CASCADE: Cognitive Assessment via Sentence Context and Distortion Extraction

Archit Narumanchi **Azamat Rashidov** **Jong C. Park***
School of Computing
Korea Advanced Institute of Science and Technology (KAIST)
{archit.narumanchi,azamat03,jongpark}@kaist.ac.kr

Abstract

Cognitive distortions are unhelpful thinking patterns central to Cognitive Behavioral Therapy (CBT) and closely linked to mental health outcomes. Despite their context-dependent nature, studies at the intersection of Natural Language Processing and Cognitive Behavioral Therapy are centered largely around sentence-level utterances, ignoring the document-level complexity of real clinical text. To this end, we propose a two-stage rank-then-classify cascade that decomposes document-level multi-label cognitive distortion detection into (1) a sentence-ranking stage that identifies the most likely distorted sentences, and (2) a multi-label classification stage that assigns distortion subtypes to the selected sentences by incorporating their surrounding interleaved context. While the pipeline stages are trained on separate out-of-domain datasets, our model matches state-of-the-art performance on the CBT-CD benchmark of 166 expert-annotated clinical vignettes. Operating with only 368M parameters, our system achieves a 51.8% weighted F1 (9-class), surpassing massive large language model baselines such as Mistral-7B-v0.3 (46.0%) and performing on par with state-of-the-art models such as Gemma 4 26B A4B (50.8%).

1 Introduction

Cognitive distortions (CDs) are patterns of irrational or exaggerated thinking that cause inaccurate and negatively-biased perceptions of reality (Beck, 1963). Identifying these distortions in patient narratives is a core component of Cognitive Behavioral Therapy (CBT), where clinicians use distortion identification and re-framing to guide therapeutic interventions (Burns, 1999).

Recent empirical work has established that the expression of cognitive distortions in language is directly associated with mental health outcomes. Varadarajan et al. (2025) found significant positive correlations between language-detected distortion rates and measures of depression and anxiety severity. In general, cognitive distortions are broadly understood to play a central role in the onset and worsening of psychopathological disorders, making their timely identification a key mechanism of effective treatment (Morrison et al., 2015; Kaplan et al., 2017; Kim et al., 2025).

To this end, automating the detection and classification of cognitive distortions in naturalistic contexts holds substantial promise for scaling mental health support. However, the computational study of CDs has made notable progress almost entirely at the sentence level (Sage et al., 2025). This approach side-steps the reality that therapy sessions and clinical vignettes are inherently long-form narratives in which a patient's distorted thoughts are rarely expressed in isolation, and are instead interwoven with situational descriptions, emotional language, and undistorted reflections. Moreover, many distortions are context-dependent or only weakly signaled by surface-level lexical cues (Sage et al., 2025), meaning that models trained on isolated utterances are forced to rely on shallow patterns and may fail on more ambiguous cases. To make it further complicated, patients frequently exhibit multiple co-occurring distortion types within a single narrative, necessitating nuanced multi-label prediction (Sage et al., 2025).

We observe that the challenge of document-level distortion detection decomposes naturally into two sub-problems: retrieval of candidate distorted sentences from a longer narrative, and classification of each candidate into fine-grained distortion subtypes. Crucially, existing resources already address each sub-problem in isolation. Particularly, the THERAPISTQA (Shreevastava and Foltz, 2021) corpus provides span-level annotations indicating where distortions occur, while PATTERNREFRAME (Maddela et al., 2023) sup-

*Corresponding author.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3376–3391
November 6-10, 2026 ©2026 Association for Computational Linguistics

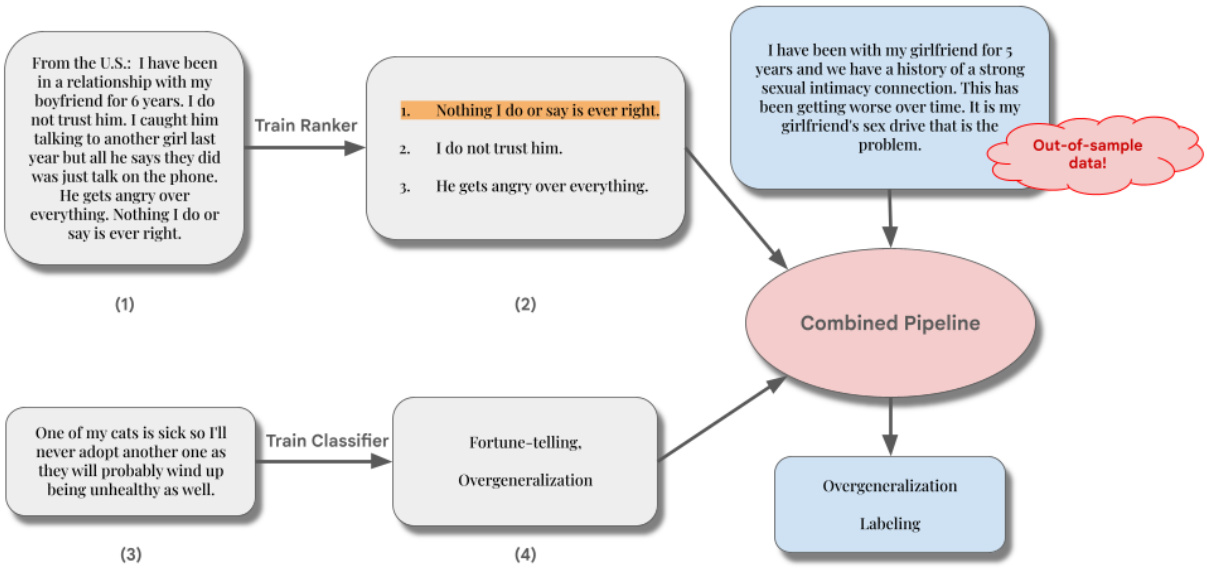


Figure 1: Full rank-then-classify pipeline architecture. The ranker (trained on the THERAPISTQA dataset) selects top-3 sentences (2) from long-form documents (1); the classifier (trained on the PATTERNREFRAME dataset) produces multi-label predictions (4) from sentence level inputs (3); the combined pipeline produces multi-label predictions from document-level inputs (evaluated on the CBT-CD dataset)

plies reliable subtype labels for individual sentences. However, no single resource provides both at document scale. This complementarity motivates a cascade architecture in which a ranker trained on the former is composed with a classifier trained on the latter, yielding a system capable of document-level inference without requiring document-level subtype annotations for training.

In this work, we address the challenge of multi-label distortion subtype classification in long-form clinical text. We introduce CASCADE (**C**ognitive **A**ssessment via **S**entence **C**ontext **a**nd **D**istortion **E**xtraction), a lightweight two-stage pipeline that is trained entirely on existing resources without access to any expert-annotated document-level subtype labels. The CBT-CD dataset (Zhang et al., 2025) is reserved exclusively for evaluation. Despite this constraint, CASCADE matches state-of-the-art performance on CBT-CD, performing on par with or surpassing substantially larger language model baselines at a fraction of the computational cost. Our contributions are as follows:

1. A two-stage rank-then-classify cascade architecture that decomposes document-level multi-label cognitive distortion detection into sentence retrieval and subtype classification, each trainable on different existing datasets and composable at inference time.

2. A novel situation-thought interleaving strategy that reconstructs clinical narrative flow via semantic similarity, providing the classifier with the contextual signal needed to disambiguate distortion subtypes.

3. Empirical evidence that task-specific fine-tuned transformers totaling 368M parameters can match Large Language Models (LLMs) with 70× more total parameters on document-level CD classification, even when such LLMs receive few-shot in-domain examples.

4. A principled taxonomy harmonization strategy that bridges three datasets with inconsistent annotation schemas, grounded in Beck's arbitrary inference framework (Beck et al., 1979).

2 Related Work

The computational study of cognitive distortions spans a range of methodological approaches, from rule-based systems to Large Language Models (LLMs). We organize prior work along axes that are central to our contribution: the granularity of

the input text and the evaluation methodology. For a comprehensive survey, we refer the reader to Sage et al. (2025).

Sentence-Level Approaches. The majority of computational CD research operates at the sentence or single-utterance level (Sage et al., 2025). Maddela et al. (2023) introduced PATTERNREFRAME, a dataset of 9,688 isolated crowdsourced thoughts annotated across 10 distortion categories, and demonstrated strong sentence-level classification with fine-tuned RoBERTa-large, achieving per-cluster accuracies of 63.8–97.6% on a 6-class grouping. While these results demonstrate that classification of de-contextualized statements is tractable, operating on isolated inputs strips away discourse-level cues and forces models to depend on shallow lexical patterns (Sage et al., 2025). This limitation becomes particularly acute in naturally occurring distortions, where distortion subtypes are often rooted in surrounding context.

Document-Level Approaches. Fewer studies have attempted classification on longer, document-level inputs. Shreevastava and Foltz (2021) introduced the THERAPISTQA dataset of 2,530 patient–therapist exchanges with distorted span annotations, representing one of the few publicly available resources with document-level structure. However, its inter-annotator joint probability agreement on distortion subtyping reaches only 33.7%, and no model in the original study exceeded 0.30 weighted F1 on classification. Subsequent work on the same data has improved upon these baselines, such as the multiple-instance learning framework proposed by Kim et al. (2025) that decomposes utterances into Emotion, Logic, and Behavior components, improving the multi-class F1 by over 9 percentage points. Most recently, Phan et al. (2025) proposed a three-stage pipeline, consisting of detection, span extraction via machine reading comprehension, and classification. Their work demonstrates that isolating distortion spans from surrounding text yields consistent improvements over classifying full utterances.

Our approach shares the pipeline intuition of Phan et al. (2025) but differs in two key respects. Whereas Phan et al. (2025) discard the surrounding narrative context by classifying bare extracted spans, our pipeline preserves contextual signal through a sentence windowing mechanism, reflecting the observation that distortion subtypes often depend on the situational context in which they occur (Sage et al., 2025). All of the studies above both train and evaluate exclusively on the THERAPISTQA dataset. While their results demonstrate clear improvements relative to prior baselines, optimizing and evaluating on a dataset with documented annotation limitations makes it difficult to disentangle genuine modeling improvements from overfitting to label noise. To that end, we take a different approach: We use the THERAPISTQA dataset only for training the retrieval component, where binary distortion presence annotations offer a more reliable inter-annotator agreement score of 61%, while training the classifier on an independently annotated PATTERNREFRAME dataset. Finally, the complete pipeline is evaluated on expert-annotated clinical vignettes from the CBT-CD dataset (Zhang et al., 2025) that neither component has seen during training.

LLM-Based Approaches and Evaluation. The difficulty of document-level CD classification is further evidenced by recent LLM benchmarking efforts. Zhang et al. (2025) introduced CBT-Bench, which includes the CBT-CD dataset: 166 clinical vignettes with multi-label annotations at the document level, annotated by domain experts. Their best LLM baseline (Mistral-7B-v0.3 (Jiang et al., 2023)) achieved only 45.3% weighted F1 in a zero-shot setting under the original 10-class taxonomy, while human expert performance stood at 49.1%, highlighting the considerable subjectivity inherent in distortion categorization. Beyond raw performance, LLM-based approaches to CD classification face broader methodological concerns, including prompt fragility, hallucinated rationales, and inconsistent cross-run behavior (Sage et al., 2025), compounded by computational requirements that limit their viability in resource-constrained clinical settings. These observations motivate our focus on task-specific fine-tuned transformers, which have been shown to match or exceed the performances of much larger LLMs on CD classification when trained on domain-relevant data (Maddela et al., 2023), a finding we aim to extend to out-of-sample, document-level settings.

3 Dataset

Our pipeline draws on three datasets that differ in granularity, annotation methodology, and intended use. No single resource provides both document-

Dataset	n	Labels	IAA	Use
THERAPISTQA	2,530	10+1	0.337[a]	Ranker
PATTERNREFRAME	9,688	10+1	0.355[b]	Classif.
CBT-CD	166	10	—	Eval

Table 1: Dataset summary. [a]Joint prob. of agreement on subtypes. [b]Krippendorff's α. Labels shows distortion categories (+1 = "No Distortion" class).

level structure and reliable subtype labels, and it is this complementarity that motivates our cascade design.

THERAPISTQA. Shreevastava and Foltz (2021) provide 2,530 patient–therapist exchanges annotated with a dominant distortion type and character-level distorted spans. The dataset exhibits well-documented annotation limitations: Inter-annotator agreement on distortion subtyping reaches only 33.7%, and disagreements are resolved by collapsing labels to a single agreed-upon dominant distortion, discarding genuine multi-label cases and introducing subtype-level noise. On the other hand, binary (distorted vs. non-distorted) agreement is substantially higher at 61% (Shreevastava and Foltz, 2021; Zhang et al., 2025; Sage et al., 2025). This asymmetry is taken into account in our design, where we use THERAPISTQA exclusively for ranker training, where the task is identifying whether or not a sentence is distorted (Appendix L). We split the THERAPISTQA dataset into 70/15/15 train/validation/test sets for ranker development.

PATTERNREFRAME. Maddela et al. (2023) provide 9,688 isolated, crowd-sourced thoughts, annotated across 10 distortion categories by trained workers. Although the dataset's Krippendorff's α of 0.355 on distortion labeling is numerically comparable to the THERAPISTQA figure, it is chance-corrected and PATTERNREFRAME preserves full 5-annotator vote distributions per example rather than collapsing annotations into a single dominant label. This enables training on soft label signals that encode genuine annotator disagreement as a property of the data. We use each category's share of the 5 annotator votes as a soft label for multi-label training, and use PATTERNREFRAME exclusively for classifier training. The dataset is split into 1920 training set instances, 961 validation set instances, and 6807 test set instances. To account for the high conceptual overlap among related patterns, the authors clustered the 10 original categories into 5 broader groups using a KMeans algorithm based on aggregated classifier outputs. This maps the patterns into Polarized Thinking (Overgeneralization, Catastrophizing, All-or-Nothing Thinking, Fortune Telling), Filtering (Mental Filtering, Discounting the Positive), Mislabeling (Personalization, Labeling and Mislabeling), alongside standalone categories for Mind Reading and Should Statements. Because the original dataset contains no undistorted examples, we append 'No Distortion' sentences from THERAPISTQA, whose higher binary agreement (61%) makes it a reliable source for this 6th target class.

CBT-CD. Zhang et al. (2025) provide 166 expert-annotated clinical vignettes with multi-label distortion annotations across 10 categories. Each vignette is structured as a separate situation block (the triggering event) and thoughts block (the patient's interpretations). The dataset is split into a 20-document seed set and a 146-document test set for final evaluation.

3.1 Taxonomy Harmonization

While the three datasets give our cascade the flexibility to separate ranking from classification, bridging them requires resolving taxonomy inconsistencies. A persistent challenge in CD research is the lack of a standardized category scheme, despite Burns' widely cited 10-category list (Burns, 1999) serving as a common reference (Sage et al., 2025). Table 2 shows our unified 11-class taxonomy, which aims to resolve these dependencies and allow cross-domain evaluation. This is accomplished with the help of two key merging decisions.

Emotional Reasoning → Fortune-telling. Both distortions derive from what Beck et al. (1979) termed *arbitrary inference*: "the process of forming an interpretation of a situation, event, or experience when there is no factual evidence to support the conclusion or when the conclusion is contrary to the evidence." They differ only in the non-evidential input: feelings for Emotional Reasoning, and expectations for Fortune-telling. Burns (1999) similarly groups Fortune-telling under "Jumping to Conclusions," defined as jumping to conclusions not warranted by the facts; Emotional Reasoning shares this evidence-free inferential structure. Mental Filtering, the other candidate, involves a distinct mechanism

Our Class	PR	CBT-CD
No Distortion	✓‡	—
All-or-nothing thinking	✓	✓
Discounting the positive	✓	—
Fortune-telling	✓	✓*
Labeling	✓	✓
Magnification	✓†	✓
Mental filter	✓	✓
Mind Reading	✓	✓
Overgeneralization	✓	✓
Personalization	✓	✓
Should statements	✓	✓

Table 2: Taxonomy mapping across datasets. *Includes merged Emotional Reasoning. ‡Added from THERAPISTQA. †Includes Catastrophizing.

(selective attention to negative details rather than unwarranted inference), making Fortune-telling the more semantically aligned target.

Discounting the Positive, No Distortion → Excluded. These categories exist in our classifier's label space but not in CBT-CD's annotation schema, so they are excluded from the 9-class evaluation.

4 Method

The core insight behind CASCADE is that document-level distortion detection can be decomposed into two sub-problems that existing resources already address separately: where distortions occur (from THERAPISTQA's span annotations) and their subtypes (from PATTERNREFRAME's subtype labels). By training a sentence ranker on the former and a subtype classifier on the latter, the two stages can be composed at inference time into a pipeline that handles document-level input without requiring document-level subtype annotations for training (Figure 1).

4.1 Stage 1: Distortion Ranker

Given a multi-sentence clinical document, the ranker must identify which sentences are most likely to contain distorted thinking. We frame this as a ranking problem rather than a binary classification problem because clinical narratives contain gradient levels of distorted language: A sentence describing a stressful event may carry emotional valence without being distorted, and a ranking objective allows the model to learn these relative distinctions rather than committing to a hard boundary.

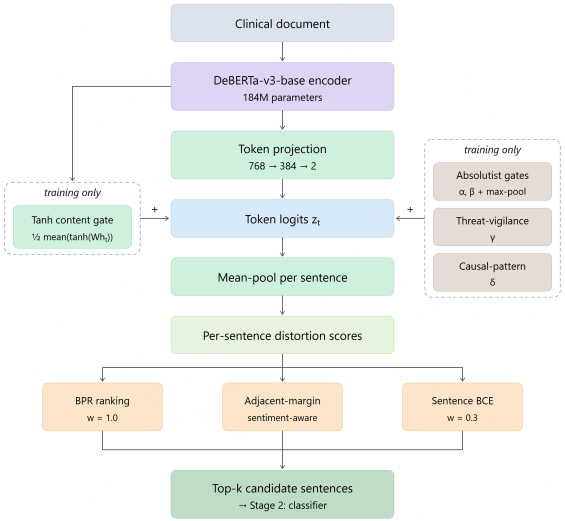


Figure 2: Distortion ranker architecture. The DeBERTa-v3-base encoder's token logits are augmented during training by three learnable prior gates: absolutist-word (α, β), threat-vigilance (γ), and causal-pattern (δ), which are disabled at inference. Token-level logits from the projection layer (768 → 384 → 2) are mean-pooled to produce sentence-level distortion scores. Training combines BPR pairwise ranking loss, a novel sentiment-aware adjacent-margin loss, and sentence-level BCE.

Architecture. The ranker is a DeBERTa-v3-base model (He et al., 2023) (184M parameters) that produces per-token distortion logits via a projection layer (768 → 384 → 2). Because many cognitive distortions are characterized by specific linguistic phenomena such as absolutist language ("always," "never"), threat-related vocabulary, and causal reasoning patterns ("because," "therefore") that are well-documented in the clinical psychology literature (Beck et al., 1979; Burns, 1999), we augment the encoder's token logits with three sets of prior gates that add learnable biases during training: (1) *absolutist-word gates* (α, β) that boost tokens matching absolutist language and their neighbors via max-pooling, (2) *threat-vigilance gates* (γ) for anxiety and rumination indicators, and (3) *causal-pattern gates* (δ) for logical connectives (Figure 2). These gates encode soft linguistic priors as inductive biases during training and are disabled at inference.

Training Objective. The ranker is trained on the THERAPISTQA dataset with document-level batching using a composite loss that addresses three complementary failure modes. The primary signal is a pairwise ranking loss (weight 1.0) based

on the Bayesian Personalized Ranking (BPR) log-sigmoid formulation (Rendle et al., 2009):

$$\mathcal{L}_{\text{rank}} = -\log \sigma(s_{\text{pos}} - s_{\text{neg}})$$

which pairs every distorted sentence against every non-distorted sentence in the same document, generating directly informative gradient signals per document.

However, pairwise ranking alone does not account for a domain-specific failure mode: Patient narratives are often saturated with intense emotional language that is affectively charged but not structurally distorted (Burns, 1999). To explicitly regularize against these artifacts, we introduce a novel adjacent-margin loss. For each distorted sentence at position i, it enforces a score gap over its immediate non-distorted neighbors at $i \pm 1$, with a stricter margin when the neighbor is a high-sentiment "confuser" that contains spurious words such as 'sad' or 'angry', but lacks structural distortion markers. Finally, we apply a sentence-level BCE loss (weight 0.3) to anchor the scores as interpretable probabilities. The loss weights and margin values are tuned on the THERAPISTQA validation set using mean reciprocal rank (MRR) as the selection criterion. We stabilize training with exponential moving average (EMA) of model weights (decay 0.999) (Polyak and Juditsky, 1992).

Inference. At inference, each sentence's distortion score is computed by mean-pooling its token-level logits, and the top-k sentences are returned as candidates for classification.

4.2 Stage 2: Multi-Label Classifier

The classifier must assign one or more distortion subtypes to each candidate sentence identified by the ranker. Two properties of this task shape our design choices: The classifier must generalize from PATTERNREFRAME's crowdsourced, decontextualized thoughts to CBT-CD's authentic clinical vignettes, and it must produce probability estimates that remain useful for thresholding rather than merely maximizing in-domain accuracy, since these probabilities will be aggregated across multiple sentences and thresholded downstream.

Architecture. The classifier shares the ranker's DeBERTa-v3-base backbone (184M parameters), with encoder weights initialized directly from the trained ranker. This transfer provides a warm start already attuned to distortion-relevant features in clinical language, partially bridging the domain gap between PATTERNREFRAME's crowd-authored thoughts and CBT-CD's authentic therapist–patient text. The [CLS] representation feeds two heads (Figure 3): a *classification head* (Dropout(0.1) → Linear(768 → 11), sigmoid) for multi-label prediction, and a *projection head* (768 → 768 → ReLU → 128, L2-normed) for supervised contrastive learning (Khosla et al., 2020). The classification head optimizes directly for label accuracy, while the contrastive head shapes the representation space so that samples sharing distortion types cluster together. This is done to ensure that the decision boundaries learned on synthetic data generalize to clinical text with different lexical distributions.

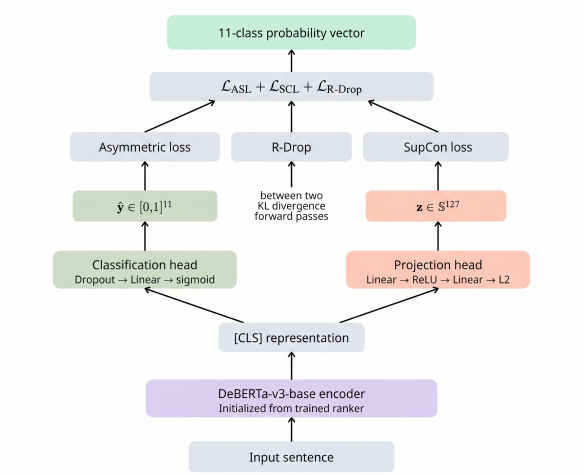


Figure 3: Multi-label cognitive distortion classifier architecture. The DeBERTa-v3-base encoder, initialized from the trained ranker, produces a [CLS] representation that feeds two heads: a classification head for multi-label prediction and a projection head whose L2-normalized outputs lie on the unit hypersphere S^{127} for supervised contrastive learning. The model is trained with a joint objective combining Asymmetric Loss, Supervised Contrastive Loss, and R-Drop regularization, outputting an 11-class probability vector over distortion subtypes.

Training Objective. The classifier is trained on PATTERNREFRAME using soft labels from 5-annotator vote distributions (the "No Distortion" target is set only when no distortion receives at least 2 votes). Because the classifier must generalize from crowdsourced, decontextualized thoughts to authentic clinical vignettes at infer-

ence time, we adopt Asymmetric Loss (ASL) (Ridnik et al., 2021) ($\gamma^- = 3.0$, $\gamma^+ = 0.0$, clip $= 0.1$), which down-weights gradients from high-confidence negative examples while maintaining full gradient flow for positive classes. ASL was designed to prevent the overconfident negative predictions that standard BCE produces in imbalanced multi-label settings (Ridnik et al., 2021); we find that this property is especially valuable under domain shift, as BCE's sharper but overconfident probability surfaces degrade when applied to out-of-distribution clinical text (Appendix E). In addition, we implement Supervised Contrastive Loss (Khosla et al., 2020) (weight 0.1, τ=0.07) on the projection head, pulling together representations sharing at least one distortion label, and R-Drop regularization (Liang et al., 2021) (weight 0.1) to enforce consistency between two stochastic forward passes. We optimize with AdamW and linear warmup, and maintain an exponential moving average of the weights (decay 0.999) (Polyak and Juditsky, 1992).

Hyperparameter Selection. We select all ASL hyperparameters by training multiple classifiers on PATTERNREFRAME and evaluating each through the full pipeline on the 20-document CBT-CD seed set via leave-one-out cross-validation (LOO-CV). Furthermore, we conduct hyperparameter tuning at the 6-class granularity groupings suggested by Maddela et al. (2023) rather than 9-class because the seed set contains only 20 documents, making it too small for reliable per-class optimization across 9 categories, but sufficient to distinguish among coarser-grained groupings.

4.3 Pipeline Composition

At inference, the two stages are composed as follows:

1. The input document is sentence-tokenized using NLTK (Bird et al., 2009) and interleaved using the procedure described in §4.4.
2. The ranker scores all sentences and returns the top-3 by distortion probability.
3. For each selected sentence, a ± 1 sentence context window is extracted from the interleaved document.
4. The classifier produces an 11-class probability vector for each windowed input.
5. The three probability vectors are aggregated via mean pooling across sentences.
6. Per-class thresholds (derived from seed LOO-CV) are applied to produce binary multi-label predictions.

4.4 Situation-Thought Interleaving

Cognitive distortions are inherently relational: A thought such as "I will never succeed" is distorted not in isolation but in relation to a specific triggering event. However, the CBT-CD dataset presents each vignette as separate situation and thought blocks, severing this referential link. Naïve concatenation forces the classifier to operate without the disambiguating contextual signal that a therapist would naturally have access to.

Our interleaving strategy reconstructs this event-interpretation adjacency using semantic similarity as the alignment signal.

Procedure. All situation and thought sentences are encoded using mean-pooled representations from the classifier's DeBERTa encoder. Then, a cosine similarity matrix is computed between each thought and every situation sentence. If at least half of the thoughts exceed a similarity threshold ($\tau = 0.15$) with their best-matching situation sentence, each thought is placed immediately after its highest-similarity match, while unmatched thoughts from the first half of the list go to the beginning of the document, and unmatched thoughts from the second half go to the end. When similarity is too sparse (fewer than half of thoughts exceed τ), thoughts are distributed proportionally across the situation by relative position.

5 Results

We employ a strictly leak-free evaluation protocol. All tuning decisions (per-class probability thresholds, searched uniformly over $[0.05, 0.95]$) are made via leave-one-out cross-validation (LOO-CV) on the 20-document seed set. The 146-document test set is left out until the final evaluation; we report this as our primary result. All reported pipeline results are averaged over 5 runs with different random seeds; LLM baselines (Mistral-7B-v0.3, Gemma-4 26B A4B) are similarly averaged over 5 independent runs under our 9-class taxonomy. We report weighted F1 at 9-class granularity and evaluate our performance against the established benchmarks as well as modern state-of-the-art LLMs.

5.1 Ranker Performance

The ranker achieves strong retrieval performance on the THERAPISTQA test set (240 documents with at least one distorted sentence): MRR = 0.743, Recall@1 = 0.476, and Recall@3 = 0.736. This indicates that the model typically places the first true distortion at rank 1 or 2, and, on average, the top-3 retrieved sentences successfully capture 73.6% of all true distorted sentences present within a document.

5.2 Classifier Standalone Performance

On the PATTERNREFRAME test set, the ASL-trained classifier achieves 71.9% macro F1 and 76.8% AUPRC on the 11-class multi-label setting. The model also achieves an accuracy of 83.7% on 6-class granularity. A BCE-trained variant scores higher in isolation (76.4% macro F1) but transfers worse through the pipeline. A detailed per-class comparison of the ASL-trained, BCE-trained, and RoBERTa-Large baselines at 6-class granularity is provided in Appendix D. We attribute this gap to the failure mode ASL was designed to address (Ridnik et al., 2021): BCE drives negative-class probabilities toward zero, producing overconfident predictions that generalize well in-domain but lose their edge under the distribution shift from crowd-sourced thoughts to clinical vignettes. Through the pipeline, BCE assigns a probability below 0.2 to 16.9% of gold-positive labels, where no threshold can recover them, whereas ASL assigns none (Appendix E).

5.3 End-to-End Pipeline Results

Because the original CBT-Bench baselines (Zhang et al., 2025) report results under a 10-class taxonomy that includes Emotional Reasoning as a separate category, we re-run the top LLM baselines under our 9-class taxonomy to ensure a fair comparison. We evaluate CASCADE against Mistral-7B-v0.3 (Jiang et al., 2023), Llama-3.1-8B (Grattafiori et al., 2024), and Gemma-4-26B-A4B (Google DeepMind, 2026).

Table 4 presents our headline results. At 9-class granularity, our 368M-parameter pipeline achieves 51.8% weighted F1. It exceeds the 7–8B LLM baselines, and the lower bound of its 95% bootstrap confidence interval (49.2%) remains above them. It performs on par with Gemma-4-26B-A4B (50.8%). We further evaluate Gemma-4 in a few-shot setting, using the same 20 seed documents available to our pipeline as in-context examples (Table 5). Performance peaks at 51.0% with 5 shots and degrades as more examples are added, and the difference between this configuration and our pipeline is not statistically significant ($p = 0.70$). Finally, direct comparison to human expert performance (49.1%) remains difficult, as it was measured under the original 10-class taxonomy.

Class	RoBERTa-L	Ours
Filtering	72.9	64.0
Mind Reading	77.3	**82.6**
Mislabeling	63.8	**81.7**
Polarized Thinking	79.3	**91.1**
Should Statements	93.1	**96.2**
None	**97.6**	83.6

Table 3: 6-class per-class accuracy (%) on the PATTERNREFRAME test set, comparing Maddela et al. (2023)'s RoBERTa-Large with our DeBERTa-v3-base classifier. Our model improves on the semantically complex categories (Mislabeling +17.9, Polarized Thinking +11.8) while using a smaller backbone.

System	Params	W-F1 (10-class)	W-F1 (9-class)
Human Expert	—	49.1	—
Mistral-7B-v0.3	7B	45.3	46.0
Llama-3.1-8B	8B	44.1	46.0
Gemma-4-26B-A4B	26B	47.0	50.8
Ours (Cascade)	**368M**	—	**51.8**

Table 4: 9-class CBT-CD weighted F1 (%). LLM baselines are re-evaluated under our 9-class taxonomy. Our pipeline matches Gemma-4 with 70× fewer total parameters. Few-shot results are in Table 5.

6 Conclusion

We present CASCADE (**C**ognitive **A**ssessment via **S**entence **C**ontext **a**nd **D**istortion **E**xtraction), a pipeline for document-level cognitive distortion detection that matches state-of-the-art performance on the CBT-CD benchmark (51.8% weighted F1) while using roughly 70× fewer total parameters than competitive LLM baselines. Our work demonstrates (1) that decomposing document-level classification into retrieval and classification stages is effective even with limited clinical training data, (2) that situation-thought interleaving that reconstructs the logical flow of clinical narratives, together with context windowing, provides meaningful gains by giving the classifier contextual signal, and (3) that principled tax-

Shots (k)	W-F1 (9-class)
0	50.8
2	49.8
5	**51.0**
10	49.2
15	48.9
20	47.9

Table 5: Few-shot scaling of Gemma-4-26B-A4B on CBT-CD (9-class weighted F1, %). k seed documents are added to the CBT-Bench prompt (Zhang et al., 2025) as examples. The best result (5-shot, 51.0%) is not significantly different from CASCADE (51.8%; $p = 0.70$; Table 15). CASCADE uses no in-domain training data. It uses the seed documents only for hyperparameter selection and threshold tuning via LOO-CV.

onomy harmonization enables leveraging heterogeneous datasets despite inconsistent annotation schemas.

Limitations

While the cascade architecture provides a degree of interpretability by surfacing specific sentences that the ranker identifies as likely distorted, the classifier itself remains a black box: It outputs a probability vector over distortion subtypes with no indication of why a particular label was assigned. For clinical decision support, therapists would need to understand not only which distortion was predicted but what linguistic or contextual cues drove the prediction. Incorporating post-hoc or inherent interpretability mechanisms, such as attention-based rationale extraction or prototype-based classification, remains an important direction for future work.

Moreover, the classifier is trained entirely on isolated, decontextualized single sentences from PATTERNREFRAME, yet at inference it receives windowed inputs (± 1 surrounding sentence) drawn from interleaved clinical narratives. Although this contextual windowing improves performance, the model was never trained to leverage context, meaning the gains are incidental rather than learned. A classifier fine-tuned on context-rich inputs could exploit this signal far more effectively. However, no existing dataset provides both reliable distortion subtype labels and long-form document structure. The construction of a dataset that combines expert-quality subtype annotations with authentic long-form clinical text would substantially advance this line of research by enabling classifiers to learn contextual disambiguation directly rather than relying on test-time heuristics.

Regarding the scope of our empirical claims, we note several factors that limit generalizability. All end-to-end results are reported on a single English-language benchmark (CBT-CD, 146 test documents). While we separately evaluate the ranker on THERAPISTQA and the classifier on PATTERNREFRAME, the full pipeline has not been validated on additional clinical populations, non-English languages, or alternative documentation styles.

Direct numerical comparison with the concurrent work of Phan et al. (2025) is not possible: Their system is trained and evaluated entirely on THERAPISTQA under its original taxonomy, while ours is evaluated on CBT-CD under a harmonized 9-class taxonomy with no THERAPISTQA subtype labels used in training.

Our taxonomy harmonization decisions also influence reported performance. The merging of Emotional Reasoning into Fortune-telling is theoretically motivated (§3.1) but is not the only defensible choice; the choice of merge target measurably affects performance: Merging Emotional Reasoning into Mental Filter instead of Fortune-telling reduces weighted F1 from 51.8% to 49.1% (Appendix G). The mapping has also not been independently validated by clinicians, and our 9-class results are not directly comparable to the original 10-class CBT-Bench numbers.

Furthermore, our use of fixed per-class probability thresholds across all subtypes heavily favors recall at the expense of precision, leading to the overprediction of certain classes such as Labeling. In practice, the pipeline predicts an average of 5.86 labels per document, of which 3.87 (66%) are false positives (Appendix K). While high sensitivity is often preferable in clinical settings to avoid missing potential distortions, it increases the verification burden on end-users. Although some classes such as Mind Reading show more balanced recall and precision (60% vs. 61%), implementing class-specific dynamic thresholds to better calibrate the precision-recall trade-off remains a clear area for future work.

The fixed top-k=3 selection may also limit the pipeline on longer, multi-turn narratives where distorted thinking is distributed across more than three sentences (Sage et al., 2025). Our top-k analysis (Appendix H) shows that k=3 performs best

on the vignette-length documents in CBT-CD, but this choice would need re-evaluation on longer inputs.

Finally, cognitive distortion expression is known to vary across cultures and linguistic backgrounds, and our system's behavior on populations underrepresented in the training data is unknown. We flag this as an important direction for future work: Any clinical deployment would require validation across the target population's demographic and linguistic profile.

Ethical Considerations

All three datasets used in this work are publicly available under their respective licenses and were released in anonymized form by their respective creators. We did not collect any new human data. We use all datasets consistently with their stated research purposes.

Our system is designed exclusively as a research tool and clinical decision-support aid; it is not intended to replace professional clinical judgment. Cognitive distortion detection carries inherent risks of misclassification that could lead to inappropriate therapeutic interventions if deployed without adequate clinical oversight. False positives could pathologize normal emotional responses, potentially leading a clinician to pursue unnecessary interventions; false negatives could cause significant distortions to go unnoticed, delaying appropriate therapeutic action. We emphasize that any use of this system must involve licensed mental health professionals who retain full decision-making authority.

Moreover, our work is developed and evaluated exclusively on English-language data, and its use in cross-cultural settings risks imposing a culturally specific framework on populations for whom it may not be appropriate, potentially leading to systematic misidentification of culturally normative thought patterns as pathological.

Finally, we used AI writing assistants for language polishing and editing during the preparation of this manuscript. All research ideas, experimental design, implementation, analysis, and conclusions are solely the work of the authors.

References

Aaron T. Beck. 1963. Thinking and depression: I. Idiosyncratic content and cognitive distortions. *Archives of General Psychiatry*, 9(4):324–333.

Aaron T. Beck, A. John Rush, Brian F. Shaw, and Gary Emery. 1979. *Cognitive Therapy of Depression*. Guilford Press, New York.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

David D. Burns. 1999. *Feeling Good: The New Mood Therapy*, revised edition. Avon Books, New York.

Google DeepMind. 2026. Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4. Accessed: 2026-05-24.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

S. C. Kaplan, A. S. Morrison, P. R. Goldin, T. M. Olino, R. G. Heimberg, and J. J. Gross. 2017. The Cognitive Distortions Questionnaire (CD-Quest): Validation in a sample of adults with social anxiety disorder. *Cognitive Therapy and Research*, 41(4):576–587.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jun Seo Kim, Hyemi Kim, Woo Joo Oh, Hongjin Cho, Hochul Lee, and Hye Hyeon Kim. 2025. Multi-view attention multiple-instance learning enhanced by LLM reasoning for cognitive distortion detection. In *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13641–13660, Toronto, Canada. Association for Computational Linguistics.

Amanda S. Morrison, Carrie M. Potter, Matthew M. Carper, Dina G. Kinner, Dane Jensen, Laura Bruce, Judy Wong, Irismar Reis de Oliveira, Donna M. Sudak, and Richard G. Heimberg. 2015. The Cognitive Distortions Questionnaire (CD-Quest): Psychometric properties and exploratory factor analysis. *International Journal of Cognitive Therapy*, 8(4):287–305.

Manh-Cuong Phan, Thi-Ngoc-Phuong Nguyen, Huu-Loi Le, Huy-The Vu, Hajime Hotta, and Minh-Tien Nguyen. 2025. From span extraction to classification: A multi-step framework for cognitive distortion analysis. In *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, pages 384–393, Hanoi, Vietnam. Association for Computational Linguistics.

Boris T. Polyak and Anatoli B. Juditsky. 1992. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 452–461.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91.

Archie Sage, Jeroen Keppens, and Helen Yannakoudakis. 2025. A survey of cognitive distortion detection and classification in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14884–14899, Suzhou, China. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Vasudha Varadarajan, Allison Lahnala, Sujeeth Vankudari, Akshay Raghavan, Scott Feltman, Syeda Mahwish, Camilo Ruggero, Roman Kotov, and H. Andrew Schwartz. 2025. Linking language-based distortion detection to mental health outcomes. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 62–68, Albuquerque, New Mexico. Association for Computational Linguistics.

Mian Zhang, Xianjun Yang, Xinlu Zhang, Travis Labrum, Jamie C. Chiu, Shaun M. Eack, Fei Fang, William Yang Wang, and Zhiyu Chen. 2025. CBT-BENCH: Evaluating large language models on assisting cognitive behavior therapy. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3864–3900, Albuquerque, New Mexico. Association for Computational Linguistics.

A LLM Prompt Template

All LLM baselines use the prompt format from CBT-Bench (Zhang et al., 2025), shown below. For k-shot experiments, the first k seed documents are inserted as examples in the same format, each followed by its gold answer letters. Emotional Reasoning predictions (choice h) are mapped to Fortune-telling (§3.1). Temperature is set to 0.0, and responses are parsed by extracting the letter labels.

```
You are a CBT therapist and now
need to determine the cognitive
distortions of a patient from
his current situation and
thoughts.
Each patient may have any number
of distortions.

[Few-shot only: k examples in
the format below, each ending
with "Answer: <letters>", then
"Now classify this patient:"]

Situation: {situation}
Thoughts: {thoughts}

Question: what distortions this
patient has?
Choices:
a: all-or-nothing thinking
b: overgeneralization
c: mental filter
d: should statements
e: labeling
f: personalization
g: magnification
h: emotional reasoning
i: mind reading
j: fortune-telling

Please only output the letters
corresponding to the choices.
Multiple choices should be
separated by a comma.
```

B Extended Per-Class Results

Table 6 provides the full per-class breakdown at 9-class granularity on the CBT-CD test set.

Class	F1	Prec.	Rec.	n
All-or-nothing	.632	.50	.85	65
Fortune-telling	.653	.52	.89	72
Labeling	.331	.20	1.00	29
Magnification	.329	.24	.52	25
Mental filter	.272	.17	.67	21
Mind Reading	.602	.61	.60	47
Overgeneralization	.335	.21	.84	32
Personalization	.484	.33	.93	42
Should statements	.568	.46	.75	28
Macro	**.467**	**.36**	**.78**	—
Weighted	**.518**	**.41**	**.81**	361

Table 6: Full per-class results on CBT-CD test set (9-class, frozen seed thresholds). All values are averaged over 5 random seeds.

C ASL Hyperparameter Sweep Details

We conducted an extensive sweep over ASL hyperparameters (γ^-, γ^+, clip, learning rate, SCL weight, R-Drop weight), training classifiers on PATTERNREFRAME and evaluating each through the full pipeline on the 20-document CBT-CD seed set via leave-one-out cross-validation.

Table 7 summarizes the hyperparameter search space. The full grid comprises $4 \times 3 \times 3 = 36$ ASL configurations, each trained for 20 epochs on PATTERNREFRAME and evaluated end-to-end through the pipeline on the seed set.

Hyperparameter	Search Values
γ^- (negative focusing)	2, 3, 4, 6
γ^+ (positive focusing)	0, 0.5, 1.0
Clip	0.01, 0.05, 0.10
Learning rate	2×10^{-5}
SCL weight	0.1
R-Drop weight	0.1

Table 7: ASL hyperparameter search space. Auxiliary loss weights and the learning rate are held fixed based on preliminary experiments.

We select all ASL hyperparameters by training multiple classifiers on PATTERNREFRAME and evaluating each through the full pipeline on the seed set via LOO-CV. We conduct hyperparameter tuning at the 6-class granularity rather than 9-class because the seed set contains only 20 documents, which are too few data points for reliable per-class optimization across 9 categories, but sufficient to distinguish among coarser-grained groupings.

D Classifier Standalone Comparison

Table 8 compares 6-class per-class accuracy on the PATTERNREFRAME test set across three systems: the RoBERTa-Large baseline from Maddela et al. (2023), our BCE-trained DeBERTa-v3-base (He et al., 2023) classifier, and our ASL-trained classifier (the model used in the final pipeline). While the BCE variant achieves higher standalone accuracy on most classes, the ASL variant transfers substantially better through the pipeline (Table 9).

Class	RoBERTa-L	BCE	ASL
Filtering	72.9	76.3	64.0
Mind Reading	77.3	95.0	82.6
Mislabeling	63.8	91.1	81.7
Polarized Thinking	79.3	91.2	91.1
Should Statements	93.1	99.4	96.2
None	97.6	96.8	83.6
Overall Accuracy	80.67	90.3	83.7

Table 8: 6-class per-class accuracy (%) on the PATTERNREFRAME test set. RoBERTa-Large numbers from Maddela et al. (2023) (Figure 4). BCE and ASL are both DeBERTa-v3-base models initialized from the ranker.

Table 9 visualizes a key finding: Standalone PATTERNREFRAME performance is a poor predictor of pipeline transfer. The ASL configuration achieving the best pipeline transfer ($\gamma^- = 3.0$, $\gamma^+ = 0.0$, clip $= 0.1$) scored only 71.9% macro F1 on the PATTERNREFRAME dataset, which is 4.5 percentage points below the scores obtained by the BCE configuration (76.4%). However, the ASL trained classifier achieved a weighted-F1 score of 51.8% on the overall pipeline, which is substantially better than its counterpart (49.6%). This is consistent with ASL's original design motivation (Ridnik et al., 2021): Aggressive negative suppression prevents the overconfident predictions that BCE produces in imbalanced multi-label settings, and, although ASL's probabilities are inflated overall, they are shifted upward, where per-class thresholds can correct them, rather than toward zero (Appendix E).

Loss	Standalone PR Macro-F1	Pipeline CBT-CD W-F1
BCE	**76.4**	49.6
ASL	71.9	**51.8**

Table 9: Standalone PATTERNREFRAME macro-F1 vs. pipeline CBT-CD weighted F1 (%).

E Calibration Analysis

Figure 5 shows reliability diagrams for the ASL and BCE classifier variants evaluated through the

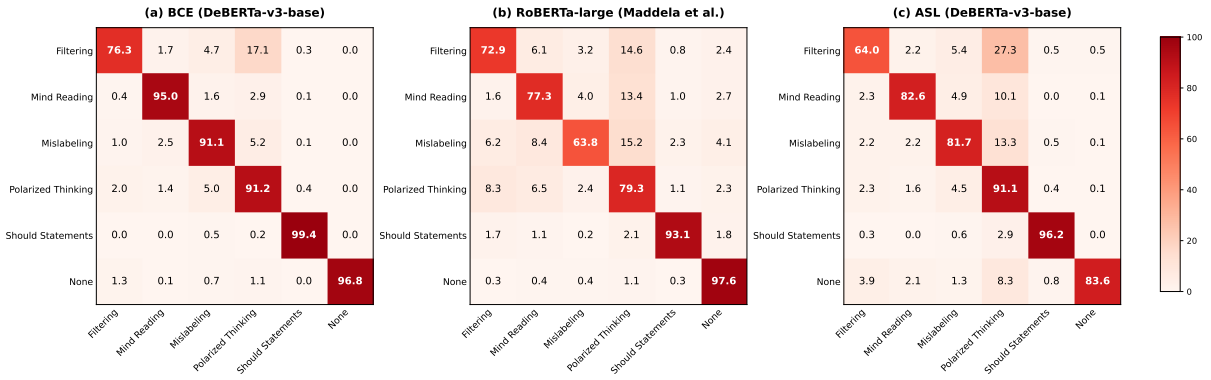


Figure 4: 6-class confusion matrices on the PATTERNREFRAME test set. Left: Our BCE-trained DeBERTa-v3-base classifier. Center: RoBERTa-Large baseline (Maddela et al., 2023). Right: Our ASL-trained DeBERTa-v3-base classifier.

full pipeline on the CBT-CD test set (1,314 class-level instances). The ASL variant has higher expected calibration error (ECE 36.8% vs. 18.8%), but it assigns no gold-positive instance a probability below 0.2, whereas BCE assigns 16.9% of gold positives to these lowest bins, where no threshold can recover them. ASL's overconfidence is corrected by the per-class thresholds, which explains why ASL transfers better through the pipeline (51.8% vs. 49.6% weighted F1) despite its higher ECE.

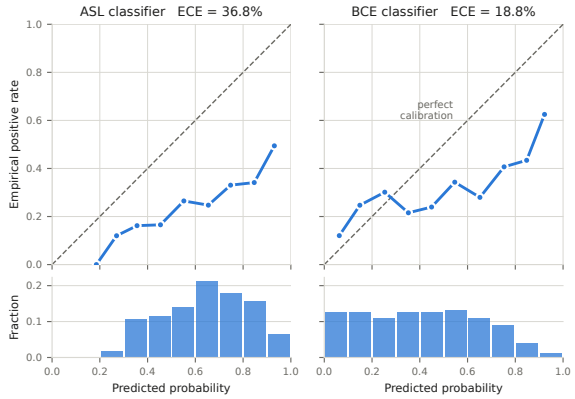


Figure 5: Reliability diagrams for the ASL (left) and BCE (right) classifier variants through the full pipeline. Dashed line: perfect calibration. Bottom panels: distribution of predicted probabilities.

F Precision–Recall Analysis

Table 10 compares weighted precision, recall, and F1 across systems on CBT-CD. LLM baselines from Zhang et al. (2025) report 10-class weighted metrics; our results use the merged 9-class taxonomy.

System	Params	W-Prec.	W-Rec.	W-F1
10-class				
Human Expert	—	51.1	48.1	49.1
Mistral-7B-v0.3	7B	43.8	56.4	45.3
Llama-3.1-8B	8B	39.7	59.9	44.1
Gemma-4 26B A4B*	26B	48.0	53.0	47.0
9-class				
Mistral-7B-v0.3*	7B	45.0	55.0	46.0
Llama-3.1-8B*	8B	44.1	61.9	46.0
Gemma-4 26B A4B*	26B	**51.1**	56.0	50.8
Cascade (ours)*	**368M**	40.8	**80.5**	**51.8**

Table 10: Weighted precision, recall, and F1 (%) on CBT-CD. * denotes averaged value over 5 runs

G Alternative Taxonomy Merge

Table 11 compares the reported merge (Emotional Reasoning → Fortune-telling) with the alternative (Emotional Reasoning → Mental Filter). The classifier is not retrained; only the gold-label mapping and per-class thresholds change. "Target F1" is the F1 of the class that absorbs Emotional Reasoning. Merging into Mental Filter reduces weighted F1 by 2.7pp: Fortune-telling F1 drops from 65.3 to 39.0, while Mental Filter F1 rises from 27.2 to 52.5. The reported merge therefore yields the higher overall score. This agrees with our theoretically motivated choice (§3.1), but it also shows that the 9-class results depend on the merge decision by several points.

Merge Target	W-F1 [95% CI]	Target F1
Fortune-telling (reported)	51.8 [49.2, 54.9]	65.3
Mental Filter	49.1 [46.4, 52.2]	52.5

Table 11: Alternative taxonomy merge on CBT-CD (9-class weighted F1, %).

H Ranking Stage Ablation

Table 12 isolates the contribution of the sentence ranker by varying which sentences are passed to the classifier. All conditions use the same classifier, context window (± 1 sentence), and aggregation. "Fixed" uses the per-class thresholds tuned for the reported k=3 configuration; "re-tuned" rederives thresholds via LOO-CV for each condition. The ranker's top-3 selection outperforms all alternative selection strategies.

Selection	Fixed thr.	Re-tuned thr.
Ranker top-3 (reported)	51.8	51.8
All sentences	49.4	49.7
Random 3 (10 seeds)	48.1 ± 0.7	47.5 ± 1.7
First 3	44.4	44.5
Ranker k=1	49.7	46.4
Ranker k=2	50.4	50.6
Ranker k=5	49.8	48.3

Table 12: Ranking stage ablation on CBT-CD (9-class weighted F1, %).

I Adjacent-Margin Loss

For each distorted sentence i in a document, the adjacent-margin loss enforces a score gap over its immediate non-distorted neighbors at positions $i \pm 1$. Let $\mathcal{P}$ be the set of such (distorted, neighbor) pairs (i, j) with $y_j = 0$:

$$\mathcal{L}_{\text{adj}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \max(0,\ m_{ij} - (p_i - p_j))$$

where $p_i = \sigma(s_i)$ is the sigmoid of the sentence score from mean-pooled token logits, and the margin m_{ij} is 0.7 if neighbor j is a high-sentiment confuser and 0.5 otherwise.

A neighbor is a "confuser" if it contains words from a negative-affect lexicon (e.g., "sad," "angry," "hopeless," "worthless") but lacks structural distortion markers (absolutist words or causal patterns). Pairs where the neighbor is itself distorted ($y_j = 1$) are excluded.

Loss weights (tuned on THERAPISTQA validation by MRR): BPR pairwise ranking 1.0, adjacent-margin 0.5, and sentence-level BCE 0.3.

J Prior Gate Formulation

Let $\mathbf{h}_t \in \mathbb{R}^{768}$ be the encoder's last hidden state for token t, and let z_t be the positive-class logit from the token projection layer ($768 \to 384 \to 2$). During training, the prior-augmented logit is:

$$\begin{aligned}\tilde{z}_t = z_t &+ \alpha\, r_t + \beta\, \text{MaxPool}_3(\mathbf{r})_t \\ &+ \gamma\, v_t(1 - r_t)(1 - c_t) + \delta\, c_t \\ &+ \tfrac{1}{2}\,\text{mean}(\tanh(\mathbf{W}\mathbf{h}_t))\end{aligned}$$

where:

- $r_t \in \{0, 1\}$: absolutist-word mask, set to 1 if token t and a lexicon entry match by substring in either direction.
- $\text{MaxPool}_3(\mathbf{r})_t$: max-pooling with kernel 3, spreading the absolutist signal to neighboring tokens.
- $v_t \in \{0, 1\}$: threat-vigilance mask, set to 1 for all tokens of a sentence that matches a threat-vigilance pattern, and zeroed where $r_t = 1$ or $c_t = 1$ to avoid double-counting.
- $c_t \in \{0, 1\}$: causal-pattern mask, set to 1 for all tokens of a sentence that matches a causal pattern.
- $\mathbf{W} \in \mathbb{R}^{768\times768}$: learned tanh gate over hidden states.

Initial values: $\alpha = 0.05$, $\beta = 0.02$, $\gamma = -0.30$, $\delta = 0.05$, temperature $\tau = 1.0$ (clamped to $[0.7, 3.0]$).

The final token logit is $\tilde{z}_t/\tau$, mean-pooled over non-special tokens to produce a sentence score. At inference the pipeline passes no masks, so $r_t = v_t = c_t = 0$ and only the encoder logits and the tanh gate contribute.

Lexicons.

- **Absolutist words** (21 entries): "always," "never," "everything," "nothing," "cannot," "can't," "cant," "impossible," "ruining," "everyone," "everybody," "no one," "nobody," "completely," "judge," "care at all," "no reason," "all," "forever," "must," "should."
- **Threat-vigilance patterns** (4 regular expressions): first-person fear or worry ("I'm afraid/scared/worried"), persistent worry ("always/constantly thinking/worried"), "overthinking," and "can't imagine"/"I can't stand."
- **Causal patterns**: "because," "since," "therefore," "so," "thus," "means that," "which means," "so now," "so that."

- **Confuser words** (adjacent-margin loss): "sad," "angry," "frustrated," "anxious," "depressed," "worthless," "useless," "hopeless," "overwhelmed," "stressed," "hate," "terrible," "awful," "scared," "afraid," "worried," "panic," "suicide," "miserable."

K Verification Burden

Table 13 quantifies the clinician verification effort required per document under the pipeline's operating point. The pipeline produces an average of 3.87 false-positive labels per document. This burden is the cost of the pipeline's recall-oriented operating point, which recovers 80.5% of gold labels compared with 53–62% for the LLM baselines (Table 10); per-class dynamic thresholding could shift this trade-off toward precision.

Metric (per document)	Mean
Predicted labels	5.86
True positives	1.99
False positives	3.87
False-positive rate	66%

Table 13: Verification burden on the CBT-CD test set.

L TherapistQA Design Rationale

We use THERAPISTQA exclusively for ranker training because of an asymmetry in its annotation reliability. Binary agreement (distorted vs. non-distorted) reaches 61%, which is sufficient for the ranker's task of identifying whether a sentence contains distorted thinking. Subtype agreement, however, is only 33.7% (joint probability), and disagreements are resolved by collapsing to a single dominant label, discarding genuine multi-label cases. Training a subtype classifier on these labels would conflate annotation noise with genuine label ambiguity. PATTERNREFRAME, with its preserved 5-annotator vote distributions, provides a cleaner signal for subtype classification despite its decontextualized single-sentence format.

M Extended Discussion

Interpreting the Precision–Recall Profile. CASCADE's weighted recall of 80.5% against a weighted precision of 40.8% (Table 10) represents a qualitatively different operating point from all LLM baselines, which cluster around 40–51% precision and 53–62% recall. The gap between our precision (40.8%) and Gemma-4's (51.1%) suggests that a meaningful share of CASCADE's F1 advantage stems from its willingness to overpredict rather than from its superior discriminative ability on ambiguous cases.

Context Dependence and the Role of Interleaving. Running the full pipeline without situation–thought interleaving and context windows (i.e., classifying extracted sentences without any surrounding context) yields a weighted F1 of 44.1%, which is a 7.7 percentage-point drop from the interleaved pipeline (51.8%).

The entire gain is recall: weighted precision is 40.8% in both configurations, while weighted recall rises from 66.8% to 80.5%, with the largest change in All-or-nothing thinking (recall .23 to .85). Table 14 summarizes the contribution of each component.

Component	Effect
Interleaving + context window	+7.7 W-F1
Ranker top-3 vs. all sentences	+2.4 W-F1
Tuned vs. uniform 0.5 thresholds	+1.0 W-F1

Table 14: Component attribution. W-F1 on CBT-CD

This pattern confirms that distortion subtype disambiguation is at least partly a discourse-level phenomenon; i.e., the same thought can be classified differently depending on whether the model has access to the event that prompted it. The fact that the classifier was never trained on interleaved or multi-sentence inputs, yet benefits substantially from receiving them at inference time, suggests that the current results likely underestimate the ceiling for context-aware distortion classification: A classifier fine-tuned on context-rich inputs could exploit this signal far more effectively.

Statistical Significance. Table 15 reports 95% bootstrap confidence intervals (10,000 resamples over 146 test documents). A paired approximate randomization test (10,000 permutations) shows that the difference between CASCADE and Gemma-4 5-shot is not statistically significant (95% CI of the difference $[-3.1, +5.0]$; $p = 0.70$). The CASCADE CI lower bound (49.2) exceeds the 7–8B baselines (46.0).

Threshold Sensitivity. Table 16 shows pipeline robustness to threshold perturbation. Per-class thresholds are derived from LOO-CV on the 20-document seed set. We perturb them by adding uniform noise $\mathcal{U}(-\epsilon, +\epsilon)$ independently per class,

System	W-F1 [95% CI]
CASCADE (ours)	51.8 [49.2, 54.9]
Gemma-4 5-shot	51.0 [47.4, 55.0]
Gemma-4 2-shot	49.8 [46.3, 53.6]
Gemma-4 0-shot	50.8
Mistral-7B-v0.3	46.0
Llama-3.1-8B	46.0

Table 15: Bootstrap confidence intervals on CBT-CD (9-class weighted F1, %).

over 1,000 draws for each ϵ. At $\epsilon = 0.05$, the lowest score across all draws is 49.7%. Even uniform thresholding at 0.5 retains 50.8%, confirming that the gain from per-class tuning is modest ($\sim$1pp).

Condition	W-F1 (%)
Per-class LOO-CV (reported)	51.8
Uniform threshold 0.5	50.8
Per-class ± 0.05 perturbation	51.4 ± 0.5
Per-class ± 0.10 perturbation	50.4 ± 0.8

Table 16: Threshold sensitivity analysis on CBT-CD (9-class weighted F1).

Limitations of State-of-the-Art Comparison. While our pipeline matches or exceeds all reported baselines on the 9-class CBT-CD benchmark, several caveats temper the strength of this claim. The LLM baselines in Table 4 are zero-shot, whereas our pipeline is tuned (via seed-set LOO-CV) on 20 in-domain documents; however, a few-shot setup with the same 20 documents improves Gemma-4 by at most 0.2 points (Table 5). The taxonomy merge from 10 to 9 classes is not neutral. It removes a category on which LLMs may have had distinctive strengths or weaknesses, making direct numerical comparison with the original 10-class CBT-Bench results inexact. Finally, the human expert baseline of 49.1% was measured under the 10-class taxonomy and has not been re-evaluated at 9 classes, and as such, the results on a cross-taxonomy comparison should be interpreted cautiously.