Uncertainty-Guided Large Language Model Data Augmentation for Cognitive Distortion Classification

Anonymous ACL submission

Abstract

Automated cognitive distortion classification is limited by data scarcity, since expert-labeled clinical text is costly, privacy-sensitive, and often anonymized for regulatory compliance. We investigate whether uncertainty-guided augmentation improves downstream classification. A baseline classifier flags its least-confident training instances, and an LLM generates synthetic variations of them under PII-preserving constraints. We cross two generators (LLaMA 3.3 70B, Qwen 3 235B) with two seed-selection strategies (entropy-based, random) on TherapistQA, against a non-augmented baseline and two traditional augmenters (EDA-SR, ContextualBERT). The best seed-selection strategy depends on the generator. In our experiments LLaMA favors random seeds and Qwen entropy-selected ones, contrary to the assumption that uncertainty-based selection transfers across consuming models. We treat this as directional, since only the cross-generator Random contrast survives Bonferroni correction (W=0.0, p=0.002). Offline LLM augmentation also beats both traditional augmenters by 7 to 9pp, and a 125M RoBERTa fine-tuned on the augmented data surpasses zero-shot LLaMA 70B and Qwen 235B by roughly 9pp at 560× to 1,880× fewer parameters, which favors local deployment for clinical text that cannot leave the institution. We release a reproducible, PII-preserving pipeline with entropy-guided seed selection.

1 Introduction

Cognitive Behavioral Therapy builds on the identification of *cognitive distortions*, biased patterns of thought that maintain depression and anxiety (Beck, 2011). Automating this identification step would help scale internet-delivered CBT, where clinicians are scarce (Chen et al., 2023). However, classifying *which* distortion is present remains hard. Even strong supervised models report weighted F1 below 0.30, owing to class imbalance, low annotator agreement, and semantic overlap between adjacent categories such as *Fortune-Telling* and *Emotional Reasoning* (Shreevastava and Foltz, 2021; Pico et al., 2025).

Large language models (LLMs) reason well zero-shot, but neither way of serving them fits this clinical setting under privacy constraints. Running a model of that size locally is costly, and the common alternative of a hosted API sends sensitive patient text to a third party, in violation of GDPR Article 9 (Hua et al., 2025; European Parliament and Council of the European Union, 2016). Compact encoders such as RoBERTa (Liu et al., 2019) run cheaply on-premise but struggle with data scarcity. A natural combination uses an LLM *offline*, on already-anonymized text, to augment the labeled set and deploys only the compact classifier at inference (Hsieh et al., 2023).

Uncertainty-guided seeding, i.e., selecting which existing labeled instances to use as prototypes for LLM generation, follows active learning, which targets the least-confident instances (Settles, 2009). The premise is that high-entropy instances mark an epistemic gap that the augmenter should fill. We test whether this uncertainty-guided seeding heuristic transfers to LLM-based augmentation, varying both the generator and the seed-selection strategy. We further investigate whether offline LLM augmentation improves a compact classifier overall, which classes it helps, and how the gains depend on these design choices.

We run a 2×2 factorial crossing two open-weight on-premise generators, the dense LLaMA 3.3 70B Instruct and the mixture-of-experts Qwen 3 235B-A22B Instruct, with two seed-selection strategies (entropy-based and random). Each cell forms an *LLM condition*. All conditions fine-tune RoBERTa-base on TherapistQA (Shreevastava and Foltz, 2021) under Relaxed Macro F1 (defined in Section 4.1), against a non-augmented

baseline and two traditional augmenters (EDA-SR, ContextualBERT). We compare the four LLM conditions with Wilcoxon signed-rank tests under Bonferroni correction.

We report three results. First, LLM augmentation outperforms both traditional augmenters and the non-augmented baseline (three of four LLM conditions above baseline, LLaMA Random highest at 0.3552 vs. 0.3331), and a fine-tuned 125M RoBERTa beats zero-shot LLaMA 70B and Qwen 235B by roughly 9 percentage points (pp) despite carrying 560× to 1,880× fewer parameters, which favors local deployment for privacy-sensitive clinical text that cannot leave the institution. Second, gains concentrate on topic-independent classes (Labeling, Overgeneralization, Personalization), evidence that augmentation works mainly because it erases dataset-specific lexical shortcuts (Geirhos et al., 2020). Third, the better seed-selection strategy depends on the generator, e.g., LLaMA prefers random and Qwen entropy, counter to the implicit assumption that seed-selection effects carry over across generators, with the cross-generator Random contrast the only pair to survive correction (W=0.0, p=0.002).

Contributions. (i) Evidence that the value of uncertainty-guided seed selection depends on the generator. LLaMA benefits from random seeds and Qwen from entropy-selected ones, which shows that the active-learning premise of transferable informativeness breaks once an LLM mediates generation. (ii) An empirical evaluation in which offline LLM augmentation beats two traditional augmenters and a non-augmented baseline, and a 125M classifier surpasses zero-shot LLaMA 70B and Qwen 235B by 9 to 10pp at 560× to 1,880× fewer parameters. (iii) Per-class evidence that gains track the topic-independence of class structure, which indicates augmentation helps mainly by erasing dataset-specific lexical shortcuts. (iv) A reproducible, privacy-aware pipeline with entropy-guided seed selection, PII-tag preservation, and multi-seed evaluation, with code released.[1]

2 Related Work

Cognitive distortion classification. Shreevastava and Foltz (2021) introduce the TherapistQA dataset and demonstrate that detecting the *presence* of a distortion is far easier than classifying its *type*. SVMs over Sentence-BERT embeddings reach weighted F1 below 0.30 in the multi-class setting. Ding et al. (2022) show on a clinical text-message dataset that a domain-specific MentalBERT outperforms several augmentation strategies, with mixup as the only augmentation that helps minority classes. Rasmy et al. (2025) apply EDA, BERT-based substitution, and back-translation on the C-Journal dataset (14 classes) and find that combinations of techniques yield up to a 5.9% F1 improvement. Pico et al. (2025) compare zero-shot LLMs and report non-trivial but limited performance without fine-tuning, while Chen et al. (2023) propose Diagnosis-of-Thought (DoT) prompting for multi-stage reasoning. Lim et al. (2024) likewise show that LLMs alone perform poorly on multi-class distortion classification and propose a multi-agent reasoning pipeline.

LLM-based data augmentation. Surface-level augmentation methods such as EDA (Wei and Zou, 2019) and contextual word substitution with pretrained encoders (e.g., Wu et al., 2019; Devlin et al., 2019) provide modest gains for small data, but their utility plateaus in the presence of pretrained classifiers (Li et al., 2022). LLM-based generation produces semantically diverse synthetic samples conditioned on a class label or seed instance (an existing labeled training example used as a generation prototype), and recent work shows it can improve downstream classification beyond what surface methods achieve (Wang et al., 2025). Generation diversity, however, is sensitive to model and decoding choices: Peeperkorn et al. (2024) show that temperature is only weakly correlated with novelty within a single LLM, and Guo et al. (2025) document substantial cross-model differences in lexical and semantic diversity even at matched decoding settings, attributable to differences in pretraining and alignment. Kirk et al. (2024) and Padmakumar and He (2024) further document how alignment procedures reduce output diversity in ways that can affect downstream usage of generated text.

Uncertainty estimation and selective augmentation. Softmax entropy is the simplest uncertainty signal but conflates aleatoric and epistemic components (Kendall and Gal, 2017). MC Dropout (Gal and Ghahramani, 2016) and Deep Ensembles (Lakshminarayanan et al., 2017) offer richer

[1] https://anonymous.4open.science/r/cd-uncertainty-aug-B5AA/

estimates at higher cost. Active-learning theory predicts that selecting high-uncertainty samples maximizes information gain per labeled instance (Settles, 2009; Li et al., 2025), but Lowell et al. (2019) report that uncertainty-selected pools transfer poorly across consuming models. Related transfer problems have been documented for uncertainty-based selection in LLM in-context learning settings (Margatina et al., 2023). To our knowledge, however, the present paper is the first to document an analogous problem when the consuming entity is a generative LLM producing training data for a downstream classifier, and to show that the direction of the effect depends on the generator.

3 Method

The framework consists of three phases (Figure 1). *Phase 1*: train a baseline RoBERTa-base classifier and compute per-sample Softmax Entropy on the training set. *Phase 2*: select seed instances, generate synthetic samples via LLM API or a traditional augmenter, validate preservation of personally identifiable information (PII) placeholder tags, and assemble an augmented training set per condition. *Phase 3*: fine-tune RoBERTa-base on each augmented set with weighted cross-entropy and evaluate on a held-out test set under Relaxed Macro F1 across multiple random seeds.

3.1 Task and dataset

We use the TherapistQA dataset (Shreevastava and Foltz, 2021), which consists of 2,530 anonymized patient utterances annotated with a *Dominant Distortion* and an optional *Secondary Distortion* that cover 11 classes (10 distortion types plus *No Distortion*). The dataset is severely imbalanced. *No Distortion* contains 933 samples while *All-or-Nothing Thinking* contains 100, a 9.33:1 ratio between the most and least frequent class. The data are split into 80% train and validation (2,024 samples) and 20% test (506 samples) using stratified sampling with a fixed random seed. The test set contains only original human-annotated data. Within train and val, an 87.5/12.5 split yields the training (1,771 samples) and validation partitions used for model selection. The source data are anonymized by the original authors and then passed through Microsoft Presidio (Mendels et al., 2018), where residual PII is replaced with typed placeholders such as `<PERSON>` and `<DATE_TIME>`.

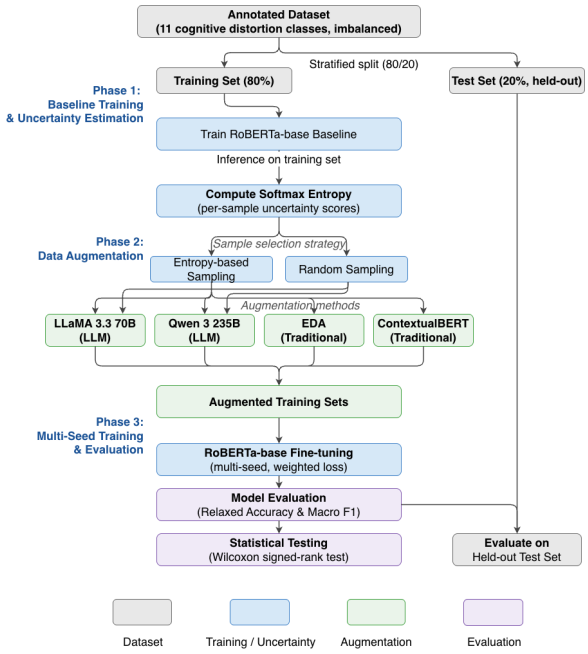

Figure 1: Three-phase experimental pipeline. Phase 1 trains a baseline RoBERTa and computes per-sample Softmax Entropy. Phase 2 selects seed instances (entropy-top-30% or uniform-random 30% per class), generates synthetic samples via an LLM (LLaMA 3.3 70B or Qwen 3 235B) or a traditional augmenter (EDA-SR, ContextualBERT), and validates PII-tag preservation. Phase 3 fine-tunes RoBERTa-base on each augmented training set and evaluates on the held-out test set across multiple random seeds.

Preserving these placeholders end-to-end ensures that the augmentation pipeline operates on the same anonymized representation that the model will see at inference time on novel clinical inputs.

3.2 Uncertainty estimation

We compare three uncertainty estimators on the trained baseline classifier: (i) Softmax Entropy from a single forward pass, (ii) MC Dropout (Gal and Ghahramani, 2016) with $T{=}10$ stochastic passes (RoBERTa default dropout $p{=}0.1$), and (iii) a Deep Ensemble (Lakshminarayanan et al., 2017) of $M{=}3$ baseline models trained with random seeds {42, 123, 2024}, where entropy is computed from the averaged predictive distribution. We retain Softmax Entropy as the operational signal because it requires only a single pass, achieves the highest Error-Detection AUC of the three (0.858 vs. 0.844 for MC Dropout and 0.848 for the ensemble), and produces uncertainty rankings that are highly correlated with the alternative estimators (Spearman $\rho > 0.89$).

3.3 LLM-based augmentation

We use two open-source instruction-tuned LLMs accessed through a single inference API: LLaMA 3.3 70B Instruct and Qwen 3 235B-A22B Instruct, both using sampling-based decoding at temperature 0.7 (full decoding parameters in Appendix A). The system prompt casts the model as a psychologist creating CBT training data, includes the definitions of all 11 distortion categories, and imposes three constraints: (1) the output must clearly exhibit the target distortion, (2) all PII placeholder tags from the seed instance must be preserved verbatim, and (3) the situation and phrasing must differ from the seed instance (no near-paraphrase). The user prompt supplies the seed text, the target label, and the list of tags to preserve. Each generation passes through two post-hoc filters. The refusal filter that drops outputs matching safety-refusal patterns, and a PII-tag check that regenerates, up to three retries, when a placeholder is missing or malformed. The full prompt is in Appendix D.

For each LLM, we evaluate two seed-selection strategies. In **Entropy-based** selection, we rank training samples within each class by Softmax Entropy and take the top 30% as the seed pool. In **Random** selection, we draw 30% of the training samples per class uniformly, using a fixed random seed for reproducibility. The augmentation target is 250 samples per class. Classes already at or above 250 (in this case, only *No Distortion*, with 653 training samples) are not augmented further. When the seed pool is smaller than the per-class generation target, the pool is cycled, with output variation arising from LLM stochasticity at temperature 0.7. After augmentation, the four LLM conditions yield 3,153 training samples each (1,382 synthetic).

3.4 Traditional baselines

EDA-SR. We use Easy Data Augmentation (Wei and Zou, 2019) restricted to Synonym Replacement, where non-stopword tokens are replaced with WordNet synonyms. Random insertion, swap, and deletion are excluded because they would shift or remove PII placeholder tags, consistent with Rasmy et al. (2025). **ContextualBERT.** A `bert-base-uncased` contextual word-substitution model (Devlin et al., 2019) replaces tokens under the same PII-stopword protection. Both methods reuse the entropy-selected seed pool (top 30% per class) and target 250 samples per class. Implementation details (`nlpaug` parameters, 13 PII stopwords) are in Appendix A.

3.5 Classifier and training

The classifier is RoBERTa-base (Liu et al., 2019) fine-tuned end-to-end with the standard `RobertaForSequenceClassification` head from HuggingFace Transformers (a two-layer MLP with $\tanh$ activation and dropout, applied to the <s> token representation and projecting to 11 classes). Training uses weighted cross-entropy with class weights $w_k = n/(K \cdot n_k)$ recomputed per condition, and AdamW. Model selection retains the checkpoint with the highest *Strict* Macro F1 on the validation set. Relaxed evaluation is reserved for the held-out test set (Section 4.2). Full training hyperparameters and hardware are in Appendix A.

4 Experimental Setup

4.1 Experimental conditions and outcome metrics

The experiments vary the augmentation method across seven settings: Baseline (no augmentation), EDA-SR, ContextualBERT, LLaMA Entropy, LLaMA Random, Qwen Entropy, and Qwen Random. Each setting comprises both the choice of generator and, for the LLM conditions, the seed-selection strategy.

Each condition is repeated across multiple random training seeds to assess result stability. The four LLM conditions are each trained with n=10 random seeds, while Baseline, EDA-SR and ContextualBERT use n=3 random seeds. The data-split seed is fixed at 42 across *all* conditions, so every model sees the same train/val/test partitions.

The primary outcome metric is **Relaxed Macro F1** on the held-out test set. A prediction is correct if it matches either the Dominant or the Secondary Distortion annotated for that sample to accommodate the multi-label nature of the annotation scheme and the documented semantic overlap between adjacent categories (Pico et al., 2025). We additionally report Relaxed Accuracy, Relaxed Weighted F1, and **Strict** Macro F1 in parallel, the latter for direct comparison with prior work that does not use the relaxed criterion.

4.2 Validation vs. test metric

Model selection during training uses *Strict* Macro F1 on the validation set, even though Relaxed Macro F1 is the primary test-time metric. The reason is that synthetic samples are generated from a prompt that conditions only on the dominant distortion. Hence, any secondary-distortion label inherited from the seed instance cannot be assumed valid for the generated text. Applying the relaxed criterion to the validation set would therefore mix verified dual labels (on original samples) with potentially invalid inherited labels (on synthetic samples), resulting in a metric whose optima would be unreliable. Strict F1, in contrast, applies the same dominant-only criterion to every sample. Relaxed evaluation is therefore reserved for the test set, where every sample carries a fully populated dual annotation.

4.3 Statistical protocol

We use the Wilcoxon signed-rank test on per-seed Relaxed Macro F1 scores. The test is restricted to the four pairwise contrasts, where each contrast is a direct comparison between two conditions, *among the four LLM-based conditions*: LLaMA Entropy vs. LLaMA Random, Qwen Entropy vs. Qwen Random, LLaMA Entropy vs. Qwen Entropy, and LLaMA Random vs. Qwen Random. With $n{=}10$ paired observations per contrast and a Bonferroni correction over four pairs, the corrected significance threshold is $\alpha{=}0.0125$. Comparisons involving Baseline, EDA-SR, or ContextualBERT are reported descriptively only (mean $\pm$ sample standard deviation, $\mathrm{ddof}{=}1$), because $n{=}3$ paired observations yield a minimum achievable two-sided p-value of 0.25 under the Wilcoxon signed-rank distribution, never significant at any conventional α.

5 Results

5.1 Main results

Table 1 reports the four outcome metrics across all seven conditions, including the three baselines and the four LLM-based. The best mean per column is bolded. Three patterns stand out. First, LLM-based augmentation cleanly outperforms traditional augmentation. All four LLM conditions exceed both EDA-SR and ContextualBERT on Relaxed Macro F1 by 0.07–0.10 absolute points, and both traditional methods drop *below* the non-augmented baseline by similar margins. Second, three of four LLM conditions exceed the Baseline mean (LLaMA Random by $+0.0221$, Qwen Entropy by $+0.0084$, LLaMA Entropy by $+0.0056$), while Qwen Random falls 0.0017 below the baseline mean, which is well within seed-to-seed variation. Third, the seed-selection strategy that wins is generator-dependent. LLaMA prefers Random ($+0.0165$ over Entropy), while Qwen prefers Entropy ($+0.0101$ over Random).

Comparing the results under Strict and Relaxed Macro F1 provides additional insight into how LLM augmentation shifts predictions. Under *Strict* Macro F1, only LLaMA Random clearly exceeds the baseline ($+0.009$). The other three LLM conditions sit at or slightly below baseline, while traditional methods drop by 0.05–0.06. The gap Δ = Relaxed $-$ Strict is systematically larger for LLM conditions ($+0.052$ to $+0.057$) than for the baseline ($+0.043$) or traditional methods ($+0.027$ to $+0.033$). A larger Δ indicates the classifier more often misses the dominant label while still landing on the annotated secondary one. LLM augmentation does not primarily sharpen dominant-label decisions. It makes the classifier more likely to predict a *plausible alternative distortion* when it does not identify the dominant one. The pattern where LLaMA benefits most from Random selection and Qwen from Entropy selection holds under both metrics.

5.2 Statistical comparison among LLM conditions

Of the four contrasts (Table 2), only LLaMA Random vs. Qwen Random reaches significance after Bonferroni correction ($W{=}0.0$, $p{=}0.002$). The $W{=}0.0$ statistic indicates that LLaMA Random outperformed Qwen Random in all 10 paired seeds, which is the strongest possible result for $n{=}10$. The within-LLM seed-selection strategy comparisons both show directional trends consistent with the LLM and strategy interaction (LLaMA: random $>$ entropy, $p{=}0.064$; Qwen: entropy $>$ random, $p{=}0.106$), but neither reaches the corrected threshold. This is unsurprising with $n{=}10$ and small effect sizes. The cross-LLM Entropy contrast is the genuine null ($p{=}0.375$). That is, when both generators are fed entropy-selected seed instances, they produce indistinguishable outputs. Figure 2 visualizes this generator and seed-selection interaction. The line crossing makes the qualitative pattern immediately visible.

Condition	n	Rlx. Acc.	Rlx. Macro F1	Rlx. Weighted F1	Strict Macro F1
Baseline	3	0.4196 ± 0.0225	0.3331 ± 0.0060	0.4301 ± 0.0137	0.2897 ± 0.0091
EDA-SR	3	0.4177 ± 0.0148	0.2608 ± 0.0142	0.3836 ± 0.0101	0.2276 ± 0.0095
ContextualBERT	3	0.4058 ± 0.0075	0.2655 ± 0.0195	0.3835 ± 0.0079	0.2387 ± 0.0132
LLaMA Entropy	10	0.4370 ± 0.0189	0.3387 ± 0.0193	0.4372 ± 0.0177	0.2819 ± 0.0171
LLaMA Random	10	**0.4461 ± 0.0154**	**0.3552 ± 0.0164**	**0.4443 ± 0.0145**	**0.2987 ± 0.0149**
Qwen Entropy	10	0.4423 ± 0.0128	0.3415 ± 0.0167	0.4396 ± 0.0127	0.2856 ± 0.0136
Qwen Random	10	0.4362 ± 0.0142	0.3314 ± 0.0166	0.4305 ± 0.0123	0.2791 ± 0.0139

Table 1: Main results on the held-out test set: mean ± sample standard deviation across n random seeds. *Rlx.* denotes Relaxed evaluation. The bolded row (LLaMA Random) is the per-column best across all four metrics.

Pair	W	p	Sig.
LLaMA Ent. vs. LLaMA Rand.	9.0	0.0645	No
Qwen Ent. vs. Qwen Rand.	11.0	0.1055	No
LLaMA Ent. vs. Qwen Ent.	18.0	0.3750	No
LLaMA Rand. vs. Qwen Rand.	0.0	0.0020	**Yes**

Table 2: Wilcoxon signed-rank tests among the four LLM-based conditions (n=10, two-sided, Bonferroni-corrected α=0.0125). Only the cross-LLM Random pair is significant.

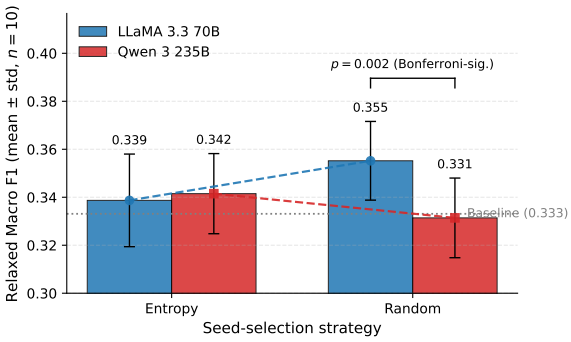

Figure 2: Generator and seed-selection interaction on Relaxed Macro F1. The bars are the mean ± sample standard deviation across $n = 10$ random seeds. The line crossing visualizes the qualitative interaction. LLaMA prefers Random selection (+0.0165), Qwen prefers Entropy selection (+0.0101). Cross-LLM Random is the only Bonferroni-significant Wilcoxon contrast in our four-pair test family (p=0.002).

5.3 Per-class analysis

Table 3 reports relaxed per-class F1 as means across random seeds. The Macro Average row reproduces the Relaxed Macro F1 column of Table 1. Three patterns are visible. (a) Both traditional methods degrade most classes. The largest drops appear on *Emotional Reasoning* (−0.192 EDA-SR), *Fortune-Telling* (−0.206 ContextualBERT), and *Should statements* (−0.145 EDA-SR). (b) LLM-based augmentation produces large improvements on *Labeling* (+0.147 for Qwen Random over baseline), *Overgeneralization* (+0.084 Qwen Entropy), *Personalization* (+0.076 Qwen Random), and *Should statements* (+0.051 LLaMA Random). (c) *All-or-Nothing Thinking*, the smallest class with 100 samples in the original data, shows the opposite pattern as the baseline achieves the highest F1 (0.304) and *every* augmentation method degrades it. The largest class *No Distortion* remains stable across all conditions (0.633–0.643).

5.4 RoBERTa vs. zero-shot LLMs

The non-augmented RoBERTa baseline achieves a Relaxed Macro F1 of 0.3331. For comparison, zero-shot LLaMA 3.3 70B and Qwen 3 235B (each averaged over three runs at temperature 0.7 with a fixed prompt) reach only 0.2391 and 0.2463, respectively. Fine-tuned RoBERTa therefore outperforms them by +9.4 pps (+39.3% relative) and +8.7 pps (+35.2% relative). The best augmented condition extends this gap further. Despite having ∼560× to ∼1,880× fewer parameters, the fine-tuned compact classifier substantially outperforms both LLMs in zero-shot use, which supports the deployment of locally-deployable classifiers as a practical alternative to large general-purpose LLMs at inference time.

6 Analysis and Discussion

6.1 The generator and seed-selection strategy interaction

The most informative finding is that the optimal seed-selection strategy is generator-dependent. We discuss three non-exclusive hypotheses. The present data cannot definitively distinguish between them.

Label fidelity under ambiguous seed instances. Entropy-selected seed instances are, by construction, those on which the baseline is least confident: semantically ambiguous, plausibly carrying multiple distortions. A generator must implicitly

Class	Base	EDA	BERT	LL-E	LL-R	Qw-E	Qw-R
All-or-nothing	**.304**	.172	.178	.191	.180	.214	.183
Emot. Reasoning	**.257**	.065	.086	.203	.253	.158	.157
Fortune-Telling	.482	.297	.276	.439	**.491**	.455	.429
Labeling	.311	.221	.289	.415	.444	.433	**.458**
Magnification	.254	**.314**	.297	.298	.303	.307	.279
Mental Filter	.164	.204	.130	.191	**.209**	.173	.179
Mind Reading	.415	.359	.347	.409	.404	**.426**	.403
No Distortion	**.643**	.636	.634	.640	.634	.634	.633
Overgeneraliz.	.173	.132	.146	.235	.216	**.257**	.210
Personalization	.322	.274	.227	.334	.381	.361	**.398**
Should stmts	.341	.196	.313	.369	**.392**	.343	.320
Macro Avg	.333	.261	.266	.339	**.355**	.342	.331

Table 3: Relaxed per-class F1 (mean across random seeds). LL-E/LL-R = LLaMA Entropy/Random; Qw-E/Qw-R = Qwen Entropy/Random. Per-class best is bolded.

resolve which distortion to preserve. Xia et al. (2024) show that format-following and content-following capabilities are decoupled in LLMs. Xie et al. (2024) show that LLMs vary substantially in how they balance contextual evidence against parametric memory under conflict. By analogy, LLaMA and Qwen may differ in how faithfully they preserve the labeled distortion when the seed instance is ambiguous, despite being comparable on standard instruction-following benchmarks (e.g., IFEval). One plausible conjecture is that Qwen stays closer to the labeled category on ambiguous inputs and therefore benefits from harder, more informative entropy-selected seed instances, while LLaMA drifts toward whichever distortion the surface text most closely resembles, making cleaner, randomly selected seed instances more useful. Direct measurement of label fidelity on the generated samples would test this.

Output diversity. At fixed temperature, generators produce different intrinsic levels of variation (Peeperkorn et al., 2024; Guo et al., 2025; Kirk et al., 2024). If LLaMA's outputs are more homogeneous than Qwen's, the informativeness gap between high-entropy and randomly-selected seed instances may be flattened in the LLaMA case because the generator produces similar text regardless of seed-instance difficulty, which makes entropy targeting less effective. Quantifying lexical and semantic diversity in the generated samples themselves would test this.

Category-boundary alignment. The entropy signal is task-specific. It reflects the decision boundary that the RoBERTa baseline learned on TherapistQA. Generators whose pretraining corpora align with TherapistQA's annotation conventions may find the entropy-selected seed instances informative. Generators whose pretraining encodes different category boundaries (i.e., different pretrained intuitions about where one distortion type ends and another begins) may find the same seed instances confusing (Lowell et al., 2019; Zhao et al., 2024). Testing this would require measuring agreement between each LLM's zero-shot classification and the gold labels on a shared seed pool.

The broader implication is robust regardless of which hypothesis turns out correct: **seed-instance selection and generator choice cannot be tuned independently in LLM-based augmentation**. The classical active-learning assumption that informative samples yield better downstream performance holds only when the entity processing them can faithfully exploit their informativeness. When an LLM mediates between a seed instance and the training set, that assumption is no longer automatic.

6.2 Why augmentation gains concentrate on attributional classes

The three classes that benefit most from LLM augmentation (*Labeling*, *Overgeneralization*, *Personalization*) share a structural property. Each is a descriptive attribution presented as a factual claim: *Labeling* attributes a fixed characteristic to a person ("I am a failure"); *Overgeneralization* attributes a universal property to a category ("all medical staff are rude"); *Personalization* attributes causation to oneself ("it rained because I planned a picnic"). None of these requires a specific tense, modal verb, emotional clause, or reference to another person's mental state. Their definition is es-

sentially *topic-independent*, unlike classes such as *All-or-Nothing Thinking* that rely on specific surface markers (e.g., "always" or "never").

This formal freedom plausibly matters via shortcut learning (Geirhos et al., 2020). With limited training data, neural classifiers latch onto surface features that happen to co-occur with class labels. For a topic-independent class, the few available training examples inevitably share incidental topics and phrasing alongside the underlying attributional pattern, and the model has no signal to tell which features are genuinely discriminative. LLM augmentation exhibits the same attributional pattern across many topics (work, family, health, friendships). In the augmented dataset, topic features become uncorrelated with the label, and the only feature that remains reliably discriminative is the attribution itself. The mechanism predicts exactly the observed contrast. Classes whose definition is topic-independent benefit most, while *All-or-Nothing Thinking* is *harmed* by augmentation precisely because LLM-generated variation weakens its lexical signal. This pattern cannot be reduced to a class-size or augmentation-volume effect. All-or-Nothing Thinking receives the largest number of synthetic samples (180; Appendix B) yet is the most degraded class, while gains concentrate on classes that receive fewer synthetic samples but share the structural property. Testing this account directly would require measuring topic diversity in the augmented data and correlating it with class-wise improvement. We leave that to future work.

6.3 Why traditional baselines hurt

EDA-SR and ContextualBERT both perform word-level substitution while leaving sentence structure, situation, and the underlying cognitive pattern unchanged. Cognitive distortions are defined by their rhetorical and semantic structure rather than by specific lexical choices, so lexical substitution generates samples that are *near-paraphrases of the seed instance* rather than *new instances of the class*. This produces three concrete failure modes consistent with the per-class drops we observe: (a) the model gains no new examples of how a distortion can manifest in different surface forms, so the decision boundary does not move where it most needs to. (b) PII placeholder tags are preserved as-is, but surrounding clinical-register vocabulary is sometimes substituted by lexically-valid-but-register-inappropriate WordNet synonyms or BERT predictions, increasing distributional drift between training and test. (c) For classes with strong lexical markers (e.g., All-or-Nothing Thinking), random word replacement sometimes erases the very signal the model was relying on, with no compensating gain. The net effect is that traditional methods add noise without adding diversity along the dimension that matters.

6.4 Practical implications

Two implications follow. First, for clinical NLP under privacy constraints, *compact classifiers fine-tuned on LLM-augmented data* are a credible deployment architecture pending performance improvements. A 125M-parameter RoBERTa beats 70B–235B zero-shot LLMs by 9 pps absolute (35–39% relative) on this task, runs locally on commodity hardware, and never requires sending patient text to an external API at inference time. The LLM is used only offline, on PII-redacted seed instances, during a controlled preparation phase. Second, for practitioners building such pipelines, both the generator and the seed-selection strategy should be treated as hyperparameters, validated empirically on held-out data rather than chosen by convention. A small-scale pilot comparing two or three (generator, strategy) combinations is cheap and would have flipped the design in our setting.

7 Conclusion

We investigated uncertainty-guided LLM augmentation for cognitive distortion classification. We built an end-to-end pipeline including initial uncertainty estimation per class, a generator, and a classification component. The winning strategy is generator-dependent. LLaMA prefers random seeds, Qwen entropy-selected ones. Generator and seed-selection choice thus behave as one joint hyperparameter. Our pipeline, with a 125M RoBERTa fine-tuned on LLM-augmented data, outperforms zero-shot LLaMA 70B and Qwen 235B by roughly 9 pp despite carrying 560× to 1,880× fewer parameters, and is a credible deployment path for privacy-sensitive clinical text. Whether the interaction holds across more than two models and a single English dataset remains open, which we leave for future work.

Limitations

Single dataset, single language. All results are on TherapistQA in English. The dataset's annotation conventions, taxonomy, and text style may favor certain augmentation strategies; replication on the C-Journal dataset (Rasmy et al., 2025) or the clinical-text-message corpus of Ding et al. (2022), and on non-English data, would strengthen external validity.

Qualitative class-structure characterization. The account in Section 6.2 characterizes class structure (topic-independent attribution patterns vs. narrow lexical markers) qualitatively, from the operational definitions of the cognitive distortion categories. A quantitative characterization (for example, per-class lexical concentration measured by pointwise mutual information between tokens and class labels) would test this account more directly, which we leave to future work.

Asymmetric random-seed budgets. The four LLM conditions use n=10 random seeds while Baseline, EDA-SR, and ContextualBERT use n=3. This asymmetry is intentional (the LLM conditions are the inferential focus), but it means we cannot apply Wilcoxon signed-rank to LLM-vs-traditional contrasts (the n=3 minimum two-sided p is 0.25). Even within the LLM conditions, n=10 provides limited power: the within-LLM strategy contrasts (p=0.065 and p=0.106) might cross threshold with a larger budget. Replicating all conditions at the same n would close this gap.

Two LLMs only. We cannot determine whether the generator and seed-selection strategy interaction generalizes to other model families, sizes, or training procedures. The three hypotheses in Section 6.1 make different predictions and would require additional generators to test.

Validation/test metric asymmetry. Synthetic samples' secondary-distortion labels cannot be assumed valid because the LLM generation prompt conditions only on the dominant distortion, so model selection uses Strict Macro F1 on the validation set while the primary test-time metric is Relaxed Macro F1. We argue this is the only consistent option (Section 4.2), but it does mean that hyperparameter selection is not directly aligned with the primary metric.

Entropy computation with dropout active. Softmax Entropy was computed with RoBERTa in training mode (dropout p=0.1 active), not in pure-eval mode. Given the low dropout rate the relative ranking of samples is expected to be stable, but the scores are not strictly deterministic.

No quality audit of synthetic samples. We did not evaluate the LLM-generated samples for label fidelity, factual coherence, or demographic diversity beyond the PII tag check and refusal filter. A quality audit would directly test the label-fidelity hypothesis in Section 6.1.

Possible pretraining contamination. TherapistQA is publicly available and may be in either LLM's pretraining corpus. The augmentation prompt instructs the model to generate *new* scenarios, the dataset is small relative to the pretraining corpus, and the observed gains are modest rather than anomalous, but membership-inference testing was not performed.

Computational cost. LLM augmentation incurs API costs that scale with the number of synthetic samples; for a single dataset of this size the cost is modest, but iterative or multi-round designs would compound it.

Ethics Statement

Data and PII. TherapistQA was anonymized by its original authors (Shreevastava and Foltz, 2021) and released publicly. We additionally pass all text through Microsoft Presidio so that residual PII is replaced with typed placeholders, and we validate placeholder preservation in every generated sample. Real patient data does not reach the LLM at inference time; LLM augmentation operates on already-anonymized text during an offline preparation phase, and the deployed classifier is small enough to run locally. This pipeline aligns with the data-minimisation principle of GDPR Article 5 (European Parliament and Council of the European Union, 2016).

Synthetic clinical-adjacent text. The LLM-generated samples are training data only; they are not clinical content and are not released or shared as patient-facing material. A safety-refusal filter removes outputs that match patterns associated with crisis content. Should the synthetic data be released in future work, downstream users should be made aware of its synthetic origin.

Deployment risks. The reported performance (∼0.355 Relaxed Macro F1 for the best condition)

is well below what would justify autonomous clinical use. Any deployment should treat the classifier as a clinician-support component, with human oversight, and should evaluate performance across demographic subgroups before fielding. Cognitive distortion detection has dual-use risks (e.g., surveillance applications); we advocate clinician-supervised use and discourage autonomous deployment.

Bias. TherapistQA reflects the demographic and topical distribution of its sources; LLMs carry their own pretraining biases. We did not perform a systematic bias audit of the augmented data; one is recommended before any fielded use.

References

Judith S. Beck. 2011. *Cognitive Behavior Therapy: Basics and Beyond*, 2nd edition. The Guilford Press, New York, NY.

Zhiyu Chen, Yujie Lu, and William Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiruo Ding, Kevin Lybarger, Justin Tauscher, and Trevor Cohen. 2022. Improving classification of infrequent cognitive distortions: Domain-specific model vs. data augmentation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 68–75, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

European Parliament and Council of the European Union. 2016. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Official Journal of the European Union, L 119, 4.5.2016, pp. 1–88. CELEX:32016R0679. Accessed: 2026-04-17.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, NY, USA. PMLR.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou, Lauren V. Moran, Sophia Ananiadou, David A. Clifton, Andrew Beam, and John Torous. 2025. Large language models in mental health care: A scoping review. *Current Treatment Options in Psychiatry*, 12(1):27. Article number 27.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, volume 30, pages 5580–5590, Red Hook, NY, USA. Curran Associates, Inc.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of RLHF on LLM generalisation and diversity. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, Vienna, Austria.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, volume 30, pages 6405–6416, Red Hook, NY, USA. Curran Associates, Inc.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. 2025. A survey on deep active learning: Recent advances and

new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):5879–5899.

Sehee Lim, Yejin Kim, Chi-Hyun Choi, Jy-yong Sohn, and Byung-Hoon Kim. 2024. ERD: A framework for improving LLM reasoning for cognitive distortion classification. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 292–300, Mexico City, Mexico. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. 2019. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 21–30, Hong Kong, China. Association for Computational Linguistics.

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active learning principles for in-context learning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5011–5034, Singapore. Association for Computational Linguistics.

Omri Mendels, Coby Peled, Nava Vaisman Levy, Sharon Hart, Tomer Rosenthal, Limor Lahiani, and 1 others. 2018. Microsoft Presidio: Context aware, pluggable and customizable pii anonymization service for text and images. Accessed: 2025-10-25.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, Vienna, Austria.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? In *Proceedings of the 15th International Conference on Computational Creativity (ICCC'24)*, pages 226–235, Jönköping, Sweden. Association for Computational Creativity.

Aaron Pico, Joaquín Taverner, Emilio Vivancos, and Ana García-Fornés. 2025. Comparative analysis of the efficacy in the classification of cognitive distortions using LLMs. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) – Volume 1*, pages 957–965, Porto, Portugal. SCITEPRESS – Science and Technology Publications.

Mohamad Rasmy, Caroline Sabty, Nourhan Sakr, and Alia El Bolock. 2025. Enhanced cognitive distortions detection and classification through data augmentation techniques. In *PRICAI 2024: Trends in Artificial Intelligence – 21st Pacific Rim International Conference on Artificial Intelligence, Kyoto, Japan, November 18–24, 2024, Proceedings, Part I*, volume 15281 of *Lecture Notes in Computer Science*, pages 134–145, Singapore. Springer Nature Singapore.

Burr Settles. 2009. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, Department of Computer Sciences, Madison, WI, USA.

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158, Online. Association for Computational Linguistics.

Zaitian Wang, Jinghan Zhang, Xinhao Zhang, Kunpeng Liu, Pengfei Wang, and Yuanchun Zhou. 2025. Diversity-oriented data augmentation with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22265–22283, Vienna, Austria. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional BERT contextual augmentation. In *Proceedings of the 3rd International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 84–95, Las Vegas, NV, USA. IEEE.

Congying Xia, Chen Xing, Jiangshu Du, Xinyi Yang, Yihao Feng, Ran Xu, Wenpeng Yin, and Caiming Xiong. 2024. FOFO: A benchmark to evaluate LLMs' format-following capability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 680–699, Bangkok, Thailand. Association for Computational Linguistics.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, Vienna, Austria.

Siyan Zhao, Tung Nguyen, and Aditya Grover. 2024. Probing the decision boundaries of in-context learning in large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, volume 37, pages 130408–130432, Red Hook, NY, USA. Curran Associates, Inc.

A Hyperparameters

Table 4 lists the training hyperparameters held constant across all conditions.

Hyperparameter	Value
Base model	RoBERTa-base (∼125M params)
Max seq. length	512 tokens
Output classes	11
Loss	Weighted CE, $w_k = n/(K \cdot n_k)$
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0.01
Batch size	16
Epochs	10
LR schedule	Linear warmup 100 + linear decay
Model selection	Strict Macro F1 on validation
Hardware	A100-40GB (cloud training)
Train time/model	∼12–13 minutes

Table 4: Training hyperparameters (constant across conditions; class weights recomputed per condition).

Augmentation hyperparameters. We fix the random seed to 42 for all data splits and seed-selection processes.

- **LLM generation:** T=0.7, `max_tokens`=256, `top-p`=1.0. We allow up to 3 retries for PII validation failures.
- **EDA-SR:** We use `nlpaug.SynonymAug` (WordNet) with α=0.1 and `aug_min`=1, `aug_max`=10. All PII tags are registered as stopwords.
- **ContextualBERT:** We use `nlpaug.ContextualWordEmbsAug` (`bert-base-uncased`, `action="substitute"`) with α=0.1 and `batch_size`=32. All PII tags are protected as stopwords.
- **Selection:** The augmentation target is 250 samples per class; classes already at or above 250 (only *No Distortion*) are not augmented. The seed pool comprises the top 30% by entropy (Entropy) or 30% selected uniformly (Random).
- **Data splitting:** All splits (train/val/test) use stratified sampling.

B Class Distribution After Augmentation

Table 5 reports per-class sample counts in the training partition, both before augmentation (*Orig.*) and after (*Aug.*), together with the number of synthetic samples added per class (*Synth.*). Augmentation targets 250 samples per class. Classes already at or above this threshold (only *No Distortion*, with 653 training samples) receive no augmentation. The most-to-least-frequent class imbalance ratio drops from 9.33:1 in the original training partition to 2.61:1 after augmentation.

Class	Orig.	Aug.	Synth.
No Distortion	653	653	0
Mind Reading	167	250	83
Overgeneralization	167	250	83
Magnification	136	250	114
Labeling	115	250	135
Personalization	107	250	143
Fortune-Telling	101	250	149
Emotional Reasoning	94	250	156
Mental Filter	86	250	164
Should Statements	75	250	175
All-or-Nothing Thinking	70	250	180
Total	**1,771**	**3,153**	**1,382**

Table 5: Class counts in the training partition before and after augmentation. The 87.5/12.5 train/validation split is applied first; counts above are training-partition only.

C Full Statistical Tests

The four Wilcoxon contrasts in Table 2 are the entire test family in this study; the Bonferroni correction divides α=0.05 by 4 to obtain α_{corr}=0.0125. We did not run pairwise Wilcoxon tests against Baseline, EDA-SR, or ContextualBERT: with n=3 paired observations, the smallest two-sided p-value attainable under the Wilcoxon signed-rank null distribution is 0.25. This is an arithmetic property of the test, not a statement about the underlying effects. Reporting such tests would have offered no inferential payoff while inviting the misreading that "Baseline vs. LLM Random has p=0.25, hence not significantly different," a conclusion that cannot be drawn from n=3 in either direction.

D Prompt Templates and Output Examples

System prompt (abbreviated, full version in supplementary material).

```
You are an expert psychologist assisting
in creating a dataset for Cognitive
Behavioral Therapy (CBT). YOUR TASK:
Rewrite the user's text to create a
new, unique scenario that exhibits the
```

```
SAME cognitive distortion. GUIDELINES:
(1) Maintain the Distortion: keep the
labeled distortion and clearly express
it. [11 distortion definitions inserted
here.] (2) Preserve PII Tags: every
placeholder tag from the input must
appear verbatim in the output. (3)
Vary the Context: change the situation,
the actors, and the sentence structure;
do NOT paraphrase. (4) Output Format:
return ONLY the new text, with no
explanation or preamble.
```

User prompt template.

```
Original Text: "{seed_text}"\n
Distortion Label: {label}\n Tags
to Keep: {tags}\n Generate a new
variation:
```

Example (LLaMA, target = Labeling).

- *Seed instance:* "I failed the interview at <DATE_TIME>, I'm just a useless person."
- *Generated:* "After my project presentation last <DATE_TIME> got cut short, I keep telling myself I'm a complete failure who can't do anything right."

Validation pipeline. Each generation passes (i) a refusal-pattern filter that drops outputs containing safety-refusal phrases, and (ii) a PII tag check that re-generates (up to 3 retries) when any placeholder from the seed instance is missing or malformed in the output.