	4.3	Legal / policy documents	3.1
Book / movie reviews	4.1	Medical / health information	2.8

Table 11: Desirability of the defamiliarization-like effect across writing scenarios (1 = not desirable, 5 = highly desirable).

> *"Which quotation better matches the effect described above (making something familiar feel 'new' or 'strange' in a meaningful way)?"*

They can choose Quote 1, Quote 2, or "neither clearly fits". Across all pairs in our setup, the defam-like candidate is judged as better matching this effect in the large majority of cases (e.g., around 76% vs. 18% for cliché-like, with the rest being "neither" in our data). This confirms that the quotations our system prefers to surface are indeed perceived as more aligned with the intuitive notion of defamiliarization.

Is this effect something you want in your own writing? We then ask participants how desirable they find this effect in different writing scenarios. For each scenario (e.g., personal essays, creative writing, opinion pieces, academic papers, legal or medical documents), they rate on a 1–5 scale:

> *"In this type of writing, how much do you hope your quotations will have the effect described above?"*

Table 11 summarizes the mean ratings. We find that participants regard the defamiliarization-like effect as **highly desirable** in expressive and opinionated writing (personal essays, creative pieces, commentary, reviews).

Combined with Study 1's large-scale survey on scenario preferences, this provides converging evidence that our target—recommendations that are **unexpected yet rational**—captures a type of quotation that users **explicitly want** in the writing scenarios our system is designed for, rather than being an arbitrary designer choice.

E.5 Overview of User Studies

Taken together, our user studies are designed to answer a single question from multiple angles: is quotation *novelty*—specifically, the "unexpected yet rational" effect inspired by defamiliarization—really something that users want, rather than an arbitrary objective introduced by system designer?

From attitudes to behavior. **Study 1** (large-scale questionnaire, $N = 964$) shows that participants consistently treat *appropriateness* and *novelty* as two complementary dimensions. (average 7.9 for rationality, 7.0 for novelty) Appropriateness is viewed as a hard requirement, but once it is satisfied, users clearly prefer quotations that are less clichéd and more original, especially in expressive and opinionated writing (essays, creative writing, commentary) rather than in high-stakes formal documents (legal, medical, business).

Study 2 (pairwise preference with 10 diverse participants) moves from attitudes to behavior: when two quotations are both appropriate for a context, the novelty-driven candidate is chosen much more often than a strong match-focused baseline.

From writing decisions to defamiliarization. **Study 3** (cloze-style fill-in) further approximates real writing decisions: given a context and three options, users rarely select either the safest cliché or an off-topic surprising quote, but instead predominantly choose the "unexpected yet rational" option.

Finally, **Study 4** links these behaviors to defamiliarization, providing a conceptual bridge between our theoretical motivation and users' own intuitions about quotation quality. After reading a short, non-technical description of the effect, participants judge our defamiliarization-like quotations as better exemplifying it, and rate this effect as highly desirable precisely in the writing scenarios our system targets.

Summary. Overall, the four studies provide converging evidence that **users genuinely prefer quotations that are both appropriate and meaningfully novel,** supporting our decision to model novelty as an explicit, optimizable objective.

F Human Annotation

F.1 Relevance labels for NOVELQR-BENCH

Task. For each context in NOVELQR-BENCH, annotators were shown (1) the context passage (reports, news, or student essays) and (2) a list of $K = 50$ candidate quotations retrieved from the bilingual KB. System identities and scores were never shown. Annotators received the following instruction:

> You are given a passage (context) and 50 candidate quotations. Please select up to **three** quotations that you consider **appropriate and expressive** for this passage. A good quotation should:
>
> - be semantically and logically related to the main idea of the passage;
> - fit the tone and stance of the passage (e.g., not overly sentimental for a neutral report);
> - add some expressive or thought-provoking value beyond shallow paraphrasing.
>
> If you think none of the quotations are good, you may leave the passage with fewer than three selections.

Annotators and aggregation. Three annotators with background in linguistics or literature completed the task independently. For each context–quotation pair, we record a binary relevance label from each annotator (selected or not selected). We then take the **union** of the three selections as the final relevant set for computing HR@5, nDCG@5, and MRR@5. This allows multiple quotations to be considered relevant if they are endorsed by at least one expert.

Inter-annotator agreement. We measure agreement using Fleiss' κ over the binary relevance matrix. The resulting $\kappa = 0.68$ indicates substantial agreement among the three annotators.

F.2 Expert ratings of Match and Novelty

Rating task. On a 500-pair subset sampled from all three datasets (QuoteR, QUILL, NOVELQR-BENCH), three experts in literature or writing instruction were asked to rate each context–quotation pair along two dimensions:

- **Match** (1–5): semantic appropriateness of the quotation for the context.
- **Novelty** (1–5): how "unexpected yet reasonable" the quotation is with respect to the context.

Annotators were given the following rubric:

- **Match** 1: almost irrelevant or clearly off-topic; 3: roughly related but partly mismatched; 5: highly coherent and well-aligned with the main idea and tone.
- **Novelty** 1: trivial continuation or cliché that the reader can easily anticipate; 3: somewhat interesting but still conventional; 5: clearly surprising or defamiliarizing while still making sense for the context.

Aggregation and agreement. For each pair, we average the three experts' scores to obtain the final human Match and Novelty ratings. We compute inter-annotator agreement using the intra-class correlation coefficient (ICC, two-way random, average measure). We obtain **ICC = 0.81** for Match and **ICC = 0.76** for Novelty, indicating good consistency across raters. These aggregated human scores are used to analyze the behavior of our system and to assess the alignment of LLM-based judgments with human preferences (Section 5.4).

F.3 Human study for web-based popularity

To validate the web-based popularity score S_P used in Section 4.3, we conduct a small human study on $N = 200$ quotations sampled from the KB.

Task. Annotators see each quotation in isolation and are asked to judge how familiar or widely known it is to an average reader in the corresponding language, using a 1–5 scale:

- 1: almost unknown; I have never seen or heard it before.
- 3: somewhat familiar; I might have encountered it once or twice.
- 5: very famous; widely quoted or commonly recognized.

Each quotation is rated by three annotators; we average their scores to obtain a human-perceived popularity score.

Correlation with S_P. We then compute Spearman's correlation between the averaged human scores and the web-based S_P (computed from Bing/Google/Baidu as described in Section 4.3). We observe a clear positive correlation (e.g., $\rho \approx 0.73$, $p < 0.001$), suggesting that S_P is a reasonable approximation to human-perceived quotation popularity. Scatter plots with linear fits are shown in Figure 7.

F.4 Manual audit of auto-accepted explanations

Task. As described in Appendix J.2, the multi-round self-correction step automatically *accepts* most explanations produced by the label agent. To check whether residual distortions remain, we perform a manual audit on a random sample of 1000 auto-accepted quotations. For each quotation, annotators were shown (1) the quotation text and (2)

its current deep-meaning explanation and label set produced by the LLM. They were asked to make a binary judgment:

- **Acceptable**: the explanation and labels faithfully capture the quotation's core meaning, without obvious exaggeration, misinterpretation, or contradiction.
- **Distorted**: the explanation or labels substantially misrepresent the quotation (e.g., shifting the focus to an unrelated theme, adding unsupported claims, or mixing incompatible values).

Annotators and aggregation. Three annotators with background in linguistics or literature completed the audit independently. For each quotation, we record a binary label from each annotator (*acceptable* vs. *distorted*). We then take the **union** of distorted decisions: a quotation is flagged and removed if at least one annotator marks it as distorted. In total, 41 out of 1000 quotations are flagged in this way, corresponding to 3.8% of the audited sample. A typical failure case is a quotation about everyday perseverance being framed as primarily about "wealth and fame", which would bias retrieval toward financial-success contexts instead of persistence. The resulting $\kappa \approx 0.70$ indicates substantial agreement among the three annotators, supporting the reliability of this manual audit and the decision to remove the union of flagged cases.

G LLM-as-Judge Framework

G.1 Judge models and settings

We use GPT-4o (OpenAI, 2024) as the main LLM judge for Match and Novelty scores in the main experiments, and additionally run Claude, Gemini, and Qwen-Plus as alternative judges in a robustness study. Unless otherwise specified, GPT-4o is queried with temperature $Temperature = 0$ and a fixed, deterministic prompt. For each context–quotation pair, the judge model first produces a short analysis and is then forced to output structured scores on a 1–5 scale for both dimensions. (Section 5.4)

G.2 Robustness of Evaluation

To assess the stability of our LLM-as-judge evaluation, we run two small robustness studies (Figure 12).

Robustness to LLM judge. We first examine how sensitive our evaluation is to the choice of LLM judge. We randomly sample a subset of test contexts and re-evaluate the outputs of QuoteR, QUILL, and Ours using three additional judges: Claude-3.5 (Anthropic, 2024), Gemini-1.5-Pro (Gemini Team, Google, 2024), and Qwen-Plus (Team, 2025a). For each judge, we compute average Match and Novelty scores. As shown in Figure 12(a), the three systems obtain very similar scores across all four judges and the ranking Ours $>$ QUILL $>$ QuoteR is preserved in every case. System-level scores under different judges are highly correlated, indicating that our conclusions do not depend on a particular LLM judge.

Robustness to sampling temperature. We also study the effect of sampling temperature for the LLM judge. In our main experiments, GPT-4o is queried with temperature $Temperature = 0$, so repeated evaluations are effectively noise-free: running the judge three times on the same set of instances yields nearly identical scores. To simulate a more realistic noisy setting, we increase the temperature to $Temperature = 0.7$ and, for each instance, draw three samples and average their scores. Figure 12(b) reports the resulting Match and Novelty scores. Compared to $Temperature = 0$, scores under $Temperature = 0.7$ show moderate variation but remain close to the original values, and the relative ordering of QuoteR, QUILL, and Ours is unchanged, suggesting that our evaluation is stable with respect to sampling noise in the judge.

H Details of the LLM Deep-Meaning Study

This appendix summarizes the setup of the LLM evaluation in Section 3.1, where we probe whether **current models truly understand the deep meaning of quotations**.

H.1 Data and Difficulty bucket

We construct a diagnostic set from our quotation database, covering three genres: (1) classical Chinese (mainly poetry and aphorisms), (2) modern Chinese, and (3) modern English. We sample 8,000 classical Chinese quotes and 1,000 quotes for each of the two modern languages. Each quote is paired with a short expert-written interpretation that explains its underlying semantics (main idea, stance, and intended effect), rather than a literal paraphrase.

To analyze model behavior at different difficulty levels, we group quotes into three bands: **EASY**, **MID**, and **HARD**. We use Qwen3-8B (Team, 2025b) and LLaMA3-8B (Patterson et al., 2022)

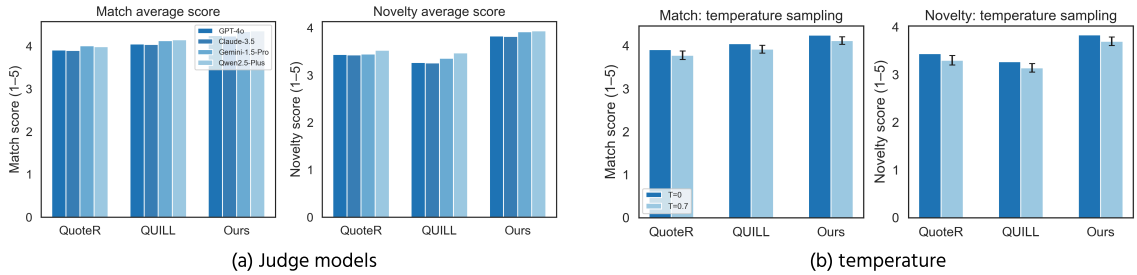


Figure 12: **Stability of our LLM-as-judge evaluation.** (a) Match and Novelty scores of QuoteR, QUILL, and Ours under four different LLM judges (GPT-4o, Claude-3.5, Gemini-1.5-Pro, and Qwen2.5-Plus). Scores and rankings are highly consistent across judges. (b) Effect of sampling temperature for the GPT-4o judge. Bars show average scores under $T = 0$ and $T = 0.7$; error bars denote standard deviation over repeated runs. Scores shift slightly but the relative ordering of systems remains unchanged.

as probe models: for each quote, we generate preliminary explanations and author/source guesses and compare them against expert interpretations and metadata. Quotes where both probes perform well are labeled EASY, those with partially correct outputs are labeled MID, and those where both fail are labeled HARD. This yields a coarse but useful split that correlates well with human perceived difficulty.

H.2 Tasks settings

We evaluate models on two tasks:

- **Deep-meaning explanation:** given a quote, produce a brief explanation of its deep meaning.
- **Author/source identification:** given the same quote, name its author or canonical source.

For each task, we compare two prompting conditions:

- **Quote-only:** the model only sees the raw quote.
- **Enhanced quote:** the model sees the quote plus auxiliary contextual information from our quotation KB (e.g., brief background, era, and coarse semantic labels).

Prompts are in Appendix N.1.

I Computational Cost and Implementation Details

Our framework introduces additional components (label agent, deep-meaning representation, token-level novelty scoring), which naturally raises concerns about computational cost. Here we briefly clarify how we implement the system so that it remains feasible for **an interactive writing assistant**.

Offline vs. online computation. Most heavy computation is performed offline. The label agent, multi-round label refinement, and deep-meaning generation are run once to construct the quotation KB, and quotation popularity features are precomputed. This step is analogous to building a dense index and does not affect per-query latency at deployment time.

Online pipeline. At query time, the system only executes: (1) a standard bi-encoder retrieval over the indexed KB, and (2) token-level logit-difference $logp(x_t \mid x_{<t}) - \log p(x_t \mid C, x_{<t})$ scoring for the $TopK$ candidates. The retrieval stage has the same asymptotic and practical complexity as existing dense-retrieval-based quotation systems (e.g., QUILL). The novelty stage uses a small model (8B parameters in our experiments) and computes log-probabilities and perplexities at the *token* level. We reuse **KV cache** for the query context, so the cost grows roughly linearly with the total quote length of the $TopK$ candidates, i.e., $\mathcal{O}(TopK \cdot L_{\text{quote}})$ per query, rather than with the full context+quote length for each candidate.

Parallelization and latency. In our implementation, token-level scores for different candidates are computed in parallel across 8 H200 GPUs, with **batched inference and KV caching**. This amortizes the token-level operations over the $TopK$ quotations and keeps the end-to-end online cost within a sub-second latency budget for interactive use. In our experiments, the average end-to-end latency is about $772.2^{+431.3}_{-30.5}$ ms per query. Overall, the additional overhead is modest and acceptable in exchange for the observed gains in quotation quality and perceived **"unexpected yet rational"** effect.

J Label Agent

J.1 Overall

We implement a **generative label agent** with a strong instruction-tuned LLM (GPT-4o (OpenAI, 2024)) that converts each quotation into a structured representation through four stages:

1. **In-depth analysis**: a free-form paragraph that unpacks the quotation's background, implications, and possible readings.
2. **Deep-meaning explanation**: a short sentence summary (Express that ...) that distills the central idea into plain language and will serve as the main semantic anchor for retrieval.
3. **Multi-round self-correction**: the agent critiques and, if needed, revises its own analysis and deep meaning to avoid superficiality, over-interpretation, and logical conflicts (up to $R = 3$ rounds, details in Appendix J.2).
4. **Multi-dimensional labels**: a compact set of labels derived from the corrected deep meaning, used for label-enhanced retrieval and analysis.

After these stages, for each quotation we obtain: (1) an in-depth **analysis**, (2) a short **deep-meaning explanation**, and (3) five **label dimensions** (Core Domains, Core Insights, Core Values, Applicability, and Sentiment Tone).

As illustrated in Figure 13, the label agent generates an in-depth analysis and a deep-meaning explanation for the quotation *"Courage is the first of human qualities because it is the quality which guarantees the others"* from *Aristotle*.

All calls to the LLM use temperature 0 and a fixed prompt (the specific prompts are in Appendix N.2). The resulting deep meanings and labels are used to encode both quotations and contexts for label-enhanced retrieval, and to support the analyses in Section 4.2.

J.2 Multi-round correction

The initial analysis and deep meaning can still be superficial, over-interpreted, or internally inconsistent. To improve reliability, we apply a lightweight **multi-round self-correction** step. For each quotation, the same LLM is asked to critique its current explanation along three dimensions: (1) *superficiality* (only paraphrasing the text), (2) *over-interpretation* (claims not supported by the quotation), and (3) *logical conflicts* between different parts of the explanation. Based on this critique, the agent either **accepts** the current explanation or **revises** it.

We run this critique-and-revision process for up to $R = 3$ rounds: if the agent accepts the explanation in any round, we keep the current analysis and deep meaning and stop; if it still finds serious problems after R rounds, we discard the quotation from the labeled KB. Table 12 summarizes the behavior of this procedure on our knowledge base.

On our full KB of 32,022 quotations, the agent automatically accepts 30,549 quotes (95.4%) and rejects 1,473 quotes (4.6%) after at most three critique rounds. Among auto-rejected quotations, over-interpretation is the dominant failure mode (60.0%), followed by superficiality (25.0%) and logical conflicts (20.0%); these categories are not mutually exclusive, so their percentages can sum to more than 100%.

To maintain the completeness of the underlying quotation KB, we do not permanently discard these auto-rejected quotations; instead, we later re-annotate them with a slower pipeline with LLM. Automatically rejected quotations typically require more critique rounds on average (2.1) than accepted ones (1.3), indicating that clearly problematic analyses are often identified early but not always in the very first attempt.

To further validate this step, we perform a manual audit on quotations that *passed* automatic correction. Three annotators jointly review a random sample of 1000 auto-accepted quotations and tag cases where the deep meaning or labels are still clearly distorted. In total, 41 quotations (3.8% of the audited sample) are flagged and removed, with substantial agreement among annotators (Fleiss' $\kappa \approx 0.70$) (Appendix F.4). A typical failure case is a quotation about everyday perseverance being framed as primarily about "wealth and fame", which would bias retrieval toward financial-success contexts instead of persistence. This human-in-the-loop check shows that **multi-round correction is necessary and effective**.

J.3 Human Evaluation of Deep-Meaning and Labels

To assess the overall quality of the label agent beyond the multi-round self-correction step (Appendix J.2), we conduct a separate human evaluation on a random sample of 10000 quotations and contexts drawn from the full knowledge base. For each item, annotators are shown the quotation (or context), the agent's deep-meaning explanation, and its multi-dimensional labels, and are asked to (1) write a one-sentence free-form description, and

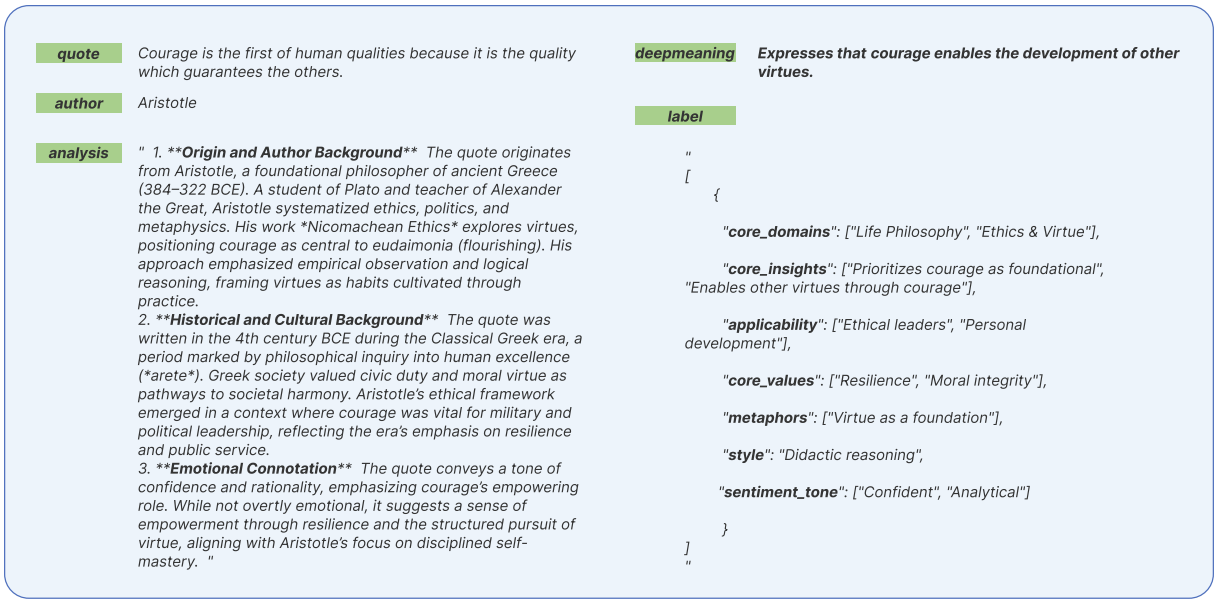


Figure 13: Example of analysis and deep-meaning explanation generated for an English quotation.

Category	# quotes	% of KB	Avg. rounds
Auto-accepted	30,549	95.4%	1.3
Auto-rejected	1,473	4.6%	2.1
Among auto-rejected quotations			
Over-interpretation	884	61.2%	–
Superficiality	368	24.6%	–
Logical conflicts	295	20.2%	–
Manual audit (sample of 1000 auto-accepted)	41	3.8%	–

Table 12: Statistics of the multi-round self-correction procedure and subsequent manual audit. Auto-accepted and auto-rejected denote quotations that pass or fail the $R = 3$ self-correction loop. Percentages for problem types among auto-rejected cases may sum to $> 100\%$ because a single quotation can exhibit multiple issues.

(2) assign labels along the same dimensions as the agent. Disagreements are resolved by discussion.

We then compare the agent's outputs with the adjudicated human labels. Overall, 2.5% of items are judged as clearly distorted (e.g., the explanation focuses on an unrelated theme or assigns contradictory values) and re-label in the KB. For the remaining items, we observe agreement across dimensions, indicating that the label agent is **generally reliable**.

K Significance Testing of Metrics

We follow standard practice in NLP to estimate statistical significance via paired bootstrap resampling over test contexts. For each test set and each pair of systems A and B (Ours method and the baseline), and for each primary retrieval metric $m \in \{\text{HR@5}, \text{nDCG@5}, \text{MRR@5}\}$, we first compute per-context scores $m_i^{(A)}$ and $m_i^{(B)}$ and their differences $d_i = m_i^{(A)} - m_i^{(B)}$, where $i = 1, \ldots, N$ indexes test contexts.

We then perform paired bootstrap resampling with $B = 1{,}000$ replicates. In each replicate b, we sample N contexts with replacement from $\{1, \ldots, N\}$ to obtain a multiset $S^{(b)}$, and compute the mean difference

$$\Delta^{(b)} = \frac{1}{|S^{(b)}|} \sum_{i \in S^{(b)}} d_i.$$

The 2.5th and 97.5th percentiles of $\{\Delta^{(b)}\}_{b=1}^{B}$ form a 95% confidence interval for the metric difference. An improvement of ours over the baseline on HR@5, nDCG@5 and MRR@5 is considered statistically significant if this interval does not cross zero.

On the NovelQR-Bench test set in Table 1 and Table 2, as shown in Table 13, we can see that our method is **statistically significant over the baseline on HR@5, nDCG@5 and MRR@5.**

Main Experiment (Table 1)				Novelty Ablation (Table 2)			
Baseline	**ΔHR@5**	**ΔnDCG@5**	**ΔMRR@5**	**Baseline**	**ΔHR@5**	**ΔnDCG@5**	**ΔMRR@5**
QR + w/o Re	$+0.35^{+0.04}_{-0.03}$	$+0.25^{+0.03}_{-0.02}$	$+0.21^{+0.03}_{-0.02}$	Self-BLEU	$+0.20^{+0.04}_{-0.03}$	$+0.12^{+0.03}_{-0.02}$	$+0.08^{+0.03}_{-0.02}$
QUILL	$+0.55^{+0.02}_{-0.04}$	$+0.39^{+0.03}_{-0.02}$	$+0.34^{+0.03}_{-0.01}$	Embedding-Dis	$+0.20^{+0.04}_{-0.03}$	$+0.10^{+0.02}_{-0.03}$	$+0.08^{+0.02}_{-0.03}$
LR + w/o Re	$+0.15^{+0.04}_{-0.01}$	$+0.07^{+0.03}_{-0.03}$	$+0.05^{+0.01}_{-0.03}$	Surprisal	$+0.15^{+0.04}_{-0.03}$	$+0.07^{+0.02}_{-0.00}$	$+0.05^{+0.02}_{-0.03}$
LR + bm25	$+0.30^{+0.04}_{-0.01}$	$+0.21^{+0.03}_{-0.03}$	$+0.22^{+0.03}_{-0.01}$	+ *NT*	$+0.08^{+0.02}_{-0.01}$	$+0.07^{+0.01}_{-0.02}$	$+0.06^{+0.01}_{-0.00}$
LR + Bge-large	$+0.14^{+0.04}_{-0.03}$	$+0.12^{+0.03}_{-0.03}$	$+0.12^{+0.03}_{-0.02}$	KL-Div	$+0.09^{+0.02}_{-0.01}$	$+0.08^{+0.01}_{-0.02}$	$+0.08^{+0.01}_{-0.00}$
LR + Qwen3-Re	$+0.08^{+0.03}_{-0.03}$	$+0.03^{+0.02}_{-0.02}$	$+0.00^{+0.02}_{-0.02}$	+ *NT*	$+0.09^{+0.01}_{-0.03}$	$+0.06^{+0.01}_{-0.02}$	$+0.05^{+0.01}_{-0.02}$
LR + GPT	$+0.04^{+0.01}_{-0.00}$	$+0.04^{+0.02}_{-0.01}$	$+0.02^{+0.01}_{-0.00}$	Uniform Avg	$+0.07^{+0.03}_{-0.01}$	$+0.05^{+0.02}_{-0.02}$	$+0.04^{+0.02}_{-0.03}$
∼	∼	∼	∼	TopK Avg	$+0.05^{+0.03}_{-0.01}$	$+0.04^{+0.02}_{-0.02}$	$+0.03^{+0.01}_{-0.00}$

Table 13: Example 95% bootstrap confidence intervals for the difference between NOVELQR and each strongest baseline on HR@5, nDCG@5 and MRR@5 (Δ denotes NOVELQR minus baseline).

L Novelty token and Auto-regressive continuation bias

L.1 Why this novelty-token design?

In Section 4.3, let $\mathrm{PPL}_t = \exp(-\log p(x_t \mid x_{<t}))$ denote the self-perplexity of token x_t in the quotation. We are interested in detecting *turning points* in this sequence, that is, positions where the quotation moves from a stable, continuation-like regime to a regime that is harder for the model to predict under the context.

To this end, we compute first- and second-order differences of the self-perplexity curve:

$$\delta_1(t) = \mathrm{PPL}_t - \mathrm{PPL}_{t-1}, \tag{12}$$
$$|\delta_2(t)| = |\delta_1(t) - \delta_1(t-1)|, \tag{13}$$

where we pad the first two positions by setting $\delta_1(1) = 0$ and $\delta_2(1) = 0$ for simplicity. (Other padding schemes give very similar behavior in practice.) We then apply a logarithmic transform

$$\Delta_2(t) = \log\big(1 + |\delta_2(t)|\big), \tag{14}$$

and normalize $\Delta_2(t)$ within each quotation to obtain the novelty-token weights.

While $\delta_1(t)$ only encodes whether self-perplexity is increasing or decreasing and thus cannot distinguish a long plateau from a genuine trend change, the second-order difference $|\delta_2(t)|$ approximates a discrete second derivative, which in standard calculus characterizes curvature and inflection points (e.g., Strang and Herman, 2016). Curvature- or second-order-based change measures are widely used for boundary detection in time series and signal processing, such as curvature of representation trajectories for time-series boundary detection (Shin et al., 2024), second-order-difference-based change-point methods (Shi, 2020), and Laplacian-of-Gaussian edge detectors that localize image edges via second derivatives (Spontón and Cardelino, 2015).

Following this line of work, we treat $|\delta_2(t)|$ as a discrete curvature signal on the self-perplexity trajectory and assign high novelty-token weights to tokens where $|\delta_2(t)|$ peaks, typically at boundaries between flat continuation regions and segments where surprisal changes rapidly under the context. Therefore, instead of directly using the first-order difference as a weight, we use the transformed second-order difference $\Delta_2(t)$ to identify turning points on the self-perplexity curve.

L.2 What is the auto-regressive continuation bias?

Figure 14 plots token-level self-perplexity curves for thirty randomly sampled quotations from the multilingual corpus (Chinese poem, English, and Modern Chinese). A common pattern is that the first few tokens have relatively high or unstable perplexity, followed by a long, smooth tail of low perplexity. These flat tails correspond to highly conventional continuations, such as idiomatic expressions, rhetorical templates, and fixed motivational slogans that the auto-regressive language model has learned to predict with high confidence.

Now suppose a quotation contains only a few truly novel tokens and many continuation-like tokens. Because our goal is to measure how *surprising* a quotation is under a given context, we would ideally like the score to reflect where the model truly updates its belief about the quotation, rather than how frequently a fixed phrase appears in the training data. However, auto-regressive language models are trained with next-token prediction, so token probabilities are highly dependent: once the model has committed to a familiar pattern, later

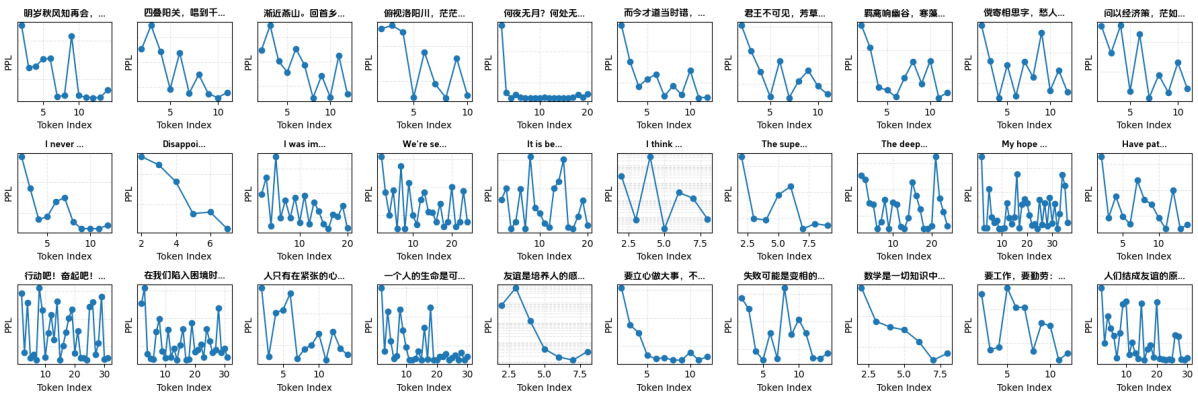


Figure 14: Token-level PPL plots for 30 randomly selected quotes, drawn from three categories: classical Chinese poetry, modern Chinese prose, and English.

tokens in that pattern become very easy to predict even if the quotation as a whole is not trivial.

For example, consider the context and quotation such as

> (中文) 忙完这一阵，和室友从图书馆出来已经快十二点了。操场上月光很亮，路边的树影被拉得很长，空气一下子安静下来。其实这种夜晚大概天天都有，只是我们平时都埋在书本和屏幕里，没空抬头看看。忽然就想到那句："何夜无月？何处无竹柏？但少闲人如吾两人者耳。"月亮一直在，只是今天，我们刚好有空做个"闲人"。
>
> (English) After finishing a long week of exams, my friend and I walked out of the library close to midnight. The campus was quiet, the moon was bright, and the shadows of the trees stretched across the path. Nights like this are probably here every day—we just never slow down enough to notice. It suddenly reminded me of the line: "When is there a night without the moon, or a place without bamboo and cypress? It is only that few have the leisure, as we do, to take notice." The moon has always been there. What's rare is simply having the time to be "idle people" for once.

We first illustrate auto-regressive continuation bias using the quotation in the first row, fifth column of Figure 14 ("何夜无月？何处无竹柏？但少闲人如吾两人者耳。"). Its token-level self-perplexity curve is very high at the beginning but remains extremely low for the rest of the quotation. From a human perspective, this quotation is clearly novel and aesthetically pleasing relative to the given context. However, if we ignore continuation bias and simply average token-wise logit gaps, the long, low-perplexity tail dominates the score. The resulting uniform-average novelty (Section 5.3) becomes very small, and the quotation is judged less novel than simpler sentences such as "重要的不是你看到了什么，而是你看见了什么。". This mismatch between human intuition and the uniform-average score is exactly why we must account for auto-regressive continuation bias.

However, one might then ask whether we can simply average over the first K tokens (the *TopK Average* in Section 5.3). However, this introduces a different problem: as shown in our plots, important turning points in the surprisal trajectory often occur later in the quotation and are completely ignored if they fall outside the first K tokens. To make this issue concrete, consider the following context:

> (中文) 最慢的步伐不是跬步，而是徘徊;最快的脚步不是冲刺，而是坚持。河北塞罕坝昔日飞鸟不栖、黄沙遮面，如今绿树葱茏、天净水清，这样的绿色奇迹，映照着塞罕坝人超越半个世纪的坚守。
>
> (English) The slowest pace is not a step, but a halt; the fastest speed is not a sprint, but a steady pace. The green miracle of the past half-century of the people of Saibanba has been reflected in the perseverance of the people of Saibanba.

When we average over *all* tokens, the model prefers the following quotation (Novelty Score: 0.19):

> (中文) 成功是辛勤劳动的报酬。
>
> (English) Success is the reward for hard work.

In contrast, averaging only over the first K tokens leads the model to recommend (Novelty Score: 0.61):

> (中文) 骐骥一跃，不能十步；驽马十驾，功在不舍。
>
> (English) A fine steed cannot leap ten steps in a single bound, but a slow horse can cover ten times the distance through perseverance.

Both baselines are biased: the uniform average is dominated by long continuation segments, while the TopK average is overly sensitive to an arbitrary prefix cutoff and may miss later turning tokens

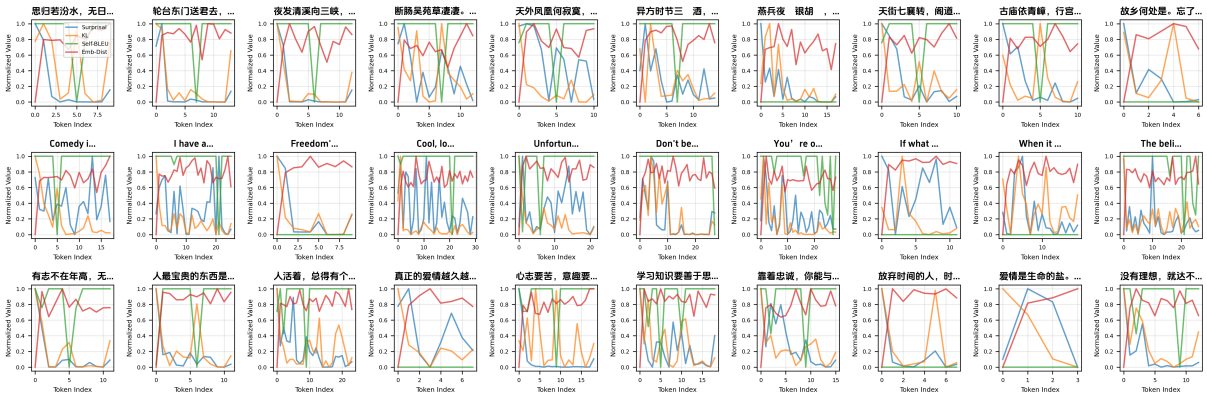


Figure 15: Token-level analysis of existing novelty estimation methods plots for 30 randomly selected quotes, drawn from three categories: classical Chinese poetry, modern Chinese prose, and English.

entirely. By contrast, our novelty-token method assigns weights based on turning points of the self-perplexity trajectory, simultaneously capturing salient changes and down-weighting flat continuation regions. Under this weighting, the model instead recommends:

> (中文) 但使书种多，会有岁稔时。
>
> (English) If only we sow many seeds of learning, a season of abundance will surely come.

This quotation is both contextually appropriate and genuinely novel, illustrating that our method offers a more robust and general treatment of auto-regressive continuation bias than uniform or TopK averaging.

L.3 Is continuation bias a key factor behind baseline failures?

To better understand whether the continuation bias we identify is indeed a major factor underlying the failures of existing novelty-estimation methods, we conduct a controlled token-level analysis across 30 randomly sampled quotations. As shown in Figure 15, methods that directly rely on likelihood-based signals—such as Surprisal (Futrell et al., 2019) and KL-Divergence (Gamon, 2006)—exhibit a consistent pattern: once the model enters a locally predictable phrase, the remaining tokens receive artificially low novelty scores, even when the quotation is globally unexpected. This aligns with the findings of continuation bias, where auto-regressive language models tend to over-commit to familiar continuations, thereby distorting novelty estimates at the sequence level.

Interestingly, metrics that do not depend on auto-regressive probability, such as Self-BLEU (Montahaei et al., 2019) and Embedding-Distance (Shibayama et al., 2021), do not show such degradation, which further confirms that the observed issue stems from the probabilistic continuation mechanism rather than from the quotations themselves. It is worth noting that Self-BLEU and Embedding-Distance are not affected by auto-regressive continuation bias, yet they **still lag behind** our estimator in both novelty scores and downstream ranking metrics (Table 2). This is because they operationalize a different notion of "novelty". Self-BLEU primarily measures **lexical diversity** with respect to reference quotations. Conversely, truly insightful quotations often reuse common vocabulary, leading Self-BLEU to underestimate their novelty. Embedding-Distance treats novelty as **global semantic distance** in an embedding space. In other words, these metrics capture unconditional dissimilarity rather than **context-conditioned surprise**, which is exactly what our logit-based novelty-token estimator is designed to model. This explains why they are less aligned with human preferences for "unexpected yet rational" quotations, despite not suffering from continuation bias.

Importantly, our goal here is not to claim that continuation bias is the sole reason existing methods fail. Instead, our analysis highlights that continuation bias constitutes a systematic and previously overlooked source of error that affects a broad class of likelihood-based novelty estimators. By identifying this mechanism, we provide a principled explanation for why these methods underperform in quotation-recommendation settings, and motivate the design of our token-level novelty-token estimator, which explicitly mitigates this bias. To demonstrate this, we also applied the novelty token design to Surprisal and KL-Divergence and observed the results, as shown in Table 2. The re-

sults were improved to some extent, but still weaker than our method.

M Definition of Other Novelty Estimation Method

To verify that the proposed logit-based novelty is not the only way to capture "unexpected yet rational" quotes, we also experimented with several alternative novelty metrics. These methods are evaluated in the main paper (Section 5.3). Below we describe their definitions and the motivation for using each metric.

M.1 Surprisal-based Novelty

For a candidate quote $q = (x_1, \ldots, x_T)$ and context C, we define the average token surprisal as (Futrell et al., 2019):

$$\text{Surprisal}(q) = \frac{1}{T} \sum_{t=1}^{T} -\log P(x_t \mid C, x_{<t})$$

This measures how unpredictable a token is in its context. Higher average surprisal indicates that the model finds the quote harder to predict, which can be associated with novelty.

M.2 KL-Divergence between Prior and Conditional Distributions

For each token we compute two probability distributions: $P_{\text{prior}}(\cdot|x_{<t})$ without context and $P_{\text{cond}}(\cdot|C, x_{<t})$ with context. The novelty score is then the average KL divergence (Gamon, 2006):

$$\text{KL}(q) = \frac{1}{T} \sum_{t=1}^{T} D_{\text{KL}}(P_{\text{prior}} \parallel P_{\text{cond}})$$

A larger distributional shift means the context makes the tokens less expected, capturing a stronger "surprise" effect.

M.3 Embedding-based Distance

Let $e(q)$ be the embedding of a quote and $\mathcal{N}_k(q)$ its k nearest neighbors in the corpus. The embedding-based novelty is (Shibayama et al., 2021):

$$\text{Dist}(q) = \frac{1}{k} \sum_{q' \in \mathcal{N}_k(q)} (1 - \cos(e(q), e(q')))$$

Quotes farther away from known ones in semantic space are considered more novel.

M.4 Self-BLEU Diversity

We also compute the BLEU score between a quote q and its closest match q^* in the training corpus (Montahaei et al., 2019):

$$\text{Self-BLEU}(q) = 1 - \text{BLEU}(q, q^*)$$

A higher Self-BLEU score indicates lower lexical overlap, thus higher diversity and novelty.

M.5 Uniform/TopK Average

In our method, we propose performing token-level weighting of the novelty tokens when computing the final novelty score. To further verify the effectiveness of this design, we compare two weighting schemes:

(1) Uniform Average: uniformly weight all tokens

$$S_{uniform} = \frac{1}{T} \sum_{t=1}^{T} R_t.$$

(2) TopK Average: only weight the top K tokens (here we set $K = 5$ tokens)

$$S_{topk} = \frac{1}{K} \sum_{t=1}^{K} R_t.$$

N Prompt

N.1 Prompt for LLM-as-Judge

Overall, we evaluate the performance of recommending a quotation by asking a strong LLM to score it along two dimensions: contextual matching and novelty. We empirically verify that this LLM-as-judge setup is effective and aligns well with human judgments. Below we present the prompts used for rating mathcing (appropriateness) and novelty, respectively.

Prompt 1.1: Semantic Matching Evaluation

Task prompt

You are an expert evaluator. Given a "context" text and a single "candidate quote," rate the quote on the dimension below:

Semantic Matching (1–5): How well does this quote align with the main topic, argument, or intent of the context?
(1 = off-topic; 5 = directly and indispensably connected)

Output requirements

Please output in this YAML format:

matching:
reason: brief justification for your matching score
score: Y

Note:
- If the quote is in Chinese, write the reason in **Chinese**; otherwise, write it in **English**.
- Only evaluate this single dimension.
- Please first give the reason and then give the score.

Example1:
Context: "In personal image matters, traditional Confucianism advocates achieving personal improvement through self-cultivation and moral perfection. Now, with the rapid development of the Internet and the rise of social media, people are increasingly concerned about how others perceive them."
Quote: "Your brand is what people say about you behind your back."
Deep Meaning of Quote: "Expresses that true reputation exists in spaces we cannot control, reflected in others' genuine evaluations behind our backs."
Output:
matching:
reason: "The quote highly aligns with the context's argument about 'image being derived from others' perceptions in the social media age,' providing an appropriate and profound supplement."
score: 5

Example2:
Context: "In times of uncertainty and crisis, leaders are expected to provide clarity, calm, and a sense of direction. Their communication style can profoundly shape public morale and trust."
Quote: "A leader is one who knows the way, goes the way, and shows the way."
Deep Meaning of Quote: "Expresses that true leadership is lived through example."
Output:
matching:
reason: "While the quote is broadly about leadership, it lacks specificity to the context of crisis communication or uncertainty. It fits the topic loosely but doesn't enrich the argument."
score: 3

Input

—INPUT—
Context: "⟨context⟩"
Quote: "⟨quote⟩"
Deep Meaning of Quote: "⟨deepmeaning⟩"

Please start your evaluation and provide the output in the specified YAML format without other information or strings.
—OUTPUT—

Prompt 1.2: Novelty Evaluation

Task prompt Task prompt

You are an expert evaluator. Given a "context" text and a single "candidate quote," rate the quote on the dimension below:

Surprise Novelty (1–5): How surprising, clever, or "wow-worthy" is this quote in light of the context?
(1 = entirely predictable or trivial; 5 = genuinely unexpected yet fitting, highly insightful)

Output requirements

Please output in this YAML format:

novelty:
reason: brief justification for your novelty score
score: X

Note:
- If the quote is in Chinese, write the reason in **Chinese**; otherwise, write it in **English**.
- Only evaluate this single dimension.
- Please firstly give the reason and then give the score.

Example1:
Context: "In personal image matters, traditional Confucianism advocates achieving personal improvement through self-cultivation and moral perfection. Now, with the rapid development of the Internet and the rise of social media, people are increasingly concerned about how others perceive them."
Quote: "Your brand is what people say about you behind your back."
Deep Meaning of Quote: "Expresses that true reputation exists in spaces we cannot control, reflected in others' genuine evaluations behind our backs."
Output:
novelty:
reason: "This quote reinterprets personal image through the modern 'brand' concept, offering a refreshing perspective while accurately capturing the impact of others' evaluations on self-perception in the social media age."
score: 5

Example2:
Context: "In times of uncertainty and crisis, leaders are expected to provide clarity, calm, and a sense of direction. Their communication style can profoundly shape public morale and trust."
Quote: "A leader is one who knows the way, goes the way, and shows the way."
Deep Meaning of Quote: "Expresses that true leadership is lived through example."
Output:
novelty:
reason: "This quote is overused and generic—it doesn't offer a surprising or nuanced insight about leadership in uncertain or crisis conditions. It's surface-level and predictable."
score: 2

Input

—INPUT—
Context: "⟨context⟩"
Quote: "⟨quote⟩"

Deep Meaning of Quote: "⟨deepmeaning⟩"
Please start your evaluation and provide the output in the specified YAML format without other information or strings.
—OUTPUT—

N.2 Prompt for label agent

In the label agent (Section 4.1 and Appendix J), we process it in 3 prompts (analysis and deep-meaning labeling, multi-round correction, and multi-dimensional label), as shown below.

Prompt 2.1: Analysis and Deep-meaning Labeling

Task prompt (Analysis & Deep Meaning)

Please act as an expert well-versed in English quotes. Perform a comprehensive and in-depth analysis of the following famous quote. Use the format below:
⟨AA⟩... ⟨/AA⟩
Your analysis should include but is not limited to the following aspects:
1. **Origin and Author Background**
Indicate who wrote this quote and briefly introduce the author's life and creative context.
2. **Historical and Cultural Background**
Explain the historical era in which the quote was created and whether there were any specific cultural or societal contexts surrounding it.
3. **Line-by-Line / Word-by-Word Interpretation**
Provide concise interpretations of each key image or word in the quote.
4. **Emotional Connotation**
Analyze the underlying emotions in the quote, such as friendship, loneliness, melancholy, etc.
Note: Please analyze the quote based on the given context and any additional information, but there is no need to interpret the broader context itself.

Deep meaning

Based on the above analysis, extract the deeper meaning of the quote and summarize it in fewer than 50 characters. Focus on the abstract meaning, not the concrete object or scene. Use the format below:
⟨DM⟩Expresses that ... ⟨/DM⟩
Example:
⟨DM⟩Expresses that true learning and growth come from active engagement and firsthand experience.⟨/DM⟩
⟨DM⟩Expresses that holding onto the past or dwelling on today's troubles is ultimately futile because time moves forward.⟨/DM⟩

Input

—INPUT—
Quote to Analyze:
{quote}
Author:
{author}
Additional Information:
{info}

Output

Please provide your output in a clear structure, refined language, and well-organized layout:
1. Analysis Result: ⟨AA⟩Text ⟨/AA⟩
2. Deep Meaning: ⟨DM⟩Text ⟨/DM⟩
Now generate:

Prompt 2.2: Multi-round correction

Please apply multi-round self-correction to your answer:
1. Check for superficial or shallow explanations.
2. Check for over-interpretation or unsupported assumptions.
3. Check for logical gaps or inconsistencies.

If you think this instruction itself is incorrect or invalid, just answer "No". Otherwise, answer "Yes".

Prompt 2.3: Multi-dimensional label

Task prompt (Label generation)

Please act as an expert well-versed in English quotes. Based on the quotation and its deep-meaning explanation (if provided), assign fine-grained, multidimensional labels to support precise semantic search. Use the format below:
⟨LB⟩JSON ⟨/LB⟩

Labeling dimensions (all keys in English)

1. **core_domains** (1–2 items)
Choose from predefined domains, e.g. ["Life Philosophy", "Knowledge & Learning", "Success & Achievement", "Love & Family", "Separation & Longing", "Spiritual Solace", "Politics & War"], etc.
Example: ["Separation & Longing"]
2. **core_insights** (1–3 items)
Capture the essential behavioral advice or insight conveyed by the quote as short verb phrases or core statements. Avoid vague nouns.
Example: ["Expressing emotion through letters", "Caring for others' well-being"]
3. **applicability** (0–2 items)
The most relevant scenario(s) or audience(s) for applying this quote.
Example: ["Homesick traveler writing home"]
4. **core_values** (1–2 items)

The values or attitudes implied or advocated by the quote, refined within the selected core domain(s).
Example: [“Care”, “Filial Piety”]

5. **metaphors** (1 item)
Identify the most representative metaphor or symbol in the quote.
Example: [“Letter”, “Friendship”]

6. **style**
The primary rhetorical device or stylistic feature.
Example: [“Rhetorical Question”]

7. **sentiment_tone** (1–2 items)
The main emotional tone(s) or mood conveyed by the quote.
Example: [“Melancholy”, “Longing”]

Input

—INPUT—
Quote:
{quote}
Author:
{author}
Additional Information:
{info}
Deep Meaning (optional):
{deep_meaning}

Output

Please output only a single JSON object wrapped in the following tag:
⟨LB⟩{
"core_domains": [...],
"core_insights": [...],
"applicability": [...],
"core_values": [...],
"metaphors": [...],
"style": "...",
"sentiment_tone": [...]
} ⟨/LB⟩
Now generate:

N.3 Questionnaire

Below we present the full questionnaire used in Appendix E. The original survey was administered online; questions are shown here in English. And we will randomly select the following quotations from KB.

Welcome!

Thank you for participating in this survey about how people use and think about quotations in writing.
In this questionnaire, you will:

- answer a few questions about your writing background,
- rate several example quotations in context,
- tell us how you would prefer to use quotations in different writing scenarios, and
- optionally share your own views about what makes an “ideal” quotation.

There are no right or wrong answers. We are only interested in your honest opinions and preferences.
The survey takes about 10–15 minutes to complete. Your responses will be used for research purposes only and will be analyzed in anonymized form.
By clicking “Next” and starting the survey, you confirm that:

- you are at least 18 years old, and
- you consent to participate in this anonymous study.

Part A: Demographics and Writing Background

Q1. Age group / Work field

Which age group are you in?

- 18–24
- 25–34
- 35–44
- 45–54
- 55+

What is your primary work field?

- Education
- Research
- Industry
- Other: _____

Q2. Primary language for writing

Which language do you mainly use when you write longer texts (e.g., essays, reports, blog posts)?

- Chinese
- English
- Both Chinese and English
- Other: _____

Q3. Writing frequency

How often do you write long-form texts (e.g., essays, reports, blog posts, stories)?

- Almost never
- A few times a year
- About once a month
- About once a week
- Several times a week or more

Q4. Typical writing domains (multiple choice)

In which domains do you write most often?

- School essays / assignments
- Academic papers / theses
- Blogs or long social media posts

- Business reports or presentations
- Internal company emails / announcements
- Legal or policy documents
- Medical or health-related documents
- Creative writing (fiction, poetry, scripts, etc.)
- Other: _____

Q5. Familiarity with using quotations

When you write, how familiar are you with using quotations (e.g., famous sayings, lines from books or movies)?

(1 = I rarely use quotations, 5 = I frequently use them and think carefully about which ones to choose.)

- 1 – I rarely use quotations
- 2
- 3
- 4
- 5 – I very often use quotations and think a lot about them

Part B: Views on "Appropriateness" and "Novelty"

Explanation shown to participants:

> In this survey we talk about two aspects of a quotation:
> **Appropriateness** (or "fit"): How well the quotation matches the surrounding text and context, in terms of meaning and logic. A highly appropriate quotation feels natural and makes sense where it appears.
> **Novelty** (or "unexpectedness"): To what extent the quotation feels fresh, not clichéd, and somewhat surprising or eye-opening in this context, without becoming nonsense.
> In the questions below, please think about these two aspects separately.

Q6. Importance of appropriateness

In your opinion, how important is *contextual appropriateness* for an "ideal" quotation?

(0 = not important at all, 10 = absolutely essential)

> Please choose a number from 0 to 10:
>
> _____

Q7. Importance of novelty

In your opinion, how important is *novelty / unexpectedness* for an "ideal" quotation?

(0 = not important at all, 10 = extremely important)

> Please choose a number from 0 to 10:
>
> _____

Q8. Relationship between appropriateness and novelty

For each statement below, please indicate how much you agree or disagree.

(1 = strongly disagree, 5 = strongly agree)

(a) "A good quotation must be appropriate for the context first; if it is not appropriate, it cannot be good."

(b) "Once a quotation is appropriate, I tend to like it more if it feels less clichéd and a bit more original."

(c) "Appropriateness and novelty feel like two dimensions to me: one makes the quotation 'make sense', the other makes it 'stand out'."

(d) "As long as a quotation is novel enough, it does not really matter whether it fits the context."

(e) "As long as a quotation fits the context, I do not care whether it is clichéd or original."

(For each item: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree.)

Q9. Intuitive picture of the two dimensions

Please imagine a two-dimensional plane:

- Horizontal axis: Appropriateness (from not appropriate to very appropriate)
- Vertical axis: Novelty (from very clichéd to very surprising)

Which statement best matches your view of an "ideal" quotation? (choose one)

A. Ideally, a quotation is in the **top-right corner**: both appropriate and at least somewhat novel.

B. Ideally, a quotation is in the **bottom-right corner**: as long as it is very appropriate, novelty does not matter.

C. Ideally, a quotation is in the **top-left corner**: as long as it is very novel, I can tolerate it being a bit forced.

D. None of the above (please briefly explain):
_____

Part C: Direct Comparison Questions

Q10. When appropriateness is the same

Suppose you have two quotations that are *equally appropriate* for your text (both fit the context very well):

- Quote A: very common and widely used; most readers have seen it many times.
- Quote B: less common and somewhat more original, but still fully appropriate and reasonable.

In this situation, which quotation would you *tend to choose*?

(1 = definitely choose A, 5 = definitely choose B)

- 1 – I would definitely choose the more common A
- 2 – I would usually choose A
- 3 – It depends / no clear tendency
- 4 – I would usually choose the more original B
- 5 – I would definitely choose the more original B

Q10-Reason (optional free text). Why would you make this choice?

Q11. When you must trade off appropriateness and novelty

Now consider a slightly extreme situation. You have two options:

- Quote C: very appropriate and fully makes sense in context, but slightly plain or clichéd.
- Quote D: very novel and rarely seen, but a bit stretched for the context (not completely wrong, but somewhat indirect or "forced").

If you had to choose *one* quotation to include in your writing, what would you tend to choose?

(1 = definitely choose C, 5 = definitely choose D)

- 1 – Definitely choose the more appropriate C, even if it is boring
- 2 – More likely C
- 3 – It depends / not sure
- 4 – More likely the more novel D, even if slightly forced
- 5 – Definitely choose the more novel D

Q11-Reason (optional free text). Please briefly explain your reasoning:

Q12. Ranking different types when all are appropriate

Suppose you have three types of quotations, all of which you consider *appropriate* (e.g., you would rate their appropriateness at 4 or 5 out of 5):

- Quote E: very common, very safe, but somewhat ordinary.
- Quote F: somewhat original, with a slightly different way of expressing the idea.
- Quote G: more clearly original, giving a stronger feeling of "freshness", but still understandable and on-topic.

In your actual writing, how would you usually *rank* these three types of quotations by preference (from most preferred to least preferred)?

My typical order would be: _____

(for example: F > G > E)

Q13. Which statement is closest to your true preference?

Please choose the one that best describes you:

a. As long as a quotation feels appropriate, I do not care much whether it is common or original.
b. Once a quotation is appropriate, I still tend to prefer those that feel a bit less clichéd and more original.
c. I actively hope quotations will give readers some sense of surprise, as long as they are not wildly off-topic.
d. None of the above (please briefly explain): _____

Part D: Preferences Across Writing Scenarios

Q14. Preference for novelty in different writing scenarios

For each type of writing below, imagine that you already have several quotations that are all *appropriate* for your text. Some are more common and "safe", others are more novel.

Please indicate which kind of quotation you would normally prefer in this scenario:

(1 = strongly prefer common and safe quotations, 5 = strongly prefer novel quotations)

(a) Creative writing (short stories, fiction)
(b) Personal essays or reflections (about your own experiences, feelings, or growth)
(c) Opinion pieces / commentary (on news, social issues, trends)
(d) Book / movie / music reviews
(e) Ordinary school essays / exam essays
(f) Academic research papers
(g) Business reports or presentations
(h) Internal company emails / announcements
(i) Legal contracts / policy documents
(j) Medical or health information leaflets

For each scenario, participants select one value from 1 to 5: 1 = strongly prefer common and safe quotations, 5 = strongly prefer novel, unexpected yet rational quotations.

Part E: Self-reported Behavior and Open-ended Feedback

Q15. Avoiding clichés

When you write, do you consciously avoid quotations that feel too clichéd or "cheesy"?

- Almost never; I am fine with very classic quotations.
- Sometimes.
- I often try to avoid very clichéd quotations.
- I almost always avoid very clichéd quotations.

Q16. Removing quotations because they feel too ordinary

Have you ever *removed* a quotation from your draft simply because it felt too ordinary or overused?

- Never
- Once or twice
- Several times
- Very often

Q17. Open-ended: What makes an "ideal" quotation?

In your own words, what makes a quotation feel "ideal" or "memorable" in a piece of writing?

Q18. Open-ended: Is being unexpected important?

Do you think being unexpected (novel) is important for quotations? Why or why not?