Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

CERA: Context-Engineered Reviews Architecture for Synthetic Dataset Generation

Kap Thang[†,*] Danial Ebrat[†,*] Luis Rueda[†,*]
[†] School of Computer Science, University of Windsor

Abstract

Aspect-Based Sentiment Analysis (ABSA) models require large-scale annotated datasets that are scarce, expensive to create, and suffer from class imbalance. While Large Language Models (LLMs) offer promising synthetic data generation, existing approaches lack factual grounding and provide limited aspect-level control. We present CERA (Context-Engineered Reviews Architecture), a *training-free* framework that generates realistic, controllable synthetic review text for ABSA through structured *context engineering*, i.e., carefully composing what an LLM receives as input rather than modifying the model itself. CERA's three-phase pipeline integrates agentic web-search factual grounding with multi-agent verification, demographic-grounded persona diversity, and configurable polarity balance. Evaluated across three review domains and four architectures, CERA achieves Real-data-level corpus diversity (Distinct-2 of 0.736 vs. Real's 0.776) while heuristic prompting collapses to 0.254, and scales to 8,000 reviews without quality degradation. Human evaluation confirms CERA reviews approach chance-level detection in a triplet Turing test (30% vs. 33% chance), nearly twice the rate of heuristic prompting (18%).

Keywords: Synthetic Data Generation, Aspect-Based Sentiment Analysis, Large Language Models, Controllable Text Generation, Multi-Agent Verification

1. Introduction

Aspect-Based Sentiment Analysis (ABSA) enables fine-grained opinion mining toward specific product attributes [1], but model development faces critical data challenges: (1) *data scarcity*, as SemEval benchmarks [1] contain only ∼3,000 sentences per domain; (2) *class imbalance*, with Amazon reviews skewing ∼65% positive [2]; and (3) *domain sparsity*, where niche domains lack annotated data [3]. Synthetic data generation via Large Language Models (LLMs) offers a promising alternative [3], but existing prompt-based approaches [4, 5] operate at document-level sentiment without aspect-level control. The "polite phenomenon" [6], where LLMs bias output toward positive sentiment, further exacerbates class imbalance, and generated text tends to be formulaic. Fine-tuning approaches (MAGIC [7], MAPLE [8], Review-LLM [6]) advance multi-aspect control but require model fine-tuning, limiting accessibility. Recent work by Hellwig et al. [9] explores few-shot prompting with GPT-3.5 and Llama-3 for synthetic ABSA data in low-resource settings, demonstrating that LLM-generated samples can improve aspect category detection; however, their approach lacks factual grounding, multi-agent verification, and corpus-level diversity enforcement, leaving generated text vulnerable to hallucination, vocabulary collapse, and the polite phenomenon at scale.

[*] thangk@uwindsor.ca, debrat@uwindsor.ca, lrueda@uwindsor.ca

We argue that *context engineering*, i.e., structuring what an LLM receives as input rather than modifying the model, can achieve controllable, high-quality generation without fine-tuning. We present CERA (Context-Engineered Reviews Architecture), a training-free framework[1] with four contributions: (1) context engineering as a training-free paradigm for synthetic ABSA data, requiring only commodity LLM API access; (2) agentic factual grounding with cross-provider multi-agent verification to reduce hallucinations; (3) demographic-grounded diversity through persona generation across seven orthogonal dimensions, with configurable polarity balance counteracting the polite phenomenon; and (4) comprehensive empirical evidence across three domains, four architectures, and human evaluation, with open-source tooling[2,3].

2. Methodology

CERA[1] is a modular, training-free framework whose pipeline flows through three phases (Figure 1). The *Composition Phase* produces three context documents that fully resolve all creative decisions before any text is generated. The Subject Intelligence Layer (SIL) operates as an agentic web-search system where LLMs gather current subject information, overcoming training data cutoff limitations. SIL's output is verified through Multi-Agent Verification (MAV), a five-round cross-provider consensus protocol using $N \geq 3$ reasoning-capable LLMs that reduces correlated hallucinations [10] through independent research, semantic deduplication (τ=0.85 via Sentence-BERT [11]), and $\lceil 2/3 \times N \rceil$ majority voting, producing the verified `Subject Context`. The Reviewer Generation Module (RGM) enforces diversity across seven orthogonal dimensions (persona identity, vocabulary variation, narrative flow, sentiment arc, connective style, evidence approach, and sentence rhythm) by dynamically sizing the persona pool to 90% of the target review count and generating domain-specific vocabulary alternatives and structure templates (`Reviewers Context`). The Attributes Composition Module (ACM) structures polarity distribution (e.g., Realistic 65/15/20% [2]), noise configuration, and generation parameters into the `Attributes Context`, enabling the polarity control that counteracts the polite phenomenon.

The *Generation Phase* is governed by *zero creative latitude*: all creative decisions are made during composition, and the generation LLM simply converts fully-resolved blueprints into natural language. The Authenticity Modeling Layer (AML) assembles per-review prompts from the three context documents, invoking the Diversity Enforcement Module (DEM) to inject diversity signals, including phrase-frequency monitoring, reference style injection, varied opening directives, and capitalization variation, without additional LLM calls. DEM's feedback loop, where earlier outputs inform later constraints, sustains diversity at scale. Post-generation noise injection via nlpaug [12] adds authentic imperfections at character, lexical, and sentence levels. The *Evaluation Phase* uses Multi-Dimensional Quality Assessment (MDQA) across three complementary axes: *Lexical Quality* (BLEU [13], ROUGE-L [14]), *Semantic Fidelity* (BERTScore [15], MoverScore [16]), and *Corpus Diversity* (Distinct-1/2 [17], Self-BLEU [18]), distinguishing genuine quality from formulaic repetition.

3. Experimental Setup

We design four experiments to evaluate CERA's effectiveness, component contributions, scaling behaviour, and cross-domain generalizability; full tables are deferred to the Appendix. We evaluate across three domains: Laptop (SemEval-2015 [19]; 2,548 sentences),

[1] https://github.com/project-cera/cera
[2] https://github.com/project-cera/cera-LADy
[3] https://github.com/project-cera/cera-human-eval

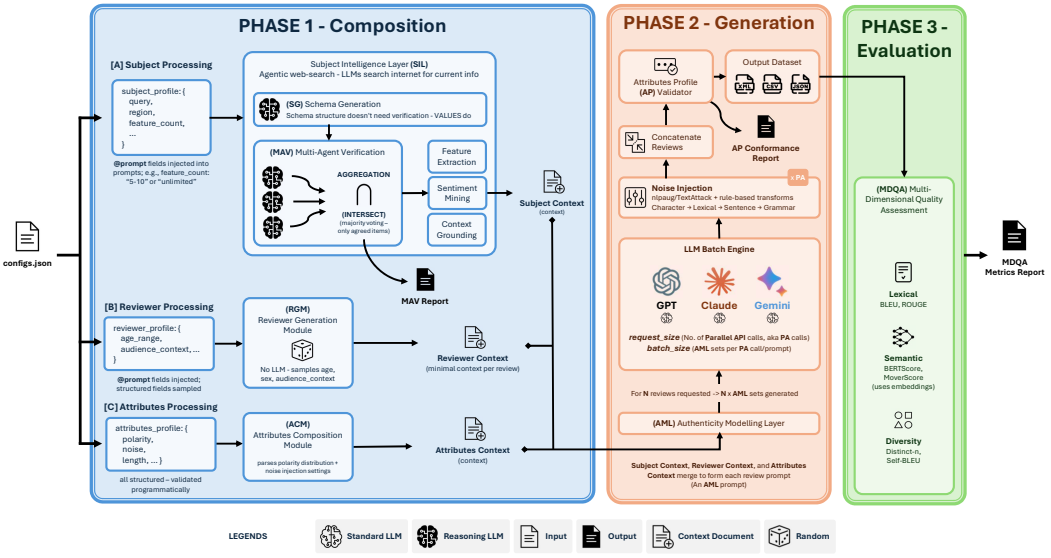


Figure 1. CERA pipeline: Composition phase (blue) researches subject via SIL, samples demographics via RGM, and structures attributes via ACM; Generation phase (orange) combines contexts in AML and produces reviews; Evaluation phase (green) assesses quality via MDQA.

Restaurant (SemEval-2015/16 [19, 20]; 4,048 sentences), and Hotel (OATS [21]; 9,793 sentences). *Intrinsic* evaluation uses MDQA against the corresponding real dataset; *extrinsic* evaluation uses cera-LADy [22], an extension of the LADy toolkit [23], with four architectures (RND, BTM [24], CTM [25], BERT [26]) and primary metric P@5 (Table 2 in the Appendix). The Heuristic baseline is *not* a naive prompt; it receives domain knowledge from the reference dataset (demographics, polarity distribution, aspect categories) but lacks CERA's composition pipeline: no SIL, MAV, RGM, or ACM. All generation uses Qwen3-235B-A22B [27] as a fixed LLM via OpenRouter [28]; composition uses three cross-provider reasoning models at approximately $0.90 per run, while generation incurs zero cost via the free tier (∼400M tokens total). End-to-end pipeline latency is approximately 8 minutes per 1,000 reviews. We use 5 runs for cross-method comparisons and 3 for within-method analyses (paired t-tests, $\alpha = 0.05$).

4. Results and Discussion

Table 1 compares CERA, Heuristic, and Real data on the Laptop domain at two representative sizes (full results in Table 3), revealing a clear *quality–diversity trade-off*. CERA's diversity metrics closely track the Real-data ceiling, while Heuristic exhibits severe vocabulary collapse (fewer than 8% of unigrams are unique at n=1,000). Heuristic scores higher on reference-similarity metrics, but this advantage is modest: at n=1,000, Heuristic leads by 8–29% on lexical/semantic metrics, whereas CERA leads by 190% on Distinct-2 and 42% on Self-BLEU ($p < .001$, d>10). This asymmetry reflects Heuristic *overfitting to the reference distribution*, recycling similar phrasing that overlaps with the fixed corpus, rather than superior quality. CERA also structurally eliminates the polite phenomenon: AML's blueprint pre-assigns each aspect its polarity, so the LLM never needs to "choose" to be negative.

Component ablation on the Laptop domain (Table 4 in the Appendix) confirms each submodule's role. RGM removal produces the most dramatic effect: Self-BLEU increases

Table 1. MDQA intrinsic quality on the Laptop domain (mean ± SD across 5 runs). **Bold** marks the best synthetic method per metric and size. Full results across all five sizes in Table 3 (Appendix). Metrics: ^{L} Lexical, ^{S} Semantic, ^{D} Diversity.

	$n = 100$			$n = 1{,}000$		
Metric	*Real*	CERA	Heur.	*Real*	CERA	Heur.
BLEU^{L} ↑	*0.077*	0.058 ± .013	**0.073** ± .017	*0.087*	0.048 ± .007	**0.055** ± .020
ROUGE-L^{L} ↑	*0.183*	0.145 ± .008	**0.185** ± .010	*0.187*	0.143 ± .004	**0.180** ± .017
BERTScore^{S} ↑	*0.415*	0.427 ± .018	**0.483** ± .019	*0.522*	0.403 ± .013	**0.519** ± .013
MoverScore^{S} ↑	*0.594*	0.553 ± .012	**0.594** ± .020	*0.585*	0.543 ± .014	**0.589** ± .009
Distinct-1^{D} ↑	*0.505*	**0.504** ± .010	0.265 ± .012	*0.268*	**0.259** ± .015	0.070 ± .001
Distinct-2^{D} ↑	*0.918*	**0.897** ± .008	0.594 ± .017	*0.776*	**0.736** ± .019	0.254 ± .001
Self-BLEU^{D} ↓	*0.493*	**0.452** ± .013	0.766 ± .014	*0.535*	**0.450** ± .011	0.776 ± .009

from 0.450 to 0.619 at n=1,000 (+37% inter-review similarity), while Distinct-2 drops from 0.736 to 0.472, approaching Heuristic-level collapse. SIL removal causes reference-similarity metrics to *increase* (BERTScore 0.445 vs. 0.403), because the model falls back on pre-training knowledge overlapping more with the 2015 corpus. SIL's primary contribution, factual accuracy, is invisible to automated metrics but critical for trustworthy training data. Figure 2 shows that CERA produces more diverse text than real reviewers at every dataset size from n=25 to n=8,000 in the Hotel domain, with Distinct-2 advantage growing from +0.8pp to +8.6pp and reference-similarity metrics showing no degradation at scale. CERA's quality profile is also *domain-invariant* across laptop, restaurant, and hotel domains (reference-similarity $\sigma < 0.01$; diversity variance $\sigma = 0.031$–0.048 reflects inherent domain characteristics).

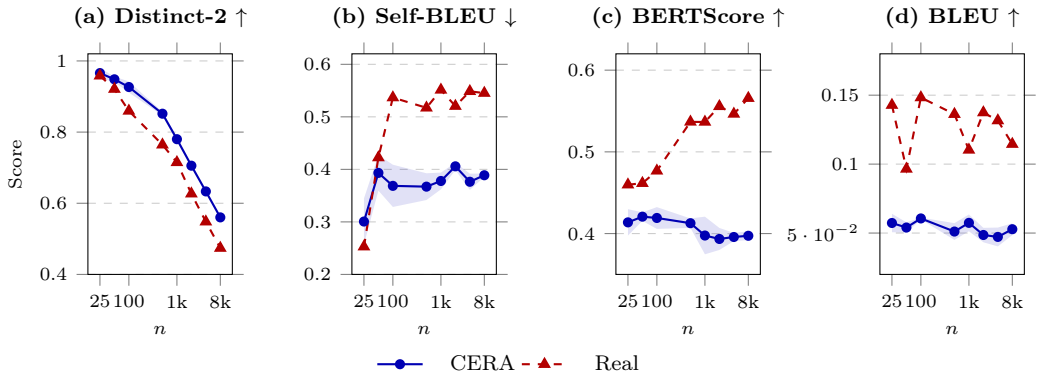


Figure 2. Scaling behavior of CERA vs. Real across eight dataset sizes (25–8,000) in the Hotel domain. CERA maintains *higher* corpus diversity than Real data at every size (a, b), while reference-similarity metrics show a stable gap (c, d). Shaded bands show ±1 SD across 3 runs; narrowing bands at larger sizes indicate increasing reproducibility. Real: single subsamples from the 9,793-sentence OATS corpus.

Finally, we conducted a human evaluation with a small sample of about 50 participants via two forced-choice tasks (Table 5 in the Appendix). In a three-way forced choice (33% chance baseline), Real reviews were correctly identified 53% of the time, CERA 30%, and Heuristic only 18%. CERA was selected as real nearly twice as often as Heuristic, confirming that structured composition produces more human-like text.

5. Conclusion

We presented CERA, a training-free framework demonstrating that context engineering, i.e., carefully structuring LLM inputs rather than modifying the model itself, can produce high-quality, controllable synthetic ABSA datasets using only commodity API access at $0.90 per pipeline run. Our experiments yield three principal findings. First, CERA achieves Real-data-level corpus diversity while heuristic prompting collapses despite receiving equivalent domain knowledge, confirming that structured composition, not model capability, drives output quality. Second, component ablation reveals that RGM is the primary diversity driver while SIL provides factual grounding that automated metrics cannot capture. Third, CERA's quality profile is both domain-invariant (σ<0.01) and scale-invariant (quality at n=8,000 matches n=25). Human evaluation (N=50) further validates these findings: CERA approaches chance-level detection (30% vs. 33%), nearly twice the rate of Heuristic (18%).

Some aspects of this work that can be further explored include: (1) SIL requires web-searchable subjects, and minimal online presence may reduce context quality, though user-provided seed data can mitigate this; (2) aspects emerge from Subject Context rather than explicit balancing; (3) LLMs may have encountered SemEval datasets during pre-training, though SIL's factual grounding partially mitigates this concern; and (4) all evaluators are CS-trained, making results conservative but demographically homogeneous. We plan a dedicated SIL factual grounding ablation comparing multi-agent verification against single-agent and no-SIL baselines across both pre- and post-training-cutoff subjects, to quantify SIL's contribution beyond what automated metrics can capture. We also intend to study the effect of varying model selection for both the MAV composition panel and the generation LLM, examining how provider diversity and model capability influence output quality.

Acknowledgements

We thank all volunteers who participated in the human evaluation study. Their willingness to assess review realism provided the perceptual validation that automated metrics alone cannot capture. AI-assisted tools supported code development, debugging, text editing, and large-scale dataset generation automation. All research direction, experimental design, and interpretive decisions remained solely with the authors, who take full responsibility for all content.

References

[1] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar. "SemEval-2014 Task 4: Aspect Based Sentiment Analysis". In: *SemEval*. 2014, pp. 27–35.

[2] J. McAuley, Q. Shi, C. Targett, and A. van den Hengel. "Image-based Recommendations on Styles and Substitutes". In: *SIGIR*. 2015, pp. 43–52.

[3] L. Long, R. Wang, R. Xiao, J. Zhao, X. Ding, G. Chen, and H. Wang. "On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey". In: *Findings of ACL*. 2024.

[4] M. Kochanek, I. Cichecki, O. Kaszyca, D. Szydło, M. Madej, and D. Jędrzejewski. "Improving Training Dataset Balance with ChatGPT Prompt Engineering". In: *Electronics* 13.12 (2024), p. 2255.

[5] H.-W. Kim and S.-B. Park. "Enhancing Imbalanced Sentiment Analysis: A GPT-3-Based Sentence-by-Sentence Generation Approach". In: *Applied Sciences* 14.2 (2024), p. 622.

[6] Q. Peng et al. "Review-LLM: Harnessing Large Language Models for Personalized Review Generation". In: *arXiv preprint arXiv:2407.07487* (2024).

[7] Y. Liu, X. Liu, X. Zhu, and W. Hu. "Multi-Aspect Controllable Text Generation with Disentangled Counterfactual Augmentation". In: *ACL*. 2024, pp. 9231–9253.

[8] C.-W. Yang, Z.-Q. Feng, Y.-J. Lin, C. W. Chen, K. da Wu, H. Xu, Y. Jui-Feng, and H.-Y. Kao. "MAPLE: Enhancing Review Generation with Multi-Aspect Prompt LEarning in Explainable Recommendation". In: *ACL*. 2025, pp. 31803–31821.

[9] L. Hellwig, M. Meilă, M. Fromm, and R. Klinger. "Exploring large language models for the generation of synthetic training samples for aspect-based sentiment analysis in low resource settings". In: *Expert Systems with Applications* 262 (2025), p. 125569.

[10] C. Chen, Y. Wen, J. Liu, S. Ding, et al. "Mitigating LLM Hallucinations Using a Multi-Agent Framework". In: *Information* 16.7 (2025), p. 517.

[11] N. Reimers and I. Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *EMNLP-IJCNLP*. 2019, pp. 3982–3992.

[12] E. Ma. *nlpaug: Data Augmentation for NLP*. https://github.com/makcedward/nlpaug. Python library for text augmentation. 2019.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. "BLEU: a Method for Automatic Evaluation of Machine Translation". In: *ACL*. 2002, pp. 311–318.

[14] C.-Y. Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. 2004, pp. 74–81.

[15] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. "BERTScore: Evaluating Text Generation with BERT". In: *ICLR*. 2020.

[16] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger. "MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance". In: *EMNLP-IJCNLP*. 2019.

[17] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. "A Diversity-Promoting Objective Function for Neural Conversation Models". In: *NAACL*. 2016.

[18] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu. "Texygen: A Benchmarking Platform for Text Generation Models". In: *SIGIR*. 2018.

[19] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos. "SemEval-2015 Task 12: Aspect Based Sentiment Analysis". In: *SemEval*. 2015, pp. 486–495.

[20] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. D. Clercq, et al. "SemEval-2016 Task 5: Aspect Based Sentiment Analysis". In: *SemEval*. 2016, pp. 19–30.

[21] S. U. S. Chebolu, F. Dernoncourt, N. Lipka, and T. Solorio. "OATS: A Challenge Dataset for Opinion Aspect Target Sentiment Joint Detection for Aspect-Based Sentiment Analysis". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, May 2024, pp. 12336–12347. URL: https://aclanthology.org/2024.lrec-main.1080.

[22] K. Thang. *cera-LADy: Latent Aspect Detection Framework for Synthetic Review Evaluation*. https://github.com/project-cera/cera-LADy. Extended from LADy [23] with unified ground truth evaluation. 2025.

[23] F. Hemmatizadeh, C. Wong, A. Yu, and H. Fani. "LADy: A Benchmark Toolkit for Latent Aspect Detection Enriched with Backtranslation Augmentation". In: *SIGIR*. 2024, pp. 1172–1178.

[24] X. Yan, J. Guo, Y. Lan, and X. Cheng. "A Biterm Topic Model for Short Texts". In: *Proceedings of the 22nd International Conference on World Wide Web (WWW)*. 2013, pp. 1445–1456. DOI: 10.1145/2488388.2488514.

[25] F. Bianchi, S. Terragni, and D. Hovy. "Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence". In: *ACL*. 2021, pp. 759–766.

[26] X. Li, L. Bing, W. Zhang, and W. Lam. "Exploiting BERT for End-to-End Aspect-Based Sentiment Analysis". In: *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT)*. 2019, pp. 34–41. DOI: 10.18653/v1/D19-5505.

[27] Qwen Team. "Qwen3 Technical Report". In: *arXiv preprint arXiv:2505.09388* (2025).

[28] OpenRouter. *OpenRouter: Unified API for LLMs*. https://openrouter.ai. Pricing accessed: February 2026. 2025.

Appendix A. Full Experimental Results

Table 2. Experimental configuration for the four cera-LADy detection architectures. A dash indicates the parameter is not applicable to that architecture. "Number of aspects" is the evaluation cutoff k used for P@k metrics, not the number of latent topics induced by the model.

Parameter	**BERT**	**BTM**	**CTM**	**RND**
Learning rate	2e-5	–	–	–
Epochs	3	–	20	–
Batch size	8	16	16	–
Iterations	–	1000	–	–
Max steps	500	–	–	–
Number of aspects	5	5	5	5
Cross-validation	5-fold	5-fold	5-fold	5-fold
Train/Test split	85/15%	85/15%	85/15%	85/15%

Table 3. Full MDQA intrinsic quality on the Laptop domain (mean ± SD across 5 runs, all dataset sizes). **Bold** marks the best synthetic method per metric and size.

		Lexical		Semantic		Diversity		
n	**Method**	BLEU ↑	ROUGE-L ↑	BERTScore ↑	MoverScore ↑	Distinct-1 ↑	Distinct-2 ↑	Self-BLEU ↓
	Real	*0.077*	*0.183*	*0.415*	*0.594*	*0.505*	*0.918*	*0.493*
100	CERA	0.058 ± .013	0.145 ± .008	0.427 ± .018	0.553 ± .012	**0.504** ± .010	**0.897** ± .008	**0.452** ± .013
	Heuristic	**0.073** ± .017	**0.185** ± .010	**0.483** ± .019	**0.594** ± .020	0.265 ± .012	0.594 ± .017	0.766 ± .014
	Real	*0.112*	*0.193*	*0.520*	*0.591*	*0.335*	*0.819*	*0.512*
500	CERA	0.054 ± .007	0.146 ± .005	0.412 ± .008	0.539 ± .016	**0.339** ± .020	**0.806** ± .018	**0.433** ± .026
	Heuristic	**0.095** ± .022	**0.153** ± .016	**0.531** ± .023	**0.572** ± .015	0.108 ± .002	0.334 ± .006	0.769 ± .008
	Real	*0.087*	*0.187*	*0.522*	*0.585*	*0.268*	*0.776*	*0.535*
1,000	CERA	0.048 ± .007	0.143 ± .004	0.403 ± .013	0.543 ± .014	**0.259** ± .015	**0.736** ± .019	**0.450** ± .011
	Heuristic	**0.055** ± .020	**0.180** ± .017	**0.519** ± .013	**0.589** ± .009	0.070 ± .001	0.254 ± .001	0.776 ± .009
	Real	*0.093*	*0.178*	*0.507*	*0.585*	*0.227*	*0.721*	*0.494*
1,500	CERA	0.049 ± .005	0.148 ± .005	0.407 ± .012	0.532 ± .007	**0.219** ± .010	**0.694** ± .015	**0.437** ± .025
	Heuristic	**0.047** ± .010	**0.178** ± .008	**0.554** ± .013	**0.583** ± .009	0.057 ± .001	0.219 ± .004	0.761 ± .013
	Real	*0.090*	*0.199*	*0.516*	*0.598*	*0.206*	*0.693*	*0.571*
2,000	CERA	0.048 ± .007	0.152 ± .007	0.402 ± .015	0.544 ± .005	**0.198** ± .015	**0.666** ± .027	**0.418** ± .026
	Heuristic	**0.043** ± .024	**0.194** ± .011	**0.569** ± .025	**0.599** ± .013	0.045 ± .001	0.183 ± .005	0.789 ± .013

Table 4. RGM ablation on the Laptop domain (mean across 5 runs). **Bold**: best condition per metric and size. Metrics: ^{L} Lexical, ^{S} Semantic, ^{D} Diversity.

	$n = 100$		$n = 500$		$n = 1{,}000$	
Metric	CERA	w/o RGM	CERA	w/o RGM	CERA	w/o RGM
BLEU^{L} ↑	0.058	**0.071**	0.054	**0.059**	0.048	**0.082**
ROUGE-L^{L} ↑	0.145	**0.172**	0.146	**0.166**	0.143	**0.175**
BERTScore^{S} ↑	0.427	**0.468**	0.412	**0.491**	0.403	**0.487**
MoverScore^{S} ↑	0.553	**0.576**	0.539	**0.578**	0.543	**0.569**
Distinct-1^{D} ↑	**0.504**	0.343	**0.339**	0.179	**0.259**	0.128
Distinct-2^{D} ↑	**0.897**	0.744	**0.806**	0.564	**0.736**	0.472
Self-BLEU^{D} ↓	**0.452**	0.627	**0.433**	0.629	**0.450**	0.619

Table 5. Human Evaluation Results ($N = 50$ evaluators). Task 1: evaluators identify the real review from an unlabelled triplet (Real, CERA, Heuristic); chance level is 33%. Task 2: evaluators select the more natural-sounding review from a pair with randomized left/right placement.

(a) Triplet Identification

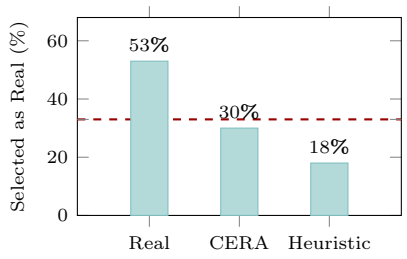
	Real	**CERA**	**Heur.**
Overall	53%	30%	18%
Laptop	48%	32%	19%
Restaurant	55%	24%	21%
Hotel	57%	35%	8%

Chance: 33%. Fleiss' $\kappa = 0.00$.

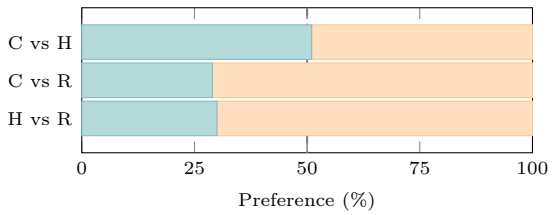
(b) Pairwise Naturalness

Pair	**Source A**	**Source B**
C vs H	CERA 51%	Heur. 49%
C vs R	CERA 29%	Real 71%
H vs R	Heur. 30%	Real 70%
Laptop: C vs H	44%	56%
H vs R	30%	70%
Rest.: C vs H	58%	42%
C vs R	24%	76%
Hotel: C vs R	34%	66%

Fleiss' $\kappa = -0.00$. C = CERA, H = Heur., R = Real.



(a) Triplet: *Which is real?* Dashed = 33% chance.



(b) Pairwise naturalness. Dashed = 50% (no pref.).

Figure 3. Human evaluation ($N = 50$ CS evaluators). (a) Triplet Turing test: CERA approaches 33% chance while Heuristic falls well below. (b) Pairwise preferences: stacked bars show the preference split. C = CERA, H = Heuristic, R = Real.