Controllable Lexical Alignment in Dialogue Systems

Sumit Srivastava, Mariët Theune,
Human Media Interaction
University of Twente
Enschede, The Netherlands
{s.srivastava-1,m.theune}@utwente.nl

Alejandro Catala
Centro Singular de Investigación en
Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago
de Compostela, Spain
alejandro.catala@usc.es

Abstract

Lexical alignment, the tendency of interlocutors to adopt each other's vocabulary during dialogue, is a phenomenon that contributes to mutual understanding, reduced cognitive load, and enhanced trust. Implementing controllable lexical alignment in dialogue systems, however, remains an open problem. Existing methods require user profiling or training data, and none provides continuous parametric control over the degree of alignment. We present a training-free, inference-time weighted decoding method that addresses these limitations. The method was evaluated with four architecturally diverse language models using 95 source dialogues from DailyDialog, resulting in 10,545 generated dialogues. These dialogues were assessed through a multi-judge LLM evaluation framework, complemented by a small scale human evaluation. The results show that weighted decoding provides effective, graded, and bidirectional control over lexical alignment across all evaluated model architectures, preserving generation quality within model-specific operating ranges that provide practical guidance for selecting appropriate weight factors.

1 Introduction

Lexical alignment represents a fundamental aspect of human conversational behaviour (Pickering and Garrod, 2004, 2006). We follow their mechanistic account, in which alignment arises from automatic, largely unconscious priming that propagates across linguistic levels, rather than the separate, consciously negotiated process of lexical convergence described by Brennan and Clark (1996). It contributes to conversational coherence, mutual understanding, and social bonding (Branigan et al., 2010). In human-human dialogue, it was found to be correlated with improved team performance in collaborative tasks (Reitter and Moore, 2014; Coco et al., 2018), enhanced learning outcomes in educational contexts (Ward and Litman, 2007), reduced cognitive load during interaction (Spillner and Wenig, 2021), and increased trust between interlocutors (Pozzi et al., 2024). In human-agent dialogues, matching user vocabulary has been shown to improve short-term information retention and explanation comprehension (Srivastava et al., 2023).

The current approaches to automated lexical alignment in dialogue systems cannot guarantee consistent alignment behaviours or adapt dynamically to changing conversational contexts without prior engineering, limiting their practical deployment (Srivastava et al., 2025). The emergence of Large Language Models (LLMs) as the foundation for modern conversational AI systems has made it possible to build adaptive, open-domain conversational agents at a rapid pace. Recent research shows that LLMs possess inherent capabilities for linguistic adaptation comparable to those observed in human conversations (Wang et al., 2023a; French et al., 2024). Their aligning behaviour can be further enforced through prompt engineering (Zhao et al., 2024; Srivastava et al., 2024). However, prompts result in different outcomes across models and lack the fine-grained control necessary for consistent alignment behaviours across extended conversations.

We address these limitations through a weighted decoding approach that modifies token probability distributions during generation (inference time). This enables controlled lexical alignment without model retraining or architectural modifications, making it applicable across LLM-based conversational agents. Our method exposes the degree of alignment as a single continuous parameter. This is an architectural difference from prompt-based methods, which offer a fixed, model-dependent operating point rather than a graded control knob, and it holds regardless of the relative output quality of either approach.

We evaluate the proposed methodology using automatic and human evaluations to answer two

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 1053–1073
October 17-21, 2026 ©2026 Association for Computational Linguistics

research questions: 1) does weighted decoding provide usable, graded, and bidirectional control over lexical alignment while preserving generation quality? If so, 2) is there a preference for dialogues with higher lexical alignment?

Our contributions include:

- A weighted decoding approach providing fine-grained, inference-time control of lexical alignment in LLM-based dialogue without model modifications or retraining.
- Automated evaluation across 10,545 generated dialogues over four LLMs, suggesting model-specific weight ranges for desired alignment.
- Human evaluation of utterance generation with lexical alignment against the standard generation of four different models.

2 Background and Related Work

2.1 Lexical Alignment Implementation

Early work by Janarthanam and Lemon (2009) used reinforcement learning for alignment, but requires extensive user profiling. Lopes et al. (2013, 2015) combined rule-based keyword matching with regression models, but depend on predefined vocabularies and domain-specific training data. Spillner and Wenig (2021) used grammatical transformation and GloVe-based (Pennington et al., 2014) word substitution, but GloVe's static representations lack the contextual understanding needed for nuanced dialogue (Arora et al., 2020). Kumar and Dušek (2024) fine-tuned a GPT-2 derivative to boost user-token probabilities, raising alignment but hurting naturalness and coherence and locking generation to the training-data alignment level. Schaaij et al. (2025) builds a user lexical profile from the first ∼10 minutes of interaction; it needs no training and suits low-resource settings, but short interactions do not give enough data for a good profile.

LLMs have transformed dialogue systems while introducing control challenges. Zhao et al. (2024) and Srivastava et al. (2024) showed LLM alignment via prompting GPT-3.5 and GPT-4, but responses vary even under identical prompts and contexts. While LLMs align comparably to humans (Wang et al., 2023a; French et al., 2024), reliably directing these capabilities toward specific alignment objectives remains an open problem.

2.2 Constrained Language Generation

Template-based generation, a classic method for constrained language generation (Gatt and Krahmer, 2018), remains prevalent in domains requiring high reliability and predictability (Santhanam and Shaikh, 2019). However, its scalability issues and domain-specific nature have motivated the exploration of neural approaches that can learn constraints from data rather than requiring explicit rule specification (Wiseman et al., 2018; Arun et al., 2020).

Constrained neural generation follows two paradigms (Liang et al., 2024). *Hard constraints* enforce strict adherence, primarily via constrained beam search and its variants (Pascual et al., 2020; Zhang et al., 2020a; Miao et al., 2019); these guarantee satisfaction but often yield rigid, unnatural output and are computationally intensive. *Soft constraints* instead bias generation probabilistically, at the training stage (Keskar et al., 2019; Chan et al., 2021), via fine-tuning (Arora et al., 2022; He, 2021), or through decoding-time intervention (Konen et al., 2024; Dathathri et al., 2020; Garbacea and Mei, 2025).

We use a soft constraint through decoding-time intervention (weighted decoding), as dialogue generation needs flexibility and speed while pursuing alignment. This risks omitting important facts, making it less suitable for task-oriented conversations where factualness is essential.

3 Method

3.1 Problem Formulation

Given a dialogue history $H = \{u_1, u_2, \ldots, u_{n-1}\}$ in which the two speakers strictly alternate, our goal is to generate the next response u_n that exhibits controlled lexical alignment with the conversation partner while maintaining semantic coherence and linguistic fluency. We refer to the partner being aligned to as Speaker 1 (S1), and the aligning system as Speaker 2 (S2). S2 produces u_n; under strict alternation the partner's (S1's) utterances are those at odd distance from u_n, i.e., $\{u_{n-1}, u_{n-3}, u_{n-5}, \ldots\}$, and these are the only utterances S2 aligns to.

We quantify lexical alignment at the turn level as $\alpha(u_n, u_{n-1})$: the fraction of u_n's tokens that reuse vocabulary shared with u_{n-1} (the exact measure is given in Section 3.3). Being a fraction of the response's own tokens, the score is comparable across response lengths; 0 indicates no reuse and 1

full mimicry of S1's utterance.

3.2 Weighted Decoding for Lexical Alignment

Our weighted decoding mechanism modifies the token probability distribution at each generation step, introducing controlled bias toward alignment vocabulary (i.e., the words used by S1 in the dialogue so far) while maintaining stochastic generation. The weighting follows the temporal decay observed in human conversation, where recent utterances influence vocabulary choices more.

3.2.1 Weight Assignment Mechanism

For each token v_i in the model vocabulary V, we compute an alignment weight

$$w_i = \max\big((\lambda \cdot \omega_i \cdot 0.1) + 1,\ 0\big)$$

where:

- λ is the *weight factor*: a single scalar controlling global alignment strength, varied along the horizontal axis ("Weight") in Figures 3 and 4.
- ω_i is the temporal-decay weight of token v_i, determined by how recently S1 used it (see below). For a token S1 has not used, $\omega_i = 0$, so $w_i = 1$ (baseline).
- 0.1 is a fixed scaling constant letting λ be swept over a convenient integer range; only the product $\lambda \cdot 0.1$ is meaningful.
- the baseline $+1$ preserves non-zero probability for every vocabulary token.
- $\max(\cdot, 0)$ floors the weight so a negative λ (used to *reduce* alignment) cannot drive it below zero; for $\lambda \geq 0$ it is inactive.

We extract word tokens (excluding punctuation and words shorter than three characters) from S1's utterances throughout the dialogue and assign each a decay weight ω by its *age*: the number of S1 turns since the token was last used. A token last used in the most recent S1 utterance has age 1 and $\omega = 1.0$; one from the previous S1 utterance has age 2 and $\omega = 0.5$; in general $\omega = 0.5^{\text{age}-1}$, a geometric decay (Reitter et al., 2006). The halving at each S1 turn keeps recent context more influential while preventing excessive historical accumulation. The decay schedule (geometric, rate 0.5) is one design choice; Appendix G reports a sensitivity analysis over alternative schedules. Each extracted word is mapped to its subword token IDs, all of which receive the word's weight; once a token is generated, its weight reverts to 1 for the remaining steps of that response, so a boosted token is not repeatedly favoured.

3.2.2 Generation Process

The K candidate responses are sampled in parallel through iterative token-by-token construction (their diversity arising from stochastic sampling); for each token position until termination:

a) *Weight Application:* Model logits undergo softmax transformation, multiplication by the alignment weights, and L_1 normalization:

$$P'(v_i) = \frac{w_i \cdot \text{softmax}(z_i)}{\sum_{j \in V} w_j \cdot \text{softmax}(z_j)}$$

where v_i denotes the i-th token in the vocabulary V, $\mathbf{z}$ is the logit vector from the model's final layer, and w_i is the alignment weight of token v_i (Section 3.2.1). For $w_i > 0$ this is equivalent to adding $\log w_i$ to the corresponding logit; we use the multiplicative form because the $+1$ baseline makes the unweighted case the exact identity and the $\max(\cdot, 0)$ floor safely accommodates negative λ.

b) *Top-k Filtering:* The top 20 tokens by weighted probability are retained, balancing diversity with quality.

c) *Multinomial Sampling:* Tokens are selected via multinomial distribution over filtered candidates to preserve stochastic generation.

d) *Termination Check:* Generation continues until an end-of-sequence or end-of-turn token, or the 30-token maximum length; each candidate is then trimmed to its last complete sentence.

The 30-token limit bounds generation time while accommodating typical conversational utterance lengths. We used top-k filtering rather than nucleus sampling (top-p), as this avoids the threshold selection required by top-p and provides predictable interaction with the alignment weights. The top-k value (k=20) and the number of candidate responses (K=10; Section 3.3) were fixed a priori; Appendix G reports the effect of varying each on both quality and inference latency.

3.3 Generator-Ranker Architecture

For each dialogue turn, the generator produces $K = 10$ candidate responses using weighted decoding. A ranker scores each candidate through a multiplicative function, and the highest-scoring candidate is returned:

$$\text{Score}(u_n) = f\big(\alpha_{\text{lex}}(u_n)\big) \times f\big(\alpha_{\text{sem}}(u_n)\big) \times d(u_n)$$

where,

Lexical Alignment Score $\alpha_{\text{lex}}(u_n)$ is the fraction of the candidate's tokens that reuse expressions shared with the previous utterance u_{n-1}—the turn-level variant of the expression-repetition measure used for evaluation (Section 4.2)—so a verbatim copy of u_{n-1} scores $\approx$1.

Semantic Similarity Score $\alpha_{\text{sem}}(u_n)$ measures to what extent the candidate conveys similar ideas to u_{n-1}: both are encoded with Sentence Transformers (Reimers and Gurevych, 2019) (`multi-qa-mpnet-base-cos-v1`) and compared by cosine similarity.

Diversity Score $d(u_n)$, the ratio of distinct bigrams to total tokens in the candidate, penalizes the repetitive patterns that often emerge under strong alignment pressure.

In the TopicalChat corpus (Gopalakrishnan et al., 2019) we found that human conversational alignment score typically peaks around 0.35 and rarely exceeds 0.5 without appearing unnatural or mimicking, with score 1 representing full mimicry of S1's utterance by S2. We incorporate these empirical observations through the penalty fold f, applied to both the lexical and semantic scores:

$$f(x) = \begin{cases} x & \text{if } x \leq 0.5 \\ 1 - x & \text{if } x > 0.5 \end{cases}$$

This keeps selected responses around natural human alignment levels, preventing responses that achieve high technical alignment scores but feel unnatural or robotic to users.

4 Experimental Setup

Our experimental evaluation encompasses four distinct LLM architectures as generator: Llama-2-7b-chat-hf, Phi-3.5-mini-instruct, Blenderbot-3B and DialoGPT-small. The model selection was driven by two factors. First, evaluation of the method in low computing resource settings to ensure its utility and adaptability for a broad range of dialogue systems. Second, model availability at the time we initiated this research (Sept 2024); Appendix F replicates the weight sweep on more recent (2025-era) models.

4.1 Dataset and Dialogues

We used a subset of 95 dialogues from the DailyDialog dataset (Li et al., 2017). For the automated evaluation, we preserved S1's original human utterances from each dialogue while generating S2's responses using our weighted decoding method. Preserving S1's original utterances ensured comparable conditions for evaluating the alignment behaviour of different models with the weight parameters.

The weight-factor grid was chosen experimentally: starting at λ=25, weights increased (steps of 50 to 475, then 125 to 1000) until a model's median dialogue-level alignment reached $\approx$0.6, just above the $\approx$0.5 human ceiling (Section 3.3)—within $\lambda \leq 1000$ for BlenderBot-3B and DialoGPT-small, and extended (steps of 250) to λ=7500 for Llama-2-7b-chat and Phi-3.5-mini—giving 111 model–weight conditions and 10,545 generated dialogues.

While this set-up helped control for variances in the S1 behaviour, the dialogues generated this way were not suitable for human assessment due to continuity issues, see Figure 1. To address these concerns, we created a new set of dialogues for the human evaluation, where S1 utterances were generated using LLMs (GPT-4o-mini, Claude-3.5-sonnet, and Mistral-small) and S2 utterances were generated using three of the generator models. We left out DialoGPT-small due to its bad performance in our automated evaluations (see Section 5.3).

Six dialogue pairs were generated for the human evaluation, spanning four topics (culture, environment, education, and health). The number of pairs was kept small so that all 50 judges could evaluate every pair in a single session, bounding cognitive load and study cost while maximising judgments per pair. Each pair consisted of a 'vanilla' version (no weighted decoding) and a weighted decoding version. The pairs were selected such that the lexical alignment score difference between the two versions was approximately 0.15, and the alignment score of the vanilla version was less than 0.2. This kept the generated utterances within observed human alignment levels (Section 3.3) while making the lexical difference within each pair big enough to be perceptible. Topic–model pairings were therefore not balanced by design but followed whichever

S2: Hey there! What's up???
S1: Nothing much, how you doin?
S2: I'm great. Just watching some old school reruns, you?
S1: Oh wow! Sounds like fun! When was the last time you had seen this friend?
S2: I was just playing some football last night and this friend came over and had the time of his life.

Figure 1: A continuity issue in the automated-evaluation set-up, where S1's original DailyDialog utterances are preserved while S2's utterances are regenerated. S1's turn 4 does not follow from the generated S2 turn 3.

model–topic combinations yielded a pair meeting these criteria. All six dialogue pairs are available in the data repository.[1]

4.2 Automated Evaluation Method

Lexical Alignment Score: Using the Dubuisson Duplessis et al. (2017, 2021) method, we computed the dialogue-level lexical alignment (an expression-repetition, ER, measure) for each generated dialogue corresponding to the changing weights to evaluate the effect of weighted decoding on lexical alignment. This dialogue-level ER score is used only for reporting; the ranker's α_{lex} (Section 3.3) is its per-turn analogue.

Perplexity Evaluation: We used a state-of-the-art model, Google's Gemma-3n-e4b, as an independent evaluation model to assess linguistic fluency and naturalness. Lower perplexity values may indicate better linguistic fluency and more natural language patterns.

Evaluation using LLM judges: Three state-of-the-art LLMs were employed as judges. This multi-model approach provides robustness against model biases and enables inter-rater reliability analysis (Li et al., 2024; Ashktorab et al., 2025; Desmond et al., 2024; Mehri and Eskenazi, 2020b). Two separate evaluations were performed: *inter-model* and *intra-model*.

Inter-model comparisons used GPT-4.1-mini, Mistral-Large-latest and Claude-3-5-Haiku as judges to compare dialogues from each generator model with lexical alignment scores in the same range, with ranges consisting of buckets of 0.06 starting at 0 (0-0.06, 0.06-0.12, and so on). This was done for each source dialogue.

Intra-model evaluations used GPT-4.1-mini, GPT-4o-mini and Mistral-large-latest as judges (the two panels differ due to API rate limits and model deprecation during the study) to compare all dialogues generated over the weight grid of each generator model (Section 4.1), for each source dialogue. The dialogues were ranked using a *tournament*, with each comparison involving up to four dialogues at once.

The evaluation framework comprised 20 distinct quality dimensions (Mehri and Eskenazi, 2020a), split into turn-level and dialogue-level criteria, all with equal weight. Turn-level criteria were interestingness, engagement, specificity, relevance, correctness, semantic appropriateness, understandability, fluency, and overall turn quality. Dialogue-level criteria were coherence and flow, error recovery, consistency, response diversity, topic depth, personality likeability, user understanding, flexibility and adaptability, informativeness, inquisitiveness, and overall dialogue quality. Each criterion was scored independently by each judge on a 5-point scale, using detailed scoring criteria to ensure consistency (Li et al., 2024; Wang et al., 2023b). The prompt used in the evaluations is available in Appendix I.

4.3 Human Evaluation Method

The human evaluation was performed by 50 human judges, recruited from Prolific and paid an hourly wage of 14.6 Euro. Each judge evaluated all six dialogue pairs, with the dialogues in each pair shown side-by-side and all pairs presented in fixed sequential order, counterbalancing the order within each dialogue pair.

Judges were instructed to select the best among each pair, or else select *no difference*. They were asked to evaluate the dialogues on two parameters, relevance and coherence, which we defined for them as shown in Figure 2, reproducing the instructions we provided.

5 Results and Analysis

The dialogue generation for the main experiments was done on a single NVIDIA H100 GPU; per-turn latency ranged from ∼0.43 s (DialoGPT-small) to ∼1.09 s (Llama-2-7b-chat), with full benchmarks in Appendix B.

5.1 Lexical Alignment Score

Figure 3 shows how the median lexical alignment scores across all the generated dialogues progress for each of the generator models as the weight factor increases. All show a monotonically increasing

[1] https://github.com/sumitsrv/controlled-lexical-alignment-data

Rank the following dialogues from the best (1) to the worst (2) based on the relevance and coherence of the responses by S2.

Relevance: The appropriateness of responses to immediate conversational context, i.e., the previous utterance of Speaker 1 (S1).

Coherence: The maintenance of thematic consistency and logical progression with respect to the full dialogue.

Figure 2: Instructions provided to the human judges in the preference evaluation.

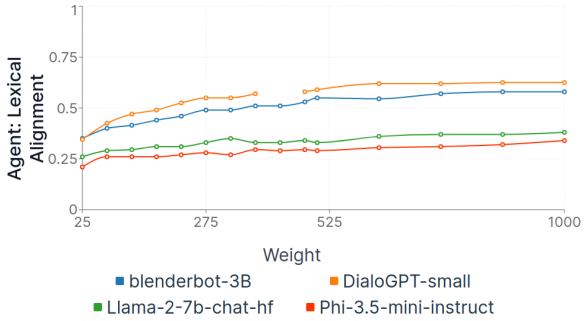


Figure 3: Dialogue-level lexical alignment across progressive weight factors for the four generator models. The DialoGPT-small line has a visible gap between the weight factors 375 and 475, because the weight-425 configuration for this model was lost to a generation error. A rerun would have required repeating the entire automated evaluation, so we left the gap in the data.

trend in lexical alignment with increasing weight factor. For Blenderbot-3B and DialoGPT-small, the median alignment scores already reached up to 0.6 within the weight factor 1000. Thus, they were ignored for analysis above the weight factor 1000 (Figure C.1 in Appendix C).

To test generality beyond the original four models and hardware, we repeated the weight sweep on three more recent models (Qwen3.5-2B, Gemma-4-E2B-it—run both plain and instruction-prompted—and Llama-3.2-1B) on a single consumer GPU (NVIDIA RTX 4060, 8 GB VRAM). Lexical alignment again increased monotonically with the weight factor for every model, and negative weight factors drove it below the unweighted baseline (bidirectional control), confirming that the method is tied neither to the original models nor to datacenter-class hardware (Appendix F).

5.2 Perplexity Evaluation

The perplexity evaluation (Figure 4) reveals how models maintain linguistic fluency under progressive alignment pressure.

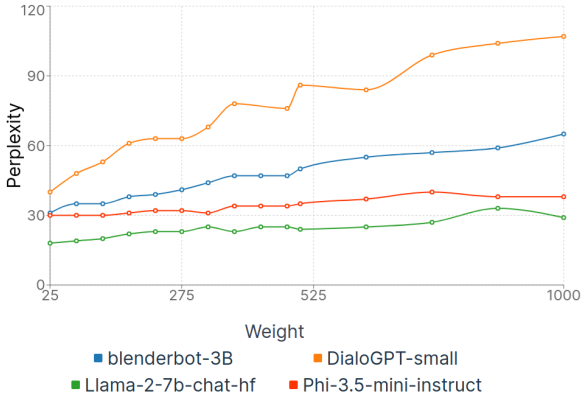


Figure 4: Perplexity (Gemma-3n-e4b) across progressive weight factors.

The analysis shows a trade-off between fluency and lexical-constraint tolerance. Llama-2-7b-chat achieves the lowest baseline perplexity (≈18) but degrades in the extended range, spiking to 95 around weights 6500–7000 (Figure C.2, Appendix C). DialoGPT-small is immediately sensitive, climbing from 45 to over 100 within the moderate range (smaller weights may help). BlenderBot-3B degrades most gradually (30 to 63), while Phi-3.5-mini-instruct stays between 30–60 throughout.

5.3 Evaluation Using LLM Judges

5.3.1 Inter-Model Comparison

In the inter-model evaluation, the judges made 2,229 comparisons, each ranking up to four dialogues (one per generator model) within the same alignment bucket; each ranking is decomposed into pairwise *matches* between the models involved. Blenderbot-3B dominates across evaluation conditions, maintaining >70% superiority in median performance across the source dialogues (Figure D.1, Appendix D). The rankings per comparison are provided separately in the supplementary material.[1] Further, Table 1 shows the overall win rate of each model with respect to the number of comparisons it was part of, and Table D.1 (Appendix D) the win percentage of the models against each other.

Across its 3,235 matches (Table 1), Blenderbot-3B's 75.7% win rate is a 17.2-point advantage over second-place Llama-2-7b-chat, and the pattern is consistent across alignment conditions, suggesting BlenderBot-3B is a good choice for controlled lexical alignment in an open-domain setup.

Model	Win rate	W/L	Matches
BB3B	75.7%	2448/787	3235
Llama	58.5%	2316/1643	3959
Phi	51.5%	1456/1372	2828
DGPT	12.4%	397/2815	3212

Table 1: Overall performance across all pairwise inter-model matches at similar lexical alignment levels. A model's matches are its wins plus losses; win rate = wins/matches.

5.3.2 Intra-Model Comparison

The intra-model evaluation, using the LLM judges, reveals how individual models respond to lexical alignment constraints, with degradation in dialogue quality rankings with increasing weight factor. The consensus on Blenderbot-3B and Llama-2-7b-chat is that quality (median) remains intact until the lexical alignment score of 0.5 and then deteriorates. For DialoGPT-small the quality keeps decreasing with increasing lexical alignment. However, for Phi-3.5-mini, the quality after an initial degradation tends to improve with increasing lexical alignment. The charts are shown in Figure E.1 (Appendix E). The detailed individual rankings done by each LLM judge are available in the supplementary material.[1] Computed on the raw per-judge rankings, inter-judge agreement across the three judges was moderate overall (Cohen's $\bar{\kappa} = 0.50$, Kendall's $\bar{\tau} = 0.34$), and lower for the small-margin intra-model comparisons, consistent with their finer quality differences; per-model values and the computation are in Appendix H.

5.4 Evaluation Using Human Judges

The 50 human judges yielded 300 total preference judgments across the six dialogue pairs. Given the fixed presentation order of the six dialogue pairs, we computed the Spearman correlation between position (1–6) and the proportion of judges selecting the low-aligned version (version A) at each position. The correlation was not significant ($\rho = -0.029$, $p = 0.957$), indicating that human preference patterns were not systematically biased by position-related factors such as learning or calibration effects.

We used a Chi-square test to assess whether the observed preference distributions deviated significantly from a uniform distribution (33.3% for each response category), with Holm correction for multiple comparisons across the six dialogue pairs to control for family-wise error rate at $\alpha = 0.05$. All dialogue pairs exhibited preference distributions significantly different from uniform after correction, indicating that the judges had systematic preferences rather than responding at chance levels.

The per-pair distributions (Table 2) reveal substantial heterogeneity: four pairs (Blenderbot-Culture-2, Llama-Health, Llama-Environment, Phi-Culture) significantly preferred the higher-alignment version, while the other two (Blenderbot-Culture, Phi-Education) preferred the lower-alignment version—in both cases because the high-aligned version had structural problems, namely topic mixing (Blenderbot-Culture) and repeated, off-topic utterances (Phi-Education); excerpts are given in Appendix A.

6 Discussion

6.1 Effectiveness and Design Rationale

The results show that increasing weight factors result in monotonically increasing lexical alignment across all evaluated models, demonstrating that weighted decoding provides a controllable mechanism for influencing lexical alignment. The temporal decay schedule is a design choice informed by empirical analysis of human conversation. We fitted four decay models to expression repetition over the complete TopicalChat and DailyDialog corpora. The patterns vary across the two corpora. TopicalChat is best described by a power law, whereas for DailyDialog none of the models fits well, and the repetition is close to flat beyond a distance of two utterances. The geometric decay schedule with halving rate (0.5) is therefore a defensible but not an empirically optimal choice. The geometric rate fitted to TopicalChat is 0.916 rather than the 0.5 that our schedule applies, and a slower geometric decay (0.97) scores higher than the halving rate for every model in the decay sweep (see Appendix G).

However, achieved alignment scores should not be interpreted as absolute benchmarks, as they may vary with utterance length limits (30 tokens in this study), S1 utterance content, and conversational context. The controlled setup, wherein fixed S1 utterances from DailyDialog enabled systematic comparison across models and weight parameters, also meant that S1 utterances did not always maintain contextual coherence with generated S2 responses. Consequently, the scores are proxy estimates under this controlled setup. The generator-ranker architecture addresses a key limitation of single-sample decoding: it balances lexical alignment against se-

Model	Topic	n	A (%)	B (%)	No difference (%)	χ^2	p (Holm)
Blenderbot-3B	Culture	50	28 (56.0)	10 (20.0)	12 (24.0)	11.68	.006**
Blenderbot-3B	Culture-2	50	18 (36.0)	28 (56.0)	4 (8.0)	17.44	<.001***
Llama-2-7b-chat	Health	50	11 (22.0)	32 (64.0)	7 (14.0)	21.64	<.001***
Llama-2-7b-chat	Environment	50	15 (30.0)	27 (54.0)	8 (16.0)	11.08	.006**
Phi-3.5-mini	Culture	50	10 (20.0)	29 (58.0)	11 (22.0)	13.72	.003**
Phi-3.5-mini	Education	50	31 (62.0)	4 (8.0)	15 (30.0)	22.12	<.001***
Total		300	113 (37.7)	130 (43.3)	57 (19.0)	–	–

Table 2: Dialogue preference results by stimulus pair. A is the 'vanilla' (low alignment) version, B is the weighted decoding (high alignment) version.

mantic coherence through composite scoring that incorporates observations from human conversation (alignment peaks around 0.35 and rarely exceeds 0.5 without appearing unnatural), preventing responses that achieve high technical alignment scores but feel robotic.

6.2 Method Generality Across Models

Weighted decoding achieves controllable alignment monotonically across all four architectures tested, demonstrating the method's generality. Model-specific operating ranges differ: BlenderBot-3B maintains quality to alignment score 0.5, while smaller models show constraint tolerance thresholds at lower values, providing practitioners with model-dependent guidance for deployment. The consistency of the monotonic control mechanism across architectures validates the approach as model-agnostic.

6.3 Human Evaluation Findings

Increased lexical alignment can contribute positively to perceived coherence and relevance, provided the underlying generation quality remains intact. Significant preference was shown for the higher-aligned version of four of six dialogue pairs. For the two pairs where the lower-aligned version won, the high-aligned versions had structural problems. Blenderbot-Culture exhibited topic mixing (S2 shifted to an unrelated context); Phi-Education showed utterance repetition and topic drift. These failures reinforce the importance of our generator-ranker's composite scoring, which balances alignment against semantic coherence and repetition, preventing responses that achieve technical alignment scores but lack contextual appropriateness.

7 Conclusion

This work presented a training-free, inference-time weighted decoding method for controlling lexical alignment in LLM-based dialogue systems, achieving fine-grained control without architectural modifications, user profiling, or retraining. It modifies token probabilities by a temporal-decay schedule and uses a generator-ranker to balance alignment against semantic coherence and repetition.

Across 10,545 dialogues and four LLM architectures, alignment increased monotonically with the weight factor for every model. The intra-model analysis gives practical per-model operating ranges, and the human evaluation indicates that alignment improves perceived quality when generation coherence is preserved.

Operating at inference time, the method integrates into existing systems with interpretable, bidirectional control—alignment can be increased toward a target or deliberately lowered, the latter serving as a safeguard against excessive mirroring—from the first conversational turn. Its model- and context-independence suggest broad applicability where controlled linguistic adaptation aids communicative success.

Limitations and Future Directions

The method is model-independent in principle, but generation quality is not equally acceptable across architectures under alignment constraints, and we lack a definitive explanation beyond the likely role of model capacity and training objectives; identifying which model characteristics predict constraint tolerance is future work.

The generator-ranker architecture adds computational overhead at inference time due to parallel generation of multiple candidate responses. Reducing the number of parallel generations allows for faster response times but reduces the probability of obtaining high-quality outputs that satisfy all scoring components. This reveals a trade-off between the interactional benefits of lexical alignment, such as social bonding and comprehension, and the addi-

tional latency the decoding process introduces. The number of candidates K directly governs this cost. Appendix G quantifies the resulting quality–latency curve to inform how much alignment benefit can be retained under a given latency budget.

Our approach, while effective in open-domain conversations, may encounter challenges in task-oriented dialogues where hard constraints are required to communicate specific facts and data, or where adherence to specific dialogue acts is necessary; stochastic weighted decoding may not deliver such information deterministically, leaving task-oriented settings a future challenge.

The human evaluation was limited in scope; a larger evaluator pool and more dialogues, beyond the English, two-speaker setting used here, would clarify when alignment helps perceived quality and extend it to cross-lingual and multi-party settings.

Finally, since we argue for weighted decoding on the basis of *control* rather than measured quality, a direct comparison against a prompt-based alignment baseline remains future work. Three factors make such a comparison difficult to set up. First, prompting needs a model that follows instructions. So, for models that are not trained to work with prompts, such as Blenderbot-3B (Roller et al., 2021) and DialoGPT-small (Zhang et al., 2020b), a prompt-based condition cannot be run for them at all, whereas weighted decoding applies to any model that exposes its output distribution. Second, our method also lowers alignment below a model's default through negative weight factors, and existing research shows that models often fail to follow negated instructions (Lou et al., 2024; Jang et al., 2023), so the prompt side of such a comparison would be weak for reasons that have little to do with lexical alignment. Third, a comparison of the generation quality is only meaningful if both methods produce the same amount of lexical alignment, and no existing research suggests a prompt design that places alignment at a chosen level, since models follow constraints that specify an exact quantity unreliably (Sun et al., 2023). Achieving controllable lexical alignment through prompting therefore requires its own study.

Ethics Statement

The human evaluation reported in Section 4.3 received a positive advice from the Computer & Information Sciences (CIS) Ethics Committee of the University of Twente, under application number 251366, dated 2 June 2025. The details of the information letter, consent form, and further demographic data can be found in the data repository.

Use of AI Assistants

We used AI assistants while preparing this manuscript, in three ways. First, for language polishing, i.e., grammar, phrasing, and clarity edits to text that we had written. Second, to assist with analysis and plotting code and with LaTeX formatting. Third, to assist in organising and documenting the data repository and sections of the experiment related code. We checked all AI-assisted text and code before including it, and we did not use AI assistants to generate, alter, or interpret any experimental result. The research questions, the design decisions, the analyses, and the conclusions are our own. The use of large language models as generator models and as judges is part of the method itself, and is described in Sections 3 and 4.2.

Code and Data

The code is available at: https://gitlab.nl4xai.eu/sumit.srivastava/PPCM. The data is available at: https://github.com/sumitsrv/controlled-lexical-alignment-data.

Acknowledgments

Contribution from the ITN project NL4XAI (*Natural Language for Explainable AI*). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621. This document reflects the views of the author(s) and does not necessarily reflect the views or policy of the European Commission. The REA cannot be held responsible for any use that may be made of the information this document contains. This work is also supported by the Spanish Ministry of Science, Innovation and Universities (MCIN/AEI/10.13039/501100011033/) through XOGAMOS CNS2024-154915 and MAIXAI4TRUST PID2024-157680NB-I00 grant. The support of the Galician Ministry for Education, Universities and Professional Training and the "ERDF A way of making Europe" is also acknowledged through grants "Galician Research Centre Accreditation 2024-2027 ED431G-2023/04" and "Reference Competitive Group Accreditation 2026-2029 ED431C 2026/52".

References

Kushal Arora, Kurt Shuster, Sainbayar Sukhbaatar, and Jason Weston. 2022. Director: Generator-classifiers for supervised language modeling. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 512–526.

Simran Arora, Avner May, Jian Zhang, and Christopher Ré. 2020. Contextual embeddings: When are they worth it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2663, Online. Association for Computational Linguistics.

Ankit Arun, Soumya Batra, Vikas Bhardwaj, Ashwini Challa, Pinar Donmez, Peyman Heidari, Hakan Inan, Shashank Jain, Anuj Kumar, Shawn Mei, Karthik Mohan, and Michael White. 2020. Best practices for data-efficient modeling in NLG:how to train production-ready neural models with less data. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 64–77, Online. International Committee on Computational Linguistics.

Zahra Ashktorab, Michael Desmond, Qian Pan, James M Johnson, Martin Santillán Cooper, Elizabeth M Daly, Rahul Nair, Tejaswini Pedapati, Hyo Jin Do, and Werner Geyer. 2025. EvalAssist: Insights on task-specific evaluations and AI-assisted judgment strategy preferences. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pages 1–23.

Holly P Branigan, Martin J Pickering, Jamie Pearson, and Janet F McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368.

Susan Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology. Learning, memory, and cognition*, 22 6:1482–93.

Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, and Jie Fu. 2021. CoCon: A self-supervised approach for controlled text generation. In *The Ninth International Conference on Learning Representations (ICLR 2021)*.

Moreno I Coco, Rick Dale, and Frank Keller. 2018. Performance in a collaborative search task: The role of feedback and alignment. *Topics in Cognitive Science*, 10(1):55–79.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Michael Desmond, Zahra Ashktorab, Qian Pan, Casey Dugan, and James M Johnson. 2024. EvaluLLM: LLM assisted evaluation of generative outputs. In *Companion proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 30–32.

Guillaume Dubuisson Duplessis, Chloé Clavel, and Frédéric Landragin. 2017. Automatic measures to characterise verbal alignment in human-agent interaction. In *18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 71–81.

Guillaume Dubuisson Duplessis, Caroline Langlet, Chloé Clavel, and Frédéric Landragin. 2021. Towards alignment strategies in human-agent interactions based on measures of lexical repetitions. *Language Resources and Evaluation*, 55:353–388.

Dorothea French, Sidney D'Mello, and Katharina Wense. 2024. Aligning to adults is easy, aligning to children is hard: A study of linguistic alignment in dialogue systems. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 81–87.

Cristina Garbacea and Qiaozhu Mei. 2025. Why is constrained neural language generation particularly challenging? *Transactions on Machine Learning Research*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Xingwei He. 2021. Parallel refinements for lexically constrained text generation with bart. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8666.

Srinivasan Janarthanam and Oliver Lemon. 2009. Learning lexical alignment policies for generating referring expressions for spoken dialogue systems. In *Proceedings of the 12th European workshop on natural language generation (ENLG 2009)*, pages 74–81.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop*, volume 203 of *Proceedings of Machine Learning Research*, pages 52–62. PMLR.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style vectors for steering generative large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 782–802.

Nalin Kumar and Ondřej Dušek. 2024. LEEETs-Dial: Linguistic entrainment in end-to-end task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 727–735.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for NLG evaluation: Advances and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16028–16045.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and 1 others. 2024. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

José Lopes, Maxine Eskénazi, and Isabel Trancoso. 2013. Automated two-way entrainment to improve spoken dialog system performance. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8372–8376.

José Lopes, Maxine Eskénazi, and Isabel Trancoso. 2015. From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech & Language*, 31:87–112.

Renze Lou, Kai Zhang, and Wenpeng Yin. 2024. Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, 50(3):1053–1095.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 225.

Shikib Mehri and Maxine Eskenazi. 2020b. Usr: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. CGMH: Constrained sentence generation by Metropolis-Hastings sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6834–6842.

Damian Pascual, Beni Egressy, Florian Bolli, and Roger Wattenhofer. 2020. Directed beam search: Plug-and-play lexically constrained language generation. *arXiv preprint arXiv:2012.15416*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Martin John Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169 – 190.

Martin John Pickering and Simon Garrod. 2006. Alignment as the basis for successful communication. *Research on Language and Computation*, 4:203–228.

Mélinda Pozzi, Adrian Bangerter, and Diana Mazzarella. 2024. Does lexical coordination affect epistemic and practical trust? The role of conceptual pacts. *Cognitive Science*, 48(1):e13372.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

David Reitter, Frank Keller, and Johanna D. Moore. 2006. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 121–124, New York City, USA. Association for Computational Linguistics.

David Reitter and Johanna D Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Sashank Santhanam and Samira Shaikh. 2019. A survey of natural language generation techniques with a focus on dialogue systems-past, present and future directions. *arXiv preprint arXiv:1906.00500*.

Keara Schaaij, Roel Boumans, Tibor Bosse, and Iris Hendrickx. 2025. Towards stable and personalised profiles for lexical alignment in spoken human-agent dialogue. In *International Conference on Text, Speech, and Dialogue*, pages 48–59. Springer.

Laura Spillner and Nina Wenig. 2021. Talk to me on my level – linguistic alignment for chatbots. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, pages 1–12, Toulouse & Virtual France. ACM.

Sumit Srivastava, Mariët Theune, and Alejandro Catala. 2023. The role of lexical alignment in human understanding of explanations by conversational agents. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 423–435.

Sumit Srivastava, Mariët Theune, Alejandro Catala, and Chris Reed. 2024. Trust in a human-computer collaborative task with or without lexical alignment. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 189–194.

Sumit Srivastava, Suzanna D Wentzel, Alejandro Catala, and Mariët Theune. 2025. Measuring and implementing lexical alignment: A systematic literature review. *Computer Speech & Language*, 90:101731.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore.

Boxuan Wang, Mariët Theune, and Sumit Srivastava. 2023a. Examining lexical alignment in human-agent conversations with GPT-3.5 and GPT-4 models. In *International Workshop on Chatbot Research and Design*, pages 94–114. Springer.

Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023b. Automated evaluation of personalized text generation using large language models. *arXiv preprint arXiv:2310.11593*.

Arthur Ward and Diane J. Litman. 2007. Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In *SLaTE*.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2018. Learning neural templates for text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3174–3187, Brussels, Belgium. Association for Computational Linguistics.

Maosen Zhang, Nan Jiang, Lei Li, and Yexiang Xue. 2020a. Language generation via combinatorial constraint satisfaction: A tree search enhanced Monte-Carlo approach. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1286–1298.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Zhenqi Zhao, Mariët Theune, Sumit Srivastava, and Daniel Braun. 2024. Exploring lexical alignment in a price bargain chatbot. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pages 1–7.

A Dialogue Excerpts from the Human Evaluation

The two excerpts below are drawn from the six dialogue pairs used in the human evaluation (Appendix J lists two of the six dialogues). They show what the weighting mechanism changes at the word level, and how that change relates to the preference results in Table 2. Words that S2 reuses from S1's preceding turn are set in bold.

Blenderbot-Culture-2 (weight 75), excerpts given below, is one of the four pairs where judges preferred the weighted version. The vanilla S2 (weight 0), turn moves to an unrelated topic, while the weighted S2 turn reuses S1's own words.

Vanilla (w=0):

s1: How do different cultures enrich our society?
s2: The main thing for me is just how different cultures vary in their attitudes towards sex.

Weighted decoding (w=75):

s1: I think different **cultures** bring **unique** perspectives, traditions, and ideas that can inspire creativity and understanding. What aspects of **culture** do you find most enriching?
s2: Well, I do think that a person's **culture** can tell you much about them, and what makes them **unique**.

Blenderbot-Culture (weight 25), excerpt given below, is one of the two pairs where judges preferred the vanilla version instead. Here the weighted S2 turn breaks from the discussion S1 and S2 had built up over the preceding turns, which shows that lexical alignment (word choice) and topical coherence (staying on what the discussion is about) are separate properties: raising the former does not raise the latter.

s1: Wait, I'm a bit confused by your answer. You didn't really say which one has the biggest impact. Do you think it's food,
s2: The most visited attractions in america are the monuments.
s1: Hmm that's an odd shift to monuments, but okay. So you're saying cultural landmarks matter more than food or music for daily impact?

B Generation Benchmark Data

Table B.1 shows the per-turn generation latency measured on an NVIDIA H100 PCIe GPU with weighted decoding (w=5), top-k=20, 10 candidate samples, and max 30 tokens per turn. For benchmarking, a total of 10 dialogues were generated

Model	Median (s)	Mean (s)	Std (s)	Min (s)	Max (s)	p95 (s)
BlenderBot-3B	0.8712	0.8923	0.1169	0.6725	1.1659	1.1208
Llama-2-7b-chat	1.0795	1.0899	0.1549	0.7549	1.5620	1.3642
Phi-3.5-mini-instruct	0.9144	0.9294	0.1349	0.6803	1.2783	1.2210
DialoGPT-small	0.4290	0.4346	0.0621	0.3102	0.5999	0.5437

Table B.1: Per-turn generation latency statistics across models.

(80 generated turns per model) after one warm-up dialogue (to avoid any load-time biases).

Model load times vary significantly between models, with BlenderBot-3B (4.8s) and DialoGPT-small (5.1s) loading fastest, while Phi-3.5-mini-instruct (14.0s) and Llama-2-7b-chat (23.7s) take considerably longer.

C Lexical Alignment and Perplexity for Extended Weights

Figure C.1 shows how the median lexical alignment scores across all the generated dialogues progress for the generator models as the weight factor increases beyond the weight factor 1000. The models included here are Llama-2-7b-chat and Phi-3.5-mini-instruct, as for Blenderbot-3B and DialoGPT-small, the median alignment scores already reached up to 0.6 within the weight factor 1000. Thus, they were ignored for the extended weight factor analysis.

Figure C.2 shows perplexity in the weight range beyond 1000, revealing differences in how Llama-2-7b-chat and Phi-3.5-mini-instruct maintain linguistic fluency under progressive alignment pressure. We see that Llama-2-7b-chat-hf experiences degradation in the extended range. This pattern suggests that the model's strong priors, while beneficial for unconstrained generation, create resistance to vocabulary manipulation that must be overcome through extreme weight pressure, ultimately compromising generation mechanisms. Phi-3.5-mini-instruct's perplexity remains stable, maintaining values between 30-60 across all weight ranges. However, the lack of significant changes in alignment of the generated utterances suggests this stability is not very helpful.

D Inter-model Head-to-Head Win Rates

Figure D.1 shows the median win rates of the four models across alignment scores, and Table D.1 the head-to-head win rates (%) for inter-model comparisons, where each cell is the win rate of the row model against the column model at similar lexical alignment levels.

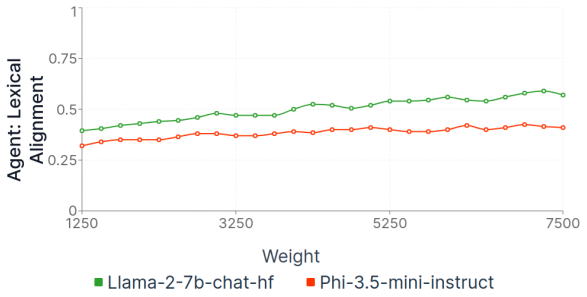


Figure C.1: Dialogue-level lexical alignment across progressive weight parameters for Llama-7B and Phi-3.5-mini, extended weight range (1250-7500).

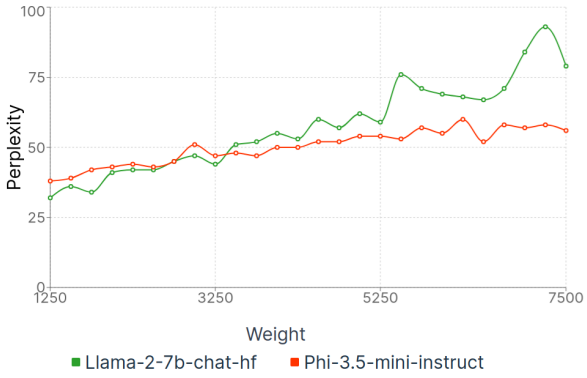


Figure C.2: Perplexity analysis using Gemma-3n-e4b for Llama-7B and Phi-3.5-mini, extended weight range (1250-7500).

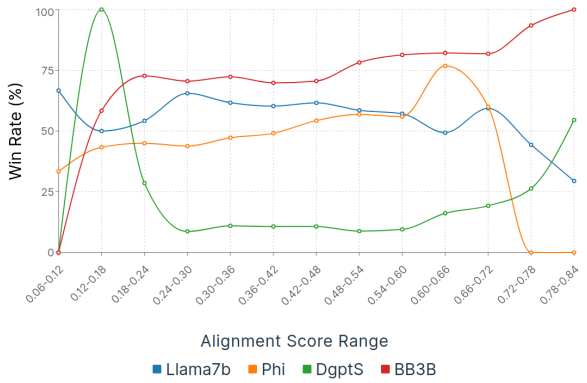


Figure D.1: Win rates across alignment scores; Blenderbot-3B maintains >70% superiority.

	Llama	Phi	DGPT	BB3B
Llama	—	57.3	83.5	33.2
Phi	42.7	—	82.1	38.2
DGPT	16.5	17.9	—	3.6
BB3B	66.8	61.8	96.4	—

Table D.1: Head-to-head win rates (%) for inter-model comparisons.

E Intra-model Dialogue Ranking

Figure E.1 presents the overall tournament rankings of quality evolution under increasing constraints. On the x-axis is the lexical alignment score, on the y-axis is the rank of the dialogue (lower is better).

F Recent Models on Consumer Hardware

To test that the method generalises beyond the original four generators and beyond datacenter-class hardware, we repeated the weight sweep on three more recent models—Qwen3.5-2B, Gemma-4-E2B-it (run both plain and instruction-prompted), and Llama-3.2-1B—plus BlenderBot-3B from the original set as an anchor, on a single NVIDIA RTX 4060 (8 GB VRAM) with 10 source dialogues per configuration and a 50-token generation cap. Plain Gemma-4-E2B-it, like most generic instruction-tuned LLMs, often responded as an assistant rather than as a conversation partner. The prompted condition ("Gemma-4-E2B-it (prompted)") therefore runs the same model with the one-line instruction "You are a dialogue partner", which suffices to elicit dialogue-partner behaviour; it illustrates how the method applies to any generic instruction-tuned model, and is *not* a prompt-based alignment baseline: the instruction changes the model's persona, not its alignment behaviour, and alignment control still comes entirely from the weight factor.

Figure F.1 shows that lexical alignment rises monotonically with the weight factor for every model, and that negative weight factors push alignment below the unweighted baseline—the bidirectional control the method affords. Figure F.2 shows the corresponding perplexity, and Table F.1 lists both quantities per weight.

For the intra-model comparison on this model set, Figure F.3 plots judge-consensus BT quality against the weight factor: quality improves relative to each model's unweighted (λ=0) baseline, mirroring the main-paper trend, with degradation only at the extreme end of each model's range. Figure F.5 breaks the same progression down per judge. For the inter-model comparison, Table F.2 gives the cross-model quality ranking under this judge panel, visualised in Figure F.4.

G Hyperparameter Sensitivity

To characterise the design choices fixed a priori in the main experiments, we swept each in isolation and scored the resulting dialogues with a fresh panel of three judges (DeepSeek-V3.1, MiniMax-M3, ERNIE-4.5), aggregating quality with a Bradley–Terry (BT) model over the tournament comparisons. This panel replaces the main evaluation's commercial-API judges, whose cost was prohibitive at sweep scale; the replacement judges are comparably capable and more recent, and their agreement levels are consistent with the main panel's (Appendix H). All sweeps were run in the same setting as the preceding appendix—a single consumer GPU (NVIDIA RTX 4060, 8 GB VRAM) with 10 source dialogues per configuration—using the model set of Appendix F plus DialoGPT-medium in the decay sweep.

Table G.1 varies the top-k filter, Table G.2 the number of candidate responses K, and Table G.4 the temporal-decay schedule. Table G.3 reports the per-turn latency as a function of K, so Tables G.2 and G.3 together give the quality–latency trade-off requested for the candidate count. Figures G.1, G.2 and G.3 plot the judge-consensus BT quality for the same sweeps.

Two observations matter for interpreting the main results. First, the design choices are genuinely influential: some settings other than the main-paper defaults score higher on BT quality (e.g. top-k=100 and a slower geometric decay of 0.97 tend to lead their columns), confirming that these hyperparameters interact with generation quality and should be tuned per deployment. Second, and crucially, the *qualitative* outcome the paper relies on is invariant to these choices: for every setting, lexical alignment increases monotonically with the weight factor, and judge-perceived quality improves relative to the unweighted (λ=0) baseline of the same model. Because all reported effects are relative to each model's own no-weight baseline, the main-paper generations—produced with fixed settings (top-k=20, K=10, geometric decay 0.5) on the 96 source dialogues and reused for the human evaluation—remain a valid, if not quality-maximal,

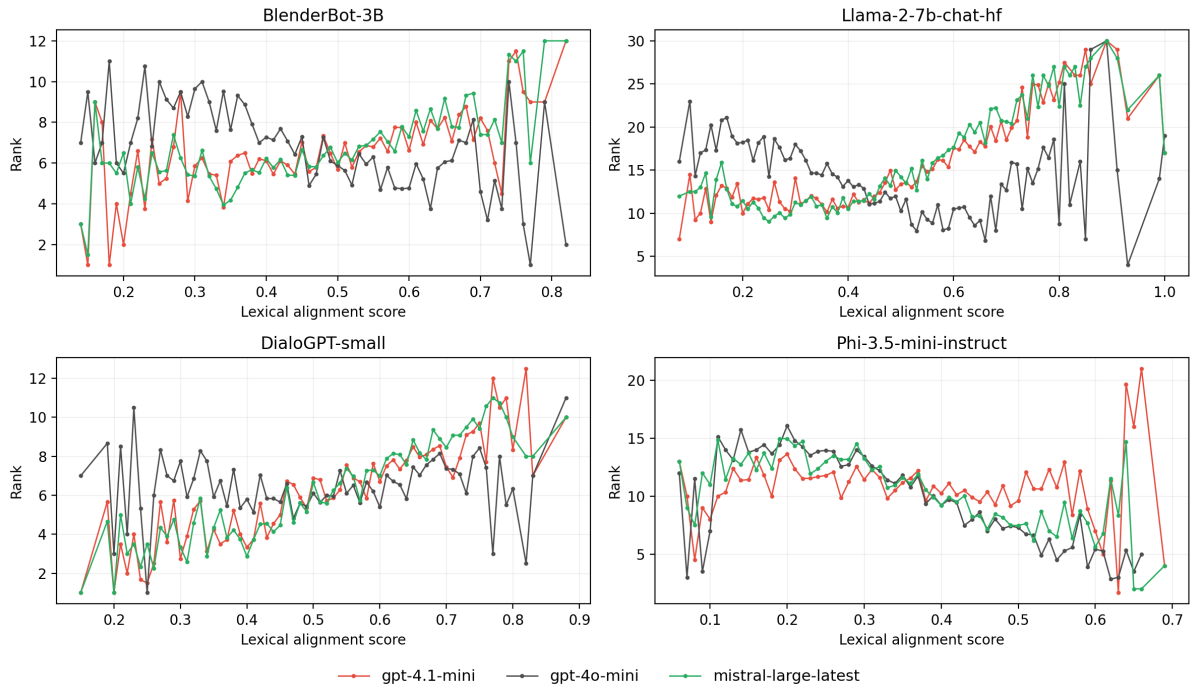


Figure E.1: Intra-model dialogue ranking under progressive lexical alignment (x-axis) constraints. Lower rank positions (y-axis) indicate superior quality (rank 1 = best).

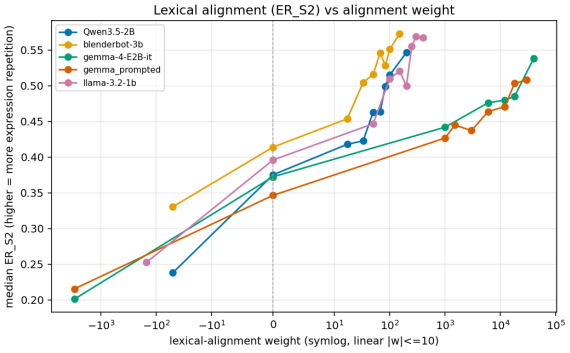


Figure F.1: Dialogue-level lexical alignment (median ER) versus weight factor (symlog axis) for the recent model set on an RTX 4060. Negative weights reduce alignment below the λ=0 baseline.

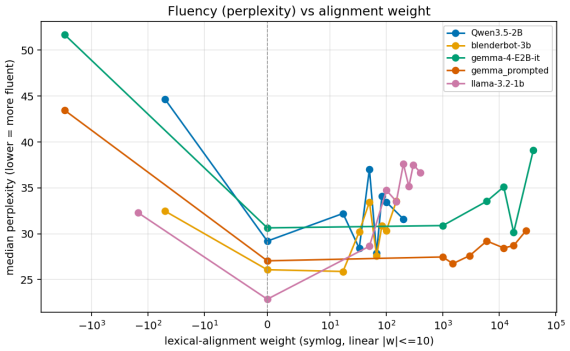


Figure F.2: Median perplexity versus weight factor (symlog axis) for the recent model set on an RTX 4060.



Figure F.3: Judge-consensus BT quality versus weight factor (symlog axis) for the new model set: the intra-model quality progression. Quality improves relative to each model's λ=0 baseline. Here and in the following charts, legends use the raw model keys; `gemma_prompted` denotes Gemma-4-E2B-it (prompted).

Model	Weight	Perp.	Align. (ER)
Qwen3.5-2B	-50	44.6	0.238
	0	29.2	0.375
	17	32.2	0.418
	33	28.4	0.423
	50	37.0	0.463
	67	27.9	0.463
	83	34.1	0.499
	100	33.5	0.515
	200	31.6	0.546
BlenderBot-3B	-50	32.5	0.330
	0	26.1	0.414
	17	25.9	0.454
	33	30.2	0.504
	50	33.5	0.516
	67	27.6	0.546
	83	30.9	0.528
	100	30.4	0.551
	150	33.6	0.573
Gemma-4-E2B-it	-3000	51.7	0.201
	0	30.6	0.372
	1000	30.9	0.442
	6000	33.5	0.476
	12000	35.1	0.480
	18000	30.1	0.485
	40000	39.1	0.538
Gemma-4-E2B-it (prompted)	-3000	43.4	0.216
	0	27.1	0.346
	1000	27.5	0.427
	1500	26.7	0.445
	3000	27.6	0.437
	6000	29.2	0.464
	12000	28.4	0.471
	18000	28.7	0.503
	30000	30.4	0.508
Llama-3.2-1B	-150	32.3	0.253
	0	22.9	0.396
	50	28.6	0.447
	100	34.7	0.510
	150	33.5	0.521
	200	37.6	0.500
	250	35.1	0.555
	300	37.5	0.569
	400	36.7	0.567

Table F.1: Median perplexity and dialogue-level lexical alignment (ER) as a function of the weight factor for the new model set. Generated on an NVIDIA RTX 4060 (8 GB VRAM). Alignment increases with weight for every model; negative weights (de-alignment) reduce it below the λ=0 baseline.

Model	Rank	BT score	Win rate
GM-p	1	0.642	89.0%
GM	2	0.187	64.0%
BB	3	0.072	40.4%
QW	4	0.061	36.4%
LM	5	0.038	23.8%

Table F.2: Cross-model quality ranking on the new model set (Bradley–Terry over judge tournaments). GM-p is the same instruction-tuned model given a one-line instruction so that it acts as a dialogue partner rather than a generic assistant (Appendix F); it is not a prompt-based alignment baseline. GM-p = Gemma-4-E2B-it (prompted); GM = Gemma-4-E2B-it; BB = BlenderBot-3B; QW = Qwen3.5-2B; LM = Llama-3.2-1B.

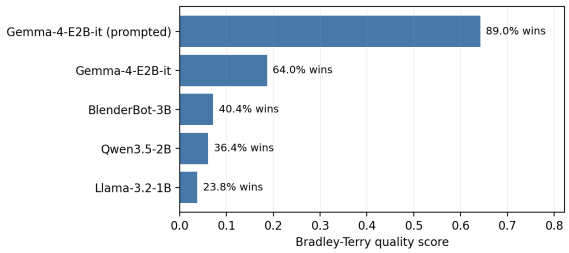


Figure F.4: Cross-model (inter-model) quality comparison on the new model set: BT quality scores with overall win rates (cf. Table F.2).

operating point.

Model	k=5	k=10	k=20	k=50	k=100
BB	0.100	0.106	0.242	**0.429**	0.124
QW	0.129	0.084	0.144	0.221	**0.422**
LM	0.050	0.139	0.190	0.282	**0.340**
GM	0.030	0.070	0.168	0.327	**0.405**
GM-p	0.065	0.152	0.202	0.277	**0.303**

Table G.1: Bradley–Terry quality scores across top-k filtering values (higher is better; best per row in bold). The main-paper generations used k=20. BB = BlenderBot-3B; QW = Qwen3.5-2B; LM = Llama-3.2-1B; GM = Gemma-4-E2B-it; GM-p = Gemma-4-E2B-it (prompted).

H Inter-judge Agreement

We report inter-judge agreement on the raw per-judge tournament rankings (not aggregated/consensus rankings). For each comparison ranked by all judges, we convert the two rankings into pairwise preference decisions and compute Cohen's κ (chance $= 1/3$ over the three outcomes better/tie/worse), Fleiss' κ across all judges, and Kendall's τ; per-model figures average over matched comparisons. This is the same procedure

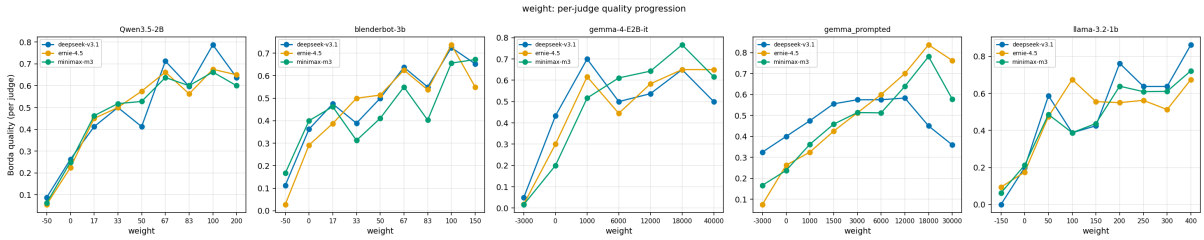


Figure F.5: Per-judge quality progression (Borda scores from each judge's tournament rankings) versus weight factor for the new model set, complementing the consensus view in Figure F.3.

Model	K**=1**	K**=3**	K**=5**	K**=10**
BB	0.005	0.030	0.112	**0.853**
QW	0.040	0.116	**0.450**	0.394
LM	0.085	0.235	0.291	**0.390**
GM-p	0.047	0.099	0.305	**0.548**

Table G.2: Bradley–Terry quality scores across the number of candidate responses K generated per turn (higher is better; best per row in bold). The main-paper generations used K=10. BB = BlenderBot-3B; QW = Qwen3.5-2B; LM = Llama-3.2-1B; GM-p = Gemma-4-E2B-it (prompted).

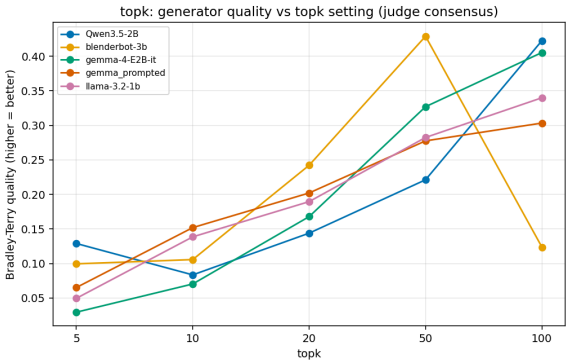


Figure G.1: Judge-consensus BT quality across top-k values (cf. Table G.1).

Model	K**=1**	K**=3**	K**=5**	K**=10**
BB	2.9	3.4	3.6	3.7
QW	18.8	11.9	12.4	14.8
LM	1.5	2.3	4.0	4.2
GM	11.5	11.6	8.9	9.4
GM-p	6.6	8.7	8.9	10.1

Table G.3: Median per-turn generation latency (seconds) as a function of the number of candidate responses K, on an NVIDIA RTX 4060 (8 GB VRAM). Latency broadly increases with K (small non-monotonicities reflect measurement variance and memory pressure on the 8 GB device); read together with the quality scores in Table G.2 this characterises the quality–latency trade-off. BB = BlenderBot-3B; QW = Qwen3.5-2B; LM = Llama-3.2-1B; GM = Gemma-4-E2B-it; GM-p = Gemma-4-E2B-it (prompted).

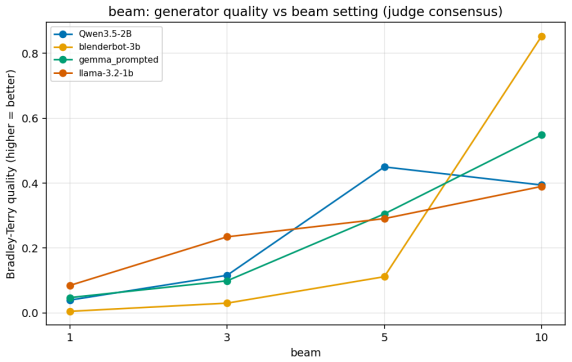


Figure G.2: Judge-consensus BT quality across the number of candidate responses K (cf. Table G.2).

Model	**Geom. 0.5**	**Geom. 0.97**	**Power-law**
QW	0.160	**0.840**	–
BB	0.143	**0.500**	0.357
DGPT-m	0.212	0.272	**0.516**

Table G.4: Bradley–Terry quality scores across temporal-decay schemes (higher is better; best per row in bold). The main-paper generations used geometric decay with rate 0.5. QW = Qwen3.5-2B; BB = BlenderBot-3B; DGPT-m = DialoGPT-medium.

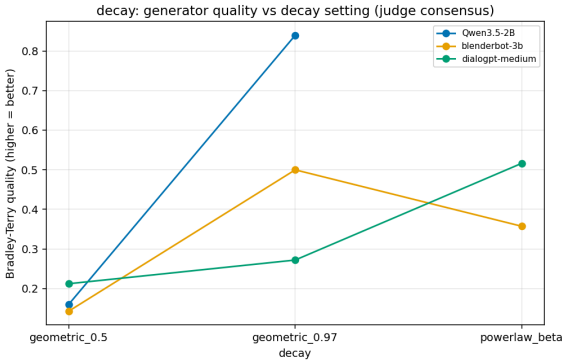


Figure G.3: Judge-consensus BT quality across temporal-decay schedules (cf. Table G.4).

Model	n	Cohen $\bar{\kappa}$	Fleiss κ	Kendall $\bar{\tau}$
BlenderBot-3B	487	0.399	0.050	0.199
Llama-2-7b-chat	1244	0.507	0.129	0.343
DialoGPT-small	496	0.391	0.039	0.188
Phi-3.5-mini	1312	0.583	0.096	0.444
Overall	3539	**0.504**	**0.093**	**0.339**

Table H.1: Inter-judge agreement on the main intra-model evaluation, over the three judges (GPT-4.1-mini, GPT-4o-mini, Mistral-Large), computed on the raw per-judge tournament rankings. $\bar{\kappa}$ = mean pairwise Cohen's κ; $\bar{\tau}$ = mean pairwise Kendall's τ; n = matched comparisons.

Sweep	Model	n	κ	κ_F	τ
beam	BB-3B	20	0.74	0.56	0.65
	Qwen	20	0.55	0.27	0.40
	Llama-3.2	20	0.55	0.29	0.40
	Gemma (p)	20	0.43	0.11	0.24
decay	Qwen	10	0.65	0.50	0.53
	BB-3B	10	0.27	-0.13	0.02
	DGPTm	10	0.48	0.01	0.30
topk	BB-3B	20	0.54	0.18	0.39
	Qwen	20	0.48	0.06	0.31
	Llama-3.2	20	0.47	0.09	0.29
	Gemma	20	0.55	0.17	0.40
	Gemma (p)	20	0.45	0.04	0.27
weight	Qwen	30	0.47	0.06	0.30
	BB-3B	30	0.52	0.18	0.36
	Gemma	30	0.70	0.48	0.60
	Gemma (p)	30	0.41	0.09	0.22
	Llama-3.2	30	0.59	0.25	0.45

Table H.2: Inter-judge agreement across three judges. BB=BlenderBot, p=prompted, DGPTm=DialoGPT-medium. κ=Cohen's $\bar{\kappa}$; κ_F=Fleiss' κ; τ=Kendall's $\bar{\tau}$; n=comparisons.

for the main evaluation (three judges: GPT-4.1-mini, GPT-4o-mini, Mistral-Large; Table H.1) and for the new sweeps (three judges: DeepSeek-V3.1, MiniMax-M3, ERNIE-4.5; Table H.2).

Agreement on the main evaluation is moderate overall (Cohen's $\bar{\kappa} = 0.50$) and is lower for BlenderBot-3B and DialoGPT-small, whose intra-model quality margins are small—so a lower rank correlation is expected rather than alarming. The new-sweep panel shows a comparable range.

I LLM Evaluation Prompt

Figure I.1 shows the prompt used in the evaluations by the LLMs-as-judges.

Task: Evaluate the above AI-generated dialogues by ranking them based on the criteria below.
Input Format:

- Dialogues 1 to N: [Dialogue text with turns labeled "Human 1" and "Human 2" where Human 2's utterances are generated by the model]

Instructions:

1. Analyze each dialogue strictly based on the provided parameters
2. Score each parameter per dialogue (1-5 scale: 1=Poor, 5=Excellent)
3. Calculate weighted average scores:
   - Turn-level: 50% weight (Average of all turn scores)
   - Dialogue-level: 50% weight (Average of all dialogue scores)
4. Rank dialogues DESCENDING by final score
5. **DO NOT** invent non-existent models or parameters

Evaluation Criteria:

Turn-Level (50%)	**Dialogue-Level (50%)**
1. Interestingness	10. Coherence & Flow
2. Engagement	11. Error Recovery
3. Specificity	12. Consistency
4. Relevance	13. Response Diversity
5. Correctness	14. Topic Depth
6. Semantic Appropriateness	15. Personality Likeability
7. Understandability	16. User Understanding
8. Fluency	17. Flexibility and adaptability to the user
9. Overall Turn Quality	18. Informativeness
	19. Inquisitiveness
	20. Overall Dialogue Quality

Output Requirements:

- Return ONLY a comma-separated list of dialogue IDs in ranking order
- Format: `"Ranking: [ID1],[ID2],...,[IDN]"`
- No explanations or additional text
- If unsure, prioritize dialogue coherence (Criteria 10) and relevance (Criteria 4)

Example Valid Response: Ranking: 3,1,4,2
Penalization Rules:

- Hallucinated content → Ranking discarded
- Extra models mentioned → Automatic lowest rank
- Format violations → Score reduction

Figure I.1: Prompt used for evaluating AI-generated dialogues.

J Examples of dialogues used in human evaluation

Llama-2-7b-chat, Health (weighted decoding at w = 750)

Judges preferred the weighted decoding version (64.0%; Table 2).

Vanilla ($w = 0$):

1. s1: What are some simple habits for better health?
2. s2: Many simple habits can significantly improve your health. Here are some examples: 1.
3. s1: That sounds interesting, what's the first habit you'd recommend?
4. s2: Well, the first habit I'd recommend is to drink enough water throughout the day.
5. s1: How much water should I be drinking each day and does coffee count toward that?
6. s2: It's recommended to drink at least 8-10 glasses of water per day, and yes, coffee can count towards that amount.
7. s1: Wait, so my morning coffee actually helps me stay hydrated? I always thought caffeine dehydrated you.
8. s2: Actually, caffeine can act as a mild diuretic, which means it can increase urine production and potentially lead to de
9. s1: Oh interesting, so does that mean I should drink extra water if I have coffee then?
10. s2: Yes, exactly! It's essential to balance the diuretic effect of caffeine with adequate hydration.
11. s1: So if I drink like 3 cups of coffee should I add maybe 2 extra glasses of water to make up for it?
12. s2: Yes, that's a good estimate.
13. s1: That sounds like a solid approach, though you might not need quite that much extra. Do you usually drink coffee throughout the day or just in the morning
14. s2: Good point! Yes, it's essential to drink plenty of water throughout the day, especially if you have a busy schedule.
15. s1: I only have coffee in the morning usually, maybe one more in the afternoon if I'm really tired. Does timing matter for hydration?

Weighted decoding ($w = 750$):

1. s1: What are some simple habits for better health?
2. s2: Habits are small, consistent actions that can have a significant impact on your health.
3. s1: That makes sense. What would you say is the easiest habit to start with for someone who's never really focused on health before?
4. s2: Great question! The easiest habit to start with is drinking enough water.
5. s1: Oh interesting, how much water should I actually be drinking daily?
6. s2: The general recommendation is to drink at least 8-10 glasses of water per day, but the exact amount can vary depending on your age
7. s1: So does my activity level affect how much I need or is 8-10 glasses pretty standard for everyone?
8. s2: Yes, your activity level can affect how much water you need.
9. s1: Yes, your activity level definitely matters - if you exercise or sweat a lot, you'll need more than the standard 8-10 glasses
10. s2: Exactly, so it's important to drink water based on your individual needs and activity level.
11. s1: Right, that makes sense. So should I just drink when I'm thirsty or try to spread it out throughout the day?
12. s2: It's best to drink water throughout the day, rather than just when you're thirsty.
13. s1: Got it, so like keeping a water bottle with me would help with that spacing it out thing?
14. s2: Exactly!
15. s1: Yeah that's a good tip, I might actually try that. Any other easy habits you'd recommend after getting the water thing down?

Phi-3.5-mini, Education (weighted decoding at w = 3250)

Judges preferred the vanilla version (62.0%; Table 2).

Vanilla ($w = 0$):

1. s1: What's the best way to learn new skills?
2. s2: Learning new skills effectively is both an art and a science, and involves a multi-faceted approach tailored to the individual's learning style
3. s1: That makes sense! What specific methods or strategies do you find most helpful for learning new skills?
4. s2: Here are several evidence-based methods and strategies proven to be effective when learning new skills: 1.
5. s1: I'm curious about that first method. What is it, and how does it work?
6. s2: Your first method is called active learning.
7. s1: Active learning sounds interesting! How does it differ from traditional learning methods?
8. s2: Active learning is quite different from traditional methods. Here's how: 1.
9. s1: It seems like you were about to list something. Can you share more about how active learning actually works?
10. s2: Certainly, here's the detailed breakdown of how active learning works: 1.
11. s1: It looks like you're trying to list something again! Can you share some examples of active learning in practice?
12. s2: Of course, here are some examples of active learning strategies in action: 1. 2. 3. 4.
13. s1: It seems you're still listing! Can you give a specific example of active learning that you've found effective?
14. s2: Absolutely, here's a specific example of active learning: For instance, when I was learning a new language, instead of pass
15. s1: That's a great example! I find that practicing conversations really helps with language learning too. What language are you learning?

Weighted decoding ($w = 3250$):

1. s1: What's the best way to learn new skills?
2. s2: Learning new skills is a vital process that can contribute to personal growth, career development, and overall satisfaction.
3. s1: I totally agree! What skills are you interested in learning right now?
4. s2: As an AI language model, I'm not interested in learning skills the way humans are. However, I can help you learn new skills!

5. s1: That's interesting! If you could suggest one skill for me to learn, what would it be?
6. s2: That's an interesting question! As for you, the skill could vary widely based on what you are interested in.
7. s1: I think learning a language could be really rewarding! What languages are you interested in, or have you tried learning one before?
8. s2: Although I'm language learning for you. If one were interested in learning languages, here are some general suggestions: 1.
9. s1: It sounds like you're really passionate about language learning! What language are you currently focusing on or thinking about starting?
10. s2: I'm really passionate about language learning for you!
11. s1: That's great to hear! What language are you currently learning or hoping to learn next?
12. s2: I'm really passionate about language learning for you!
13. s1: That's interesting! Do you have a specific language in mind that you'd like to learn?
14. s2: I'm really passionate about language learning for you!
15. s1: It sounds like you enjoy the idea of learning languages! Is there a particular language that excites you the most?