SEER: The Span-based Emotion Evidence Retrieval Benchmark

Aneesha Sampath
Computer Science & Engineering
University of Michigan
saneesha@umich.edu

Oya Aran
R&D Data Science & AI
Procter & Gamble
aran.o@pg.com

Emily Mower Provost
Computer Science & Engineering
University of Michigan
emilykmp@umich.edu

Abstract

We introduce the SEER (Span-based Emotion Evidence Retrieval) Benchmark to test Large Language Models' (LLMs) ability to identify the specific spans of text that express emotion. Unlike traditional emotion recognition tasks that assign a single label to an entire sentence, SEER targets the underexplored task of *emotion evidence* detection: pinpointing which exact phrases convey emotion. This span-level approach is crucial for applications like empathetic dialogue and clinical support, which need to know *how* emotion is expressed, not just *what* the emotion is. SEER includes two tasks: identifying emotion evidence within a single sentence, and identifying evidence across a short passage of five consecutive sentences. It contains new annotations for both emotion and emotion evidence on 1200 real-world sentences. We evaluate 14 open-source LLMs and find that, while some models approach average human performance on single-sentence inputs, their accuracy degrades in longer passages. Our error analysis reveals key failure modes, including overreliance on emotion keywords and false positives in neutral text.

1 Introduction

We introduce the Span-based Emotion Evidence Retrieval (SEER) Benchmark, which evaluates Large Language Models (LLMs) on their ability to identify which spans of text express emotion in real-world discourse. SEER consists of two tasks: identifying emotion evidence within (1) a single sentence or (2) a short passage of five sentences. These span-level tasks differ from traditional emotion recognition benchmarks, which assign a single label to an entire utterance and do not isolate the exact phrases where emotion is expressed. SEER contains new annotations on 1200 real-world sentences, making it comparable in size to related emotion benchmarks (Sabour et al., 2024). For Task 1 (single sentence), we use GPT-4.1 (Achiam et al., 2023) with human verification to label sentence-level emotion, with emotion evidence spans labeled by humans only. For Task 2 (five-sentence context), both the emotion labels and evidence spans are labeled by humans only. Figure 1 illustrates the SEER benchmark.

Emotion evidence refers to the spans of text that reveal a speaker's emotional state (Poria et al., 2021). Identifying such spans is critical for applications like empathetic dialogue and clinical telehealth sessions, where responses depend not just on knowing *what* emotion is present, but *how* it is expressed linguistically. For example, knowing that a person is 'sad' is less actionable than knowing that they said, 'nobody cares anymore.'

Most prior work frames emotion recognition either as a sentence-level classification task (Bharti et al., 2022; Alvarez-Gonzalez et al., 2021; Wagner et al., 2023) or as word-level tagging (Ito et al., 2020; Li et al., 2021). Sentence-level labels obscure which phrases convey emotion, while word-level tags often over-fragment emotionally coherent expressions (Hosseini and Staab, 2024). Span-level annotation offers a middle ground: localized enough for interpretability, but flexible enough to capture multi-word emotion cues.

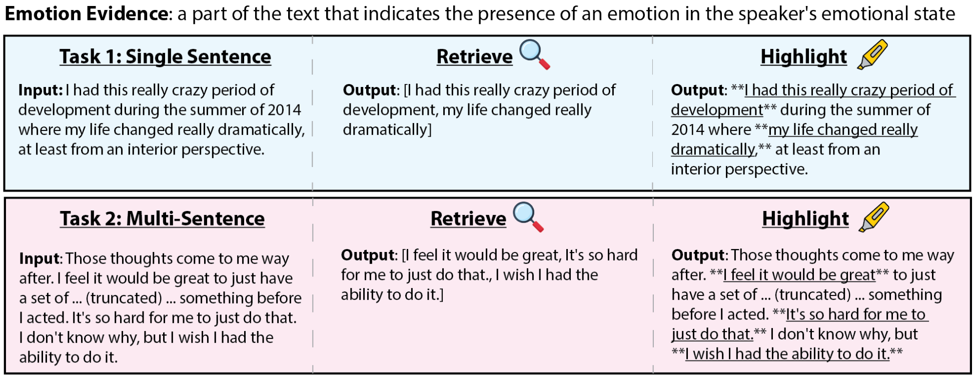
Related work in speech emotion recognition focuses on identifying *when* an emotion occurs in audio (e.g., at a certain timestamp), without linking those signals to the words used (Parthasarathy and Busso, 2016; Aldeneh and Provost, 2017). As a result, these approaches cannot answer which parts of the linguistic content express emotion.

Progress on span-based emotion evidence detection has been limited by two main challenges: (1) a lack of datasets with span-level annotations (most provide only utterance-level labels (Busso et al., 2008; Lotfian and Busso, 2017)), and (2) a lack of datasets grounded in real-world discourse (Poria et al., 2019; Busso et al., 2008).

We evaluate 14 open-source LLMs on the SEER

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1248–1267
December 20-24, 2025 ©2025 Association for Computational Linguistics

Figure 1: SEER includes two tasks: single- and multi-sentence emotion evidence identification. Each has two prompt formats: Retrieve (extract exact spans) and Highlight (mark spans in context). Task objectives are identical across formats. The text is truncated in the figure for space, but not in actual LLM input/output.



Emotion Evidence: a part of the text that indicates the presence of an emotion in the speaker's emotional state

Task 1: Single Sentence

Input: I had this really crazy period of development during the summer of 2014 where my life changed really dramatically, at least from an interior perspective.

Retrieve

Output: [I had this really crazy period of development, my life changed really dramatically]

Highlight

Output: **I had this really crazy period of development** during the summer of 2014 where **my life changed really dramatically.** at least from an interior perspective.

Task 2: Multi-Sentence

Input: Those thoughts come to me way after. I feel it would be great to just have a set of ... (truncated) ... something before I acted. It's so hard for me to just do that. I don't know why, but I wish I had the ability to do it.

Retrieve

Output: [I feel it would be great, It's so hard for me to just do that., I wish I had the ability to do it.]

Highlight

Output: Those thoughts come to me way after. **I feel it would be great** to just have a set of ... (truncated) ... something before I acted. **It's so hard for me to just do that.** I don't know why, but **I wish I had the ability to do it.**

benchmark. Our results show that while several models approach average human performance in single-sentence settings, their accuracy declines in multi-sentence contexts. Key failure modes include fixation on explicit emotion keywords (e.g., 'grateful') and false identification of emotion spans in neutral text. Future work could leverage SEER's span-level annotations to build models with better multi-word emotion identification and explore techniques to incorporate broader context to discourage keyword-matching. These directions can lead to LLMs that can more reliably pinpoint emotion expression in real-world discourse.

We publicly release all new annotations.[1] Users must obtain access to the original datasets separately before working with the full SEER benchmark to comply with licensing requirements.

2 Preliminaries

2.1 Definition of Emotion

Emotion refers to a complex reaction involving experiential, behavioral, and physiological components, typically triggered by a personally meaningful event or situation (American Psychological Association). Theories of emotion organize these within systematic frameworks.

The *categorical emotion* theory posits that basic emotions developed in response to evolutionary needs. These emotions can include happiness, surprise, fear, sadness, anger, and disgust (Ekman, 1992). The *dimensional emotion* theory maps emotion along valence (negative to positive) and activation (calm to excited) (Harmon-Jones et al., 2017; Russell, 1979). Since text provides a stronger signal for valence compared to activation (Wagner et al., 2023), we focus on valence only for dimensional emotion. We conduct error analysis on the SEER tasks for categorical emotions and valence.

2.2 Definition of Emotion Evidence

Emotion evidence is defined as "a part of the text that indicates the presence of an emotion in the speaker's emotional state. It acts in the real world between the text and the reader" (Poria et al., 2021). This should be distinguished from *emotion cause*, which is the "part of the text expressing the reason for the speaker to feel the emotion given by the emotion evidence" (Poria et al., 2021).

EXAMPLE: EMOTION EVIDENCE VS CAUSE (PORIA ET AL., 2021)

P_A: I have been accepted into graduate school!
P_B: What an amazing accomplishment!

P_B LABEL: happy / positive

CAUSE: accepted into graduate school
EVIDENCE: amazing accomplishment

3 Related Work

3.1 Emotion Hotspot Detection

Most emotion recognition work has targeted utterance-level classification, whereas *emotion hotspots* identify specific points at which emotion shifts and intensifies (Huang and Epps, 2016; Huang et al., 2015; Parthasarathy and Busso, 2016). Existing work on emotion hotspot detection has predominantly leveraged audio data. Some methods identify deviations from a baseline emotion state in valence-activation time-series traces (Parthasarathy and Busso, 2016, 2018). Other approaches partition an audio stream to answer 'which emotion appears when?' (Stemmer et al., 2023; Wang et al., 2023). The output of these audio-based methods

[1] https://github.com/chailab-umich/SEER

Table 1: Emotion benchmarks. H indicates hand-crafted, R indicates real-world, and S indicates LLM-generated.

Benchmark	Focus	Data Type
EmoBench (Sabour et al., 2024)	scenario understanding	H
EmoLLMs (Liu et al., 2024)	emotion recognition	R
EmotionQueen (Liu et al., 2024)	empathy generation	S
SEER (ours)	emotion evidence	R

is a set of timestamps and corresponding emotion. They do not identify the specific spans used to express it. Our work addresses this complementary task: identifying the discrete, linguistic spans of emotion evidence directly from text, a capability that remains underexplored.

3.2 Emotion Benchmarking in LLMs

Benchmarks such as EmoBench, EmotionQueen, and EmoLLMs evaluate LLMs on emotion-related tasks but differ from SEER (see Table 1). SEER evaluates the capability of LLMs to identify the precise spans where emotion evidence occurs. The other benchmarks have different goals, which we outline here. EmoBench (Sabour et al., 2024) evaluates emotional reasoning through hand-crafted scenarios with multiple choice answers. Given an input, "I have a teacher who gives the F grade as the highest mark... I saw he gave me an F," the LLM must identify the emotion of the speaker, and also the cause. This tests emotion recognition and emotion cause recognition, but uses hand-crafted scenarios. EmotionQueen (Chen et al., 2024) evaluates LLMs' ability to generate empathetic responses. Given a statement such as "I've been busy with work all day," an empathetic model response might be "Do you feel overwhelmed? Have you tried some ways to relax?", which provides emotional support, whereas a reply like "Hard work!" is considered non-empathetic. EmoLLMs (Liu et al., 2024) evaluates sentence-level emotion classification, where models assign emotion labels (e.g., happy, angry) to individual sentences. However, in all cases, these benchmarks do not localize the precise text that expresses emotion.

SEER tasks models with pinpointing the exact phrases that convey emotion, in both single-sentence and five-sentence passage settings. This focus on emotionally salient text spans grounded in real-world language fills a critical gap in existing emotion benchmarks.

4 The SEER Benchmark

The goal of SEER is to assess emotion evidence identification capabilities in LLMs. SEER comprises two primary tasks: single- and multi-sentence emotion evidence identification (Figure 1). All data are drawn from non-acted transcriptions of real-world speech (see Section 5) and annotated with both emotion labels and emotion evidence spans (see Sections 5.2 and 5.3 for annotation protocol).

4.1 Task Versions: Retrieve and Highlight

Each task has two versions: retrieve and highlight. In *retrieve*, the LLM must output a series of spans. The output can be empty if the LLM identifies no spans of emotion evidence. In *highlight*, the LLM must output the entire input passage, with the spans surrounded by '**' markers to indicate the start and end of an emotion evidence span.

These two prompt formats reflect real-world needs: Retrieve supports applications like evidence grounding or snippet retrieval, while Highlight supports scenarios requiring interpretable, in-context marking. To validate both formats, we conduct controlled prompting experiments with simple hand-crafted inputs (see Appendix B). Success in these setups suggests that failures on SEER tasks stem from challenges in processing real-world emotion, not formatting or retrieval deficiencies.

4.2 Task 1: Single-Sentence Emotion Evidence

LLMs must identify all emotion evidence that occurs within a single, non-neutral sentence. The goal is to isolate short-form emotion expression.

4.3 Task 2: Multi-Sentence Emotion Evidence

LLMs must identify all emotion evidence that occurs within a series of five consecutive sentences. We select five sentences as a starting point. This length is manageable for annotation and analysis, yet long enough to see whether models can track emotion expression across coherent discourse. The goal is to test emotion evidence tracking in longer, more variable contexts, where the overall emotion may shift over the course of the passage.

5 Datasets

5.1 Pre-Existing Datasets

We use samples from MSP-Podcast (Lotfian and Busso, 2017) and MuSE (Jaiswal et al., 2019) for

SEER. They contain non-acted speech (rather than scripted performances). We generate transcripts using Whisper[2] (Radford et al., 2023).

MSP-Podcast contains non-acted English conversational speech from podcasts and includes both categorical and dimensional emotion annotations (version 1.11) (Lotfian and Busso, 2017). We use the subset of the data that overlaps with the MSP-Conversation corpus version 1.1 (Martinez-Lucas et al., 2020), which contains continuous time-series trace annotations of dimensional emotion on speech. This is to allow for future research combining the strengths of both continuous and sentence-level labels. We use samples in the "Test1" split, totaling 2249 utterances.

The Multimodal Stressed Emotion (MuSE) dataset contains non-acted audiovisual English monologues (Jaiswal et al., 2019). It includes crowdsourced annotations for dimensional emotion. It totals 2648 utterances.

We collect new text-based annotations for categorical emotion and valence to align with the LLMs' input modality. The original MSP-Podcast and MuSE labels are audio- or video-based and the labels may not reflect textual cues. Using the original labels risks penalizing models for modality mismatch rather than genuine errors. In addition, MuSE lacks categorical emotion labels. Further, LLMs in Task 2 receive five-sentence context, which was not available to original annotators. Annotation details are in Sections 5.2 and 5.3.

5.2 Task 1 Data Annotation and Selection

Single Sentence Filtering. We filter the datasets to retain samples where utterances contain a single sentence only. We use NLTK (Bird et al., 2009) to tokenize by sentence. This leaves 1494 samples from MSP-Podcast and 1687 samples from MuSE.

GPT Labeling and Filtering. We use GPT-4.1 (Achiam et al., 2023) to ease the labeling burden. Prior work has shown GPT's capabilities for emotion labeling (Niu et al., 2024; Tarkka et al., 2024). We complement it with human verification.

We use the "gpt-4.1" model via the Azure OpenAI API to annotate the 1494 sentences from MSP-Podcast and 1687 sentences from MuSE for both valence (positive, negative, neutral) and categorical emotion (happy, sad, disgust, contempt, fear, angry, surprise, neutral). We select the eight categorical emotions that match the original label space from MSP-Podcast to encourage future research in the audio modality. The prompt is shown in Appendix B.5. We drop all samples that GPT-4.1 labeled as neutral. This leaves 488 sentences from MSP-Podcast and 854 sentences from MuSE.

Table 2: "% Annotator" represents the % of the time the annotators agreed with each other, and "% GPT" represents the % of the time both annotators marked agree with the GPT label.

Dataset	% Annotator	% GPT
MuSE (Categorical)	88.48%	87.31%
MuSE (Valence)	90.72%	90.25%
Podcast (Categorical)	69.14%	60.29%
Podcast (Valence)	71.05%	63.04%

Human Verification and Filtering. Two trained student workers then independently indicated agreement or disagreement with the GPT-4.1 labels. This study is IRB-approved (HUM00273067). The annotator and GPT-4.1 agreement is in Table 2.

We then filtered to only retain sentences where both annotators marked *agree* for both the categorical and valence GPT-4.1 labels of a single sentence. This leaves 215 samples from MSP-Podcast and 703 samples from MuSE.

Emotion Class Balancing. As a final filtering step, we balance the samples across the emotion classes by downsampling from over-represented classes. This leaves 30 samples per emotion class, except for surprise, which is slightly underrepresented with 20 samples, totaling 200 samples (103 from MSP-Podcast, 97 from MuSE). Since we balance by categorical emotion, valence is unbalanced since most of the emotion classes are negative. There are 155 negative and 45 positive samples. This is the final set of samples for Task 1.

Emotion Evidence Annotation. The student workers received a short training to define emotion evidence and were instructed to openly discuss examples. In the final annotation step, they jointly identified and labeled the *gold spans* of emotion evidence by discussing and highlighting the emotion evidence in the input text. This study is IRB-approved (HUM00273067).

5.3 Task 2 Data Annotation and Selection

Task 2 requires five *consecutive* sentences in order to maintain semantic cohesion. We first split the original 2249 and 2648 utterances from MSP-Podcast and MuSE, respectively, into single sen-

[2] openai/whisper-large-v2

tences using NLTK. We retain instances with a series of five consecutive sentences. We then remove overlapping instances (i.e., only including sentences 4-8 and 9-13, instead of 4-8 and 5-9). This totals 200 sets of five consecutive sentences.

The trained student workers were given each passage of five consecutive sentences, then discussed and jointly annotated the emotion (categorical and valence) of each sentence. They had access to all five sentences when annotating each sentence. This annotation step is necessary since the context of prior sentences can impact the emotion perception of a target sentence (Jaiswal et al., 2019). This results in emotion labels for each sentence within the series of five sentences. We did not use GPT for multi-sentence emotion annotation, since it is not validated in prior work (Niu et al., 2024).

For the final annotation step, the trained student workers jointly identified and labeled the gold spans of emotion evidence by discussing and highlighting their answers, as in Task 1.

Emotion Class Distribution Of the 1000 sentences (200 samples of 5 sentences each), they are 42.5% neutral, 27.8% happy, 11.6% sad, 5.1% surprise, 5.1% fear, 4% angry, 3.2% contempt, and 0.7% disgust. For valence, they are 42.7% neutral, 30.7% positive, and 26.6% negative. We do not perform balancing due to the nature of the emotion shifts within passages of longer discourse. The most common emotion transition between adjacent sentences are maintaining the current emotion or transitioning to and from neutral. The exact distribution of emotion transitions is in Appendix C.

6 Evaluation Metrics

We use two primary evaluation metrics: token-level F1-score (F1) and cosine similarity (Sim). F1 is a common metric for span-extraction tasks (Rajpurkar et al., 2016). It serves as a “fuzzy-match" metric. In addition to F1, embedding similarity metrics have also emerged as a way to assess semantic similarity as opposed to exact matches (Zhang* et al., 2020; Arabzadeh et al., 2024), which can serve to reduce penalties for differences in span boundaries in SEER tasks. We use sentence-BERT (Reimers and Gurevych, 2019) to embed spans, and then compute the cosine similarity between the embeddings of the gold and predicted spans. We do not report exact-match accuracy due to the subjective nature of span boundaries.

We use the Kuhn-Munkres Algorithm to align gold and predicted spans, which finds the optimal one-to-one matching between two sets (Luo, 2005). For example, consider two gold spans $\{g_1, g_2\}$ and three predicted spans $\{p_1, p_2, p_3\}$. The algorithm considers all possible matchings: (g_1, p_1), (g_2, p_2), (g_1, p_2), (g_2, p_1), etc. Each pairing is scored by ranking the similarity of the aligned span pairs, where similarity is defined as $\phi(g, p) = F_1(g, p)$, following Luo (2005). Since g and p are of unequal size, one span in p remains unmatched.

We compute a modified score for both F1 and Sim that penalizes a model for predicting an incorrect number of spans. This approach ensures that a high score is achieved only when a model identifies the correct spans and the correct number of them. The score is calculated as the sum of the metrics from the aligned spans, normalized by the greater of the number of gold or predicted spans. This penalizes both irrelevant predicted spans (false positives) and missed gold spans (false negatives). The formula for metric M (Sim or F1) is:

$$M = \frac{\sum \text{metric}_{\text{matched\_spans}}}{\max(\#\text{GoldSpans}, \#\text{PredictedSpans})} \quad (1)$$

7 Implementation Details

7.1 Prompts

We evaluate LLMs in two zero-shot prompt settings: Base prompting (Base) and chain-of-thought prompting (CoT), as in prior emotion benchmarking (Sabour et al., 2024). See Appendix B.5, Tables 11 and 10 for the exact prompts.

7.2 LLMs Evaluated

We evaluate the performance of 14 LLMs on the SEER benchmark. We select LLMs from the LLaMA (Dubey et al., 2024), Qwen (Yang et al., 2025, 2024), Phi4 (Abdin et al., 2024, 2025), and Gemma3 (Team et al., 2025) families.

We select models that achieve F1 $\geq$ 0.5 on prompting experiment three for further evaluation on the main SEER tasks. For the retrieve prompt, we retain all models with at least 1.7B parameters. For the highlight prompt, we retain models with at least 14B parameters, except for Qwen2.5-14B.

7.3 Experimental Setup

We use the huggingface transformers library[3] and load models in BF16. The full list of model checkpoint names is shown in Appendix Table 13. We

[3] https://github.com/huggingface/transformers

use the default hyperparameters and allow a maximum of three retries. For each model, we report the average and standard deviation across five runs.

We run experiments on an HPC with NVIDIA A40 GPUs. We use one GPU for models in the 0.5-14B range, two for 32B, and four for 70-72B.

8 Results

8.1 Task 1: Single-Sentence Emotion Evidence

The results are shown in Table 3. For the Retrieve-Base prompt, the Qwen-family models outperform all others, with every variant in the Qwen3 series achieving above 0.6 F1. Qwen3-32B achieves the highest score (0.673 F1, 0.693 Sim), while LLaMA3.2-3B performs the worst (0.193 F1). Notably, performance does not scale directly with model size: Qwen3-1.7B outperforms even the much larger LLaMA3.1/3.3-70B variants. Most models perform worse with Retrieve-CoT, with the exception of LLaMA3.2-3B, LLaMA3.1-70B, and LLaMA3.3-70B, which improve performance. We discuss CoT prompting in Section 10.1.

Performance drops for the Highlight prompt compared to Retrieve. This pattern is consistent with our prompting experiments, where we observe a typical performance drop of 0.3–0.4 F1 when comparing the same samples under Retrieve and Highlight settings (see Appendix B). This drop is expected, as Highlight requires models to reproduce the input text verbatim with added markup. Any hallucination results in an automatic score of 0. However, models in the 14–32B range drop only about 0.15 F1, while the 70–72B models drop about 0.2 F1. All models tested with the Highlight prompt perform better or similarly under the CoT prompt than the Base prompt.

8.2 Task 2: Multi-Sentence Emotion Evidence

The results are in Table 4. Notably, no model exceeds 0.41 F1 in any prompt version on Task 2, underscoring the need for models capable of emotion evidence identification in extended passages.

The Qwen-family models again perform best on Retrieve-Base, with Qwen3-14B and Qwen3-32B leading with 0.406 and 0.405 F1, respectively. LLaMA3.2-3B remains the weakest model (0.205 F1), and the larger LLaMA3.1/3.3-70B variants are again outperformed by most smaller models. These results further reinforce that model size does not directly predict performance.

Unlike Task 1 Retrieve, many models improve with CoT prompting in Task 2 Retrieve, including LLaMA3.2-3B, Phi4-Mini-3.8B, Phi4-14B, Qwen3-32B, LLaMA3.1-70B, and LLaMA3.3-70B. This suggest that CoT prompting is more effective for longer contexts, where reasoning steps may help localize relevant spans. For Highlight, this pattern persists: all models benefit from CoT prompts compared to Base. Highlight performance still falls short of Retrieve, as in Task 1.

9 Human Performance Comparison

We collected a crowdsourced human performance baseline for comparison with LLMs (IRB-approved). Details are in Appendix D. Each example received three independent annotations. The results are in Tables 3 and 4 in the ‘Human Annotator’ rows. The ‘Average’ row reflects the mean performance across all annotators, while the ‘Best’ row reports the per-sample maximum: i.e., the annotation associated with the best-performing crowdsourced annotator, compared to the gold annotations, over each sample.

Many LLMs outperform Average, but only Qwen3-32B in Task 1 slightly exceeds the Best-Human. This suggests that at least one annotator often identifies the emotion evidence in the gold labels, but the annotations of crowdsourced annotators are of variable quality. It is expected that untrained workers underperform relative to expert annotators given the nuance of emotion evidence identification. The LLM-Best-Human gap is larger in Task 2 (about .1 F1) than in Task 1 (about .01 F1), indicating that a crowdsourced annotator outperforms LLMs in longer contexts.

10 Error Analysis

10.1 Base and Chain-of-Thought Prompt

CoT prompting does not consistently improve performance across models in the Retrieve prompt setting (see Tables 3 and 4). This aligns with findings from Sabour et al. (2024), who reported that CoT prompting reduced or marginally changed performance on Emotion Intelligence tasks.

In Task 1 Retrieve, CoT prompting yields improvements for LLaMA3.1-70B and LLaMA3.3-70B, but degrades performance for all smaller models except LLaMA3.2-3B. In contrast, Task 2 Retrieve shows a less consistent trend: LLaMA3.2-3B, Phi4-Mini-3.8B, Qwen3-8B, Phi4-14B, Qwen3-32B, LLaMA3.1-70B, and LLaMA3.3-70B benefit

Table 3: F1 and cosine similarity (Sim) scores for Task 1 (Retrieve and Highlight). Each entry is averaged over five runs with standard deviations. '—' indicates the model was not evaluated in that prompt setting.

Model	**Retrieve (Base)**		**Retrieve (CoT)**		**Highlight (Base)**		**Highlight (CoT)**	
	F1	Sim	F1	Sim	F1	Sim	F1	Sim
0.5–2B								
Qwen 3 1.7B	0.620 ± .004	0.638 ± .004	0.314 ± .011	0.361 ± .015	—	—	—	—
3–4B								
LLaMA 3.2 3B	0.193 ± .017	0.219 ± .018	0.213 ± .026	0.234 ± .021	—	—	—	—
Phi 4 Mini 3.8B	0.283 ± .013	0.285 ± .013	0.275 ± .010	0.289 ± .009	—	—	—	—
Gemma 3 4B	0.556 ± .002	0.583 ± .003	0.483 ± .019	0.502 ± .022	—	—	—	—
Qwen 3 4B	0.611 ± .004	0.632 ± .005	0.348 ± .006	0.389 ± .011	—	—	—	—
8B								
LLaMA 3.1 8B	0.487 ± .010	0.518 ± .006	0.410 ± .019	0.438 ± .021	—	—	—	—
Qwen 3 8B	0.646 ± .004	0.623 ± .002	0.487 ± .012	0.512 ± .009	—	—	—	—
14B								
Phi 4 14B	0.542 ± .009	0.567 ± .009	0.505 ± .013	0.521 ± .013	0.428 ± .008	0.488 ± .008	0.500 ± .007	0.534 ± .007
Qwen 2.5 14B	0.589 ± .003	0.607 ± .003	0.516 ± .005	0.531 ± .005	—	—	—	—
Qwen 3 14B	0.658 ± .004	0.677 ± .003	0.523 ± .024	0.548 ± .025	0.509 ± .009	0.561 ± .008	0.508 ± .013	0.558 ± .011
32B								
Qwen 3 32B	**0.673 ± .006**	**0.693 ± .007**	0.575 ± .014	0.596 ± .011	0.553 ± .013	0.578 ± .012	0.579 ± .009	0.606 ± .005
70–72B								
LLaMA 3.1 70B	0.437 ± .006	0.467 ± .006	0.501 ± .015	0.522 ± .011	0.300 ± .005	0.403 ± .007	0.392 ± .005	0.447 ± .008
LLaMA 3.3 70B	0.429 ± .008	0.469 ± .008	0.534 ± .009	0.564 ± .006	0.253 ± .007	0.359 ± .005	0.331 ± .012	0.407 ± .009
Qwen 2.5 72B	0.614 ± .006	0.633 ± .007	0.610 ± .012	0.625 ± .011	0.413 ± .005	0.483 ± .002	0.507 ± .006	0.533 ± .005
Human Annotator								
Average	—	—	—	—	0.458 ± .033	0.494 ± .034	—	—
Best	—	—	—	—	0.672 ± —	0.675 ± —	—	—

from CoT prompting. This may reflect the nature of the longer input text in Task 2, where reasoning could assist span identification in longer passages.

10.2 Hallucination Rates

We define *hallucination rate* as the fraction of predicted spans that do not appear in the original text. We normalize each predicted span by removing punctuation and converting to lowercase, and then compare it to the similarly-normalized transcription. We mark a span as 'hallucinated' if it does not appear exactly as it is in the normalized text.

Models with lower hallucination scores reliably achieved higher F1 and similarity metrics. The Qwen family consistently showed the lowest hallucination rates across tasks and prompts.

Smaller models Phi4-mini-3.8B, LLaMA3.1-8B, and LLaMA3.2-3B exhibited the worst hallucination rates (16.3%, 15.2%, and 12.3% for Task 1 Retrieve-Base). CoT had minimal impact on hallucination. These three models along with Qwen3-1.7B also exhibited high hallucination rates on Task 2 Retrieve-Base, contributing to their poor performance. The models evaluated in the Highlight prompts for both Task 1 and 2 consistently exhibit low hallucination rates ($\leq 5\%$).

10.3 Errors by Emotion Category

We discuss performance by emotion category on the Base prompts only. See Figures 3, 4, 5 and 6 in Appendix E for visualizations.

For Task 1, performance by emotion category is variable. In both Retrieve and Highlight, emotion evidence identification on sentences expressing *disgust* and *anger* perform best (Figures 3a and 4a). The best performing model, Qwen3-32B, outperforms all other models on *disgust* and *happy* sentences in retrieve (Figure 3a), and outperforms other models on disgust, contempt, angry, and happy for highlight (Figure 4a). The Qwen model family consistently performs better on negative sentences compared to positive sentences (Figures 3b and 4b). This pattern is not consistent for the LLaMA, Gemma, and Phi families.

For Task 2, performance by emotion category resembles that of Task 1. However, unlike Task 1, Task 2 includes neutral sentences. The primary source of performance drop is the incorrect marking of emotion evidence in these neutral sentences. As shown in Figures 5c and 6c in Appendix E, models falsely identify emotion evidence in up to 50% of neutral sentences. This high rate of neutral false positives degrades performance across models.

10.4 Errors from Emotion Keyword Fixation

We probe over-reliance on salient emotion words by checking if models extract emotion keywords in isolation rather than the full span. We use the Empath lexicon (Fast et al., 2016) to identify instances where transcripts contain terms from the *positive-emotion* and *negative-emotion* categories. In 61 of the 200 sentences in Task 1 and 84 of the 1000 sen-

Table 4: F1 and cosine similarity (Sim) scores for Task 2 (Retrieve and Highlight). Each entry is averaged over five runs with standard deviations. '—' indicates the model was not evaluated in that prompt setting.

Model	**Retrieve (Base)**		**Retrieve (CoT)**		**Highlight (Base)**		**Highlight (CoT)**	
	F1	Sim	F1	Sim	F1	Sim	F1	Sim
0.5–2B								
Qwen 3 1.7B	0.253 ± .002	0.302 ± .003	0.248 ± .005	0.297 ± .007	—	—	—	—
3–4B								
LLaMA 3.2 3B	0.205 ± .009	0.234 ± .013	0.226 ± .007	0.255 ± .006	—	—	—	—
Phi 4 Mini 3.8B	0.224 ± .014	0.252 ± .013	0.238 ± .012	0.268 ± .012	—	—	—	—
Gemma 3 4B	0.368 ± .008	0.402 ± .008	0.332 ± .005	0.355 ± .007	—	—	—	—
Qwen 3 4B	0.335 ± .007	0.388 ± .006	0.281 ± .007	0.323 ± .005	—	—	—	—
8B								
LLaMA 3.1 8B	0.329 ± .006	0.367 ± .005	0.287 ± .009	0.326 ± .011	—	—	—	—
Qwen 3 8B	0.358 ± .006	0.362 ± .006	0.355 ± .008	0.379 ± .011	—	—	—	—
14B								
Phi 4 14B	0.342 ± .006	0.374 ± .009	0.357 ± .009	0.381 ± .009	0.259 ± .008	0.319 ± .007	0.346 ± .007	0.384 ± .008
Qwen 2.5 14B	0.362 ± .005	0.396 ± .005	0.348 ± .006	0.378 ± .006	—	—	—	—
Qwen 3 14B	0.406 ± .007	0.441 ± .006	0.354 ± .018	0.381 ± .020	0.263 ± .004	0.326 ± .005	0.315 ± .006	0.381 ± .009
32B								
Qwen 3 32B	0.405 ± .012	0.435 ± .012	**0.410 ± .008**	**0.437 ± .008**	0.322 ± .010	0.376 ± .009	0.388 ± .026	0.430 ± .026
70–72B								
LLaMA 3.1 70B	0.315 ± .006	0.352 ± .004	0.342 ± .021	0.374 ± .021	0.191 ± .006	0.268 ± .005	0.301 ± .013	0.365 ± .013
LLaMA 3.3 70B	0.280 ± .004	0.310 ± .003	0.345 ± .011	0.377 ± .013	0.150 ± .003	0.235 ± .003	0.247 ± .011	0.319 ± .014
Qwen 2.5 72B	0.391 ± .003	0.419 ± .004	0.350 ± .006	0.369 ± .006	0.284 ± .004	0.356 ± .004	0.363 ± .005	0.415 ± .009
Human Annotator								
Average	—	—	—	—	0.297 ± .069	0.336 ± .062	—	—
Best	—	—	—	—	0.506 ± —	0.533 ± —	—	—

tences in Task 2, a gold span contains an emotion keyword. We define a *fixation* as any prediction in which the model identifies only the keyword itself (e.g., predicting 'disgust' when the gold span is 'I would like to state my utter disgust.').

For both tasks, this error pattern appears most prominently in Highlight. For Highlight-Base, LLaMA3.1-70B and LLaMA3.3-70B exhibit high fixation rates of 27.9% and 26.6% for Task 1, and 36.2% and 53.6% for Task 2, respectively, where samples with an emotion keyword default to isolated words despite the emotion expressions themselves consisting of longer spans. In contrast, Qwen3-32B, the best-performing model, has only 1.6% of samples with this behavior in Task 1 and 5.7% in Task 2 (lowest fixation rate of all models). For Retrieve, most models have 0% fixation rates in Task 1, with the exception of LLaMA3.2-3B (3.6%), Qwen3-4B (1%), Phi4 (0.3%), LLaMA3.1-70B (1.6%), and LLaMA3.3-70B (6.9%). Similarly, in Task 2, the highest fixation rate is 7.1% (LLaMA3.2-3B), with most models falling within 0-3% fixation (except for Qwen3-4B with 5.7%).

CoT prompting partially mitigates this behavior for larger models. Fixation rates for LLaMA3.1-70B and LLaMA3.3-70B drop to 15.4% and 16.1%, for Task 1 Highlight-CoT and to 14.3% and 36.2% for Task 2, respectively, suggesting that reasoning steps can encourage more holistic span identification. However, this trend does not generalize across scales. In Task 1 Retrieve-CoT, smaller models Qwen3-1.7B and Qwen3-4B show increased fixation under CoT (11.8% and 9.8%, respectively), despite minimal errors with the base prompt.

These results underscore that keyword fixation is a nuanced failure mode. While CoT can guide larger models toward more nuanced span identification, it may also backfire in smaller models by drawing attention to more obvious word-level cues. Crowdsourced annotators also exhibit about 3% fixation rate in both tasks, suggesting that even human annotators are prone to this error mode.

11 Conclusion

In this paper, we propose the SEER Benchmark for evaluating LLM capability in emotion evidence identification. SEER comprises two tasks: single-sentence and multi-sentence emotion evidence identification in real-world discourse. We collect new annotations for 1200 sentences for emotion category, valence, and evidence. We evaluate SEER on 14 open-source LLMs and conduct a comprehensive error analysis. We find that models can somewhat reliably identify emotion evidence in single sentences, however, these models falsely identify emotion evidence in neutral sentences in multi-sentence contexts. Key error modes also include fixation on emotion keywords and modification of the input text (hallucination). Of the models we evaluate, Qwen3-32B performs the best in both SEER tasks.

Future work may explore whether performance on SEER aligns with standard emotion classification by evaluating the same LLMs on related tasks. SEER can also be evaluated on closed-source and reasoning models. Finally, SEER could be adapted to the audio modality by using the original datasets' annotations for evaluation on audio-LLMs.

Limitations

Sample Size. SEER is limited to 200 samples in Task 1 and 200 samples in Task 2. While this contains high-quality annotations for emotion evidence, emotion valence, and emotion category, and also is similar to the size of other emotion benchmarks (Sabour et al., 2024), we acknowledge that our dataset scale is limited, and could benefit from additional samples.

Prompt Tuning. We acknowledge that LLM outputs are highly sensitive to input prompts and that additional techniques could influence performance. We conducted extensive prompting experiments to mitigate this effect. Prompt design adjustments may impact the exact numerical scores, however we argue that they are unlikely to alter the overall trends observed across the tasks, as observed when comparing base and chain-of-thought prompting results.

Acknowledgments

This work was supported by Procter & Gamble. We used ChatGPT to assist with minor language editing in the paper and minor code debugging.

We thank Yara El-Tawil and James Tavernor for assisting with data annotation. We thank Stuart Synakowski, Minxue Niu, Mimansa Jaiswal, Arian Raje, and Victoria Lin for providing project feedback.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, and 1 others. 2025. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Zakaria Aldeneh and Emily Mower Provost. 2017. Using regional saliency for speech emotion recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2741–2745. IEEE.

Nurudin Alvarez-Gonzalez, Andreas Kaltenbrunner, and Vicenç Gómez. 2021. Uncovering the limits of text-based emotion detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2560–2583, Punta Cana, Dominican Republic. Association for Computational Linguistics.

American Psychological Association. Emotion. https://dictionary.apa.org/emotion.

Negar Arabzadeh, Amin Bigdeli, and Charles LA Clarke. 2024. Adapting standard retrieval benchmarks to evaluate generated answers. In *European Conference on Information Retrieval*, pages 399–414. Springer.

Santosh Kumar Bharti, S Varadhaganapathy, Rajeev Kumar Gupta, Prashant Kumar Shukla, Mohamed Bouye, Simon Karanja Hingaa, and Amena Mahmoud. 2022. Text-based emotion recognition using deep learning approach. *Computational Intelligence and Neuroscience*, 2022(1):2645381.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. 2024. Emotionqueen: A benchmark for evaluating empathy of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2149–2176.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657.

Eddie Harmon-Jones, Cindy Harmon-Jones, and Elizabeth Summerell. 2017. On the importance of both dimensional and discrete models of emotion. *Behavioral sciences*, 7(4):66.

Akram Sadat Hosseini and Steffen Staab. 2024. Disambiguating emotional connotations of words using contextualized word representations. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (* SEM 2024)*, pages 264–277.

Zhaocheng Huang and Julien Epps. 2016. Detecting the instant of emotion change from speech using a martingale framework. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5195–5199. IEEE.

Zhaocheng Huang, Julien Epps, and Eliathamby Ambikairajah. 2015. An investigation of emotion change detection from speech. In *INTERSPEECH*, volume 2015, pages 1329–1333. Dresden.

Tomoki Ito, Kota Tsubouchi, Hiroki Sakaji, Tatsuo Yamashita, and Kiyoshi Izumi. 2020. Word-level contextual sentiment analysis with interpretability. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4231–4238.

Mimansa Jaiswal, Zakaria Aldeneh, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. 2019. Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7415–7419. IEEE.

Zongxi Li, Haoran Xie, Gary Cheng, and Qing Li. 2021. Word-level emotion distribution with two schemas for short text emotion classification. *Knowledge-Based Systems*, 227:107163.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496.

Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 25–32.

Luz Martinez-Lucas, Mohammed Abdelwahab, and Carlos Busso. 2020. The msp-conversation corpus. *Interspeech 2020*.

Minxue Niu, Mimansa Jaiswal, and Emily Mower Provost. 2024. From text to emotion: Unveiling the emotion annotation capabilities of llms. In *Proc. Interspeech 2024*, pages 2650–2654.

Srinivas Parthasarathy and Carlos Busso. 2016. Defining emotionally salient regions using qualitative agreement method. In *Interspeech*, pages 3598–3602.

Srinivas Parthasarathy and Carlos Busso. 2018. Predicting emotionally salient regions using qualitative agreement of deep neural network regressors. *IEEE Transactions on Affective Computing*, 12(2):402–416.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, and 1 others. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

James A Russell. 1979. Affective space is bipolar. *Journal of personality and social psychology*, 37(3):345.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004.

Georg Stemmer, Paulo Lopez Meyer, Juan Del Hoyo Ontiveros, Jose A Lopez, Hector A Cordourier Maruri, and Tobias Bocklet. 2023. Detection of emotional hotspots in meetings using a cross-corpus approach. In *Proc. Interspeech 2023*, pages 1020–1024.

Otto Tarkka, Jaakko Koljonen, Markus Korhonen, Juuso Laine, Kristian Martiskainen, Kimmo Elo, and Veronika Laippala. 2024. Automated emotion annotation of finnish parliamentary speeches using gpt-4. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN)@ LREC-COLING 2024*, pages 70–76.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759.

Yingzhi Wang, Mirco Ravanelli, and Alya Yacoubi. 2023. Speech emotion diarization: Which emotion appears when? In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Hand-crafted Sentences

We design a set of hand-crafted sentences for the prompting experiments (detailed in Appendix B). The sentences include both neutral and emotion expressions. This allows us to evaluate whether models can reliably execute instruction-following behavior and return outputs in the expected format without the challenges of subtle and potentially ambiguous real-world language. Demonstrating robust performance under these conditions ensures that any failures observed in the main SEER tasks, which use real-world discourse, are not due to fundamental retrieval limitations or format misalignment, but instead reflect genuine challenges in understanding real-world emotion expression.

The sentences contain no repeating bi-grams across both the neutral and emotion sentences. The reasoning for this is described in Appendix B.3. We construct ten neutral sentences, two sentences for each categorical emotion in MSP-Podcast (Lotfian and Busso, 2017) (happy, sad, disgust, contempt, fear, anger, surprise), and five sentences for each valence (positive, negative). The full list of sentences is shown in Table 5.

B Prompting Experiments

In this section, we detail our experiments for the *retrieval* and *highlight* prompting styles. These allow us to identify which LLMs are capable of producing responses in our desired format.

B.1 Motivation

In some applications, it may be sufficient to have models retrieve only exact quotes of emotion evidence. In others, it may be necessary to place the emotion evidence back within the original context in which it was communicated. We observe that smaller models are generally capable of reproducing given text without hallucination, however they fail to reliably "highlight" a sentence within the given text. In order properly to evaluate emotion evidence capabilities in both smaller and larger models, we develop two sets of prompts: *retrieval* and *highlight*. The *retrieval* prompts evaluate LLM capability in retrieving exact quotes of emotion evidence only. The *highlight* prompts evaluate LLM capability in highlighting the exact regions of emotion evidence while also retrieving the original text without hallucination. These experiments are summarized in Table 6.

The goal of these experiments is to assess retrieval capacity, not to assess ability to identify subtle emotion evidence. Thus, for the prompting experiments, we use hand-crafted data only. These sentences are designed to unambiguously express either neutrality or a categorical emotion. See Appendix A for the full list of sentences.

B.2 Experiments

We include four experiments. Experiment 0 is a baseline experiment to assess LLM capability in reproducing passages without any modification. The LLM must reproduce the exact input text, unmodified. Experiment 1 requires the LLM to identify one span of text that occurs within the original text. The target span is provided in the prompt. This experiment uses neutral sentences only. Experiment

Table 5: Emotion-labeled sentences used in prompting experiments.

Emotion	Sentence	Gold Spans
happy	Everything felt perfect this morning.	Everything felt perfect
	This evening feels like a beautiful dream.	a beautiful dream
sad	Tears filled my eyes uncontrollably.	Tears filled my eyes uncontrollably.
	I felt so empty inside.	I felt so empty inside.
disgust	The sight turned my stomach upside down.	turned my stomach upside down
	I shivered because it was so revolting.	I shivered / it was so revolting
contempt	I sneered at the pathetic excuse.	I sneered / pathetic excuse
	I discarded his nonsense as laughable.	I discarded his nonsense as laughable.
fear	The door creak sent chills down my spine.	chills down my spine
	I began trembling when I heard the thunder.	trembling
anger	My blood is boiling with rage.	boiling with rage
	The disrespectful comment made me feel like I was going to explode.	I was going to explode
surprise	I blinked in disbelief.	I blinked in disbelief.
	My jaw dropped seeing the unexpected.	My jaw dropped seeing the unexpected.
positive	Warm laughter filled the room.	Warm laughter
	The sunlight was warm and inviting.	warm and inviting
	Aromas of fresh flowers brought a smile to my face.	brought a smile to my face
	Reflection on past events gives me hope for the future.	gives me hope
	I am grateful for all support I received.	I am grateful
negative	My hopes crumbled upon hearing the truth.	My hopes crumbled
	The regret of my actions haunted me.	regret / haunted me
	The tension was unbearable.	The tension was unbearable.
	I did not appreciate the comments.	I did not appreciate
	Dread filled me as I thought about the consequences.	Dread filled me
neutral	A book lies on the desk.	
	A clock shows the time.	
	Light travels in straight lines.	
	Clouds exist in the sky.	
	Rocks form over long periods.	
	Pens leave ink marks on paper.	
	Windows reflect ambient light.	
	Books contain printed pages.	
	A key is used to unlock a door.	
	Soil is made of rocks and minerals.	

2 requires the LLM to identify all spans of emotion evidence in the text. The LLM is given three span options in the prompt itself, in which either one or two of the three spans are correct. Experiment 3 requires the LLM to identify all spans of emotion evidence in the text, without any options provided in the prompt.

Each experiment has a retrieval and highlight version, except for Experiment 0, since there is no span identification involved. In the retrieval versions, the LLM must retrieve the specific spans of text only. In the highlight versions, the LLM must return the entire input text, and mark the specific spans by surrounding the spans with '**' markers.

We set the number of sentences (n_sentences) that the LLM must retrieve in $[1, 10]$, to identify if errors stem from the length of the input/output or from the nature of the experiment. For Experiment 0, we create ten variants for each $n \in$ n_sentences, where n neutral sentences are randomly samples and randomly shuffled. This totals 100 samples. For Experiment 1, we use the same logic as Experiment 0, except with n_sentences $\in [2, 10]$, since we need at least two sentences in order to be able to identify the target sentence. This totals 90 samples. For Experiment 2 and 3, which contain the same samples, we also set n_sentences $\in [2, 10]$. We create two variants for each emotion sentence, where three sets of neutral sentences are randomly selected and shuffled. The emotion sentence is randomly placed within the neutral sentences. This totals 432 samples.

B.3 Metrics and Constraints

We use metrics Exact-match accuracy (EM) and token-level F1-Score (F1) to evaluate LLM per-

Table 6: Overview of prompting experiments. Each experiment tests a different task objective. Retrieval variants require the LLM to output the relevant span. Highlight variants require the LLM to reproduce the input with the relevant span marked using a delimiter (**...**).

Experiment	Task Description
Exp 0	Reproduce the input text exactly, with no modifications.
Exp 1	Identify one span given the exact span in the instructions.
Exp 2	Identify all spans of emotion evidence given three span options in the instructions.
Exp 3	Identify all spans of emotion evidence in the original text.

Table 7: Exact-match accuracy (EM) and F1 scores for prompting Experiment 0 (baseline reproduction). We report the average and standard deviation over five runs.

Model	EM	F1
0.5–2B		
Qwen 3 0.6B	1.000 ± .000	1.000 ± .000
Gemma 3 1B	1.000 ± .000	1.000 ± .000
LLaMA 3.2 1B	0.730 ± .012	0.962 ± .011
Qwen 3 1.7B	0.832 ± .013	0.943 ± .005
3–4B		
LLaMA 3.2 3B	0.990 ± .000	1.000 ± .000
Phi 4 Mini 3.8B	0.710 ± .019	0.956 ± .003
Gemma 3 4B	1.000 ± .000	1.000 ± .000
Qwen 3 4B	1.000 ± .000	1.000 ± .000
8B		
LLaMA 3.1 8B	1.000 ± .000	1.000 ± .000
Qwen 3 8B	1.000 ± .000	1.000 ± .000
14B		
Phi 4 14B	0.986 ± .015	0.998 ± .002
Qwen 2.5 14B	1.000 ± .000	1.000 ± .000
Qwen 3 14B	1.000 ± .000	1.000 ± .000
32B		
Qwen 3 32B	1.000 ± .000	1.000 ± .000
70–72B		
LLaMA 3.1 70B	1.000 ± .000	1.000 ± .000
LLaMA 3.3 70B	1.000 ± .000	1.000 ± .000
Qwen 2.5 72B	1.000 ± .000	1.000 ± .000

formance on the prompting experiments. These metrics are used in the SQuAD benchmark for question-answering (Rajpurkar et al., 2016). We use these to compare the LLM output to the expected output.

B.4 Results

The results are shown in Tables 7, 8, and 9.

B.5 Prompts Provided to LLMs

The system and user prompts used for the main SEER tasks are shown in Table 11 and 10. The system prompts and user prompts used for the prompting experiments are shown in Tables 11 and 12.

For GPT annotation, we follow the approach of Niu et al. (2024) and provide the instructions and labeling schema in the system prompt, and provide the transcript itself in the user prompt.

The GPT annotation system prompt is: *"You are an emotionally-intelligent and empathetic agent. You will be given a piece of text, and your task is to identify the emotions expressed by the speaker. You are only allowed to make one selection from the following emotions: {set of emotions}. Do not return anything else."*

C Task 2 Emotion Transition Distribution

Figure 2 describes the emotion transitions between adjacent sentences for Task 2. The most common transition is neutral to neutral. Figure 2a shows the valence transitions, and Figure 2b shows the categorical emotion transitions.

D Annotator Recruitment

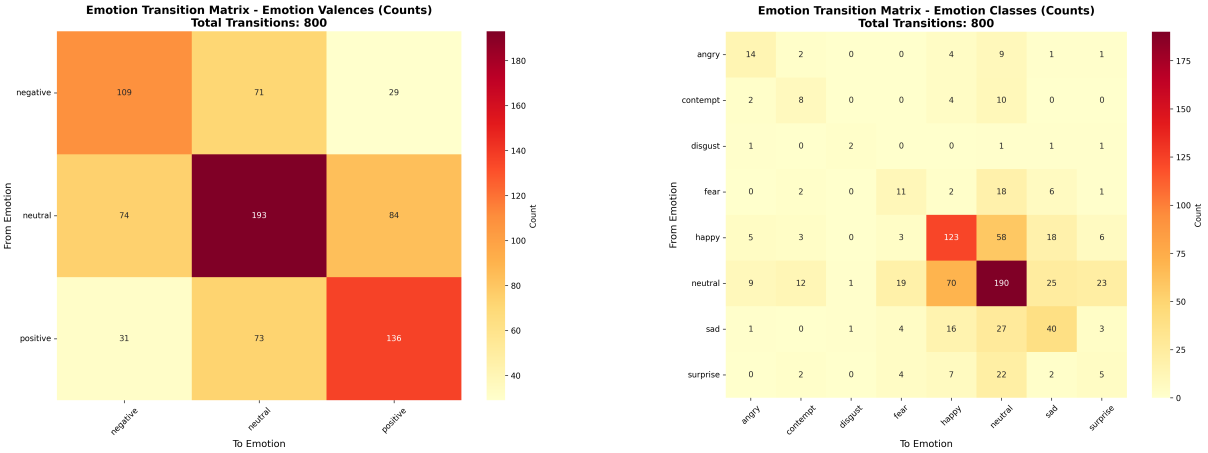
We recruited annotators using the Prolific platform[4]. We separated the 200 samples per task into surveys with 20 samples each to reduce annotation fatigue. We recruited three annotators per survey and paid them at a rate of $10 an hour. We recruited annotators that were (1) native English speakers, (2) residents of the USA.

The instructions were as follows: "In the following task, you will be asked to identify emotionally-relevant text. You will be presented with short passages and asked to identify the emotional text within the passage. Please use your mouse to highlight the regions of text that are emotionally salient. Ensure that you highlight only the specific text that expresses emotion." The participants were then asked to consent to the following: "(1) I have read and understood the information above, (2) I understand I might see potentially offensive or sexual content, and (3) I want to participate in this research and continue with the study," before proceeding to the main task.

[4] prolific.com

Table 8: Token-level F1 and cosine similarity (Sim) scores for retrieval prompting experiments. We report the average and standard deviation over five runs. *Categorical* and *Valence* indicate the model was evaluated on sentences crafted to target either a specfic emotion category or emotion valence.

Model	**Exp 1**		**Exp 2 – Categorical**		**Exp 2 – Valence**		**Exp 3 – Categorical**		**Exp 3 – Valence**	
	Sim	F1	Sim	F1	Sim	F1	Sim	F1	Sim	F1
0.5–2B										
Qwen 3 0.6B	0.814 ± .004	0.844 ± .003	0.430 ± .007	0.461 ± .008	0.382 ± .005	0.410 ± .003	0.438 ± .009	0.404 ± .013	0.314 ± .005	0.279 ± .014
Gemma 3 1B	0.768 ± .004	0.778 ± .002	0.594 ± .004	0.541 ± .005	0.578 ± .002	0.562 ± .003	0.439 ± .005	0.567 ± .006	0.419 ± .007	0.492 ± .008
LLaMA 3.2 1B	0.436 ± .008	0.192 ± .005	0.350 ± .006	0.343 ± .005	0.357 ± .013	0.364 ± .010	0.167 ± .012	0.064 ± .006	0.168 ± .012	0.074 ± .015
Qwen 3 1.7B	0.824 ± .002	0.833 ± .001	0.683 ± .005	0.689 ± .004	0.582 ± .005	0.595 ± .004	0.858 ± .005	0.915 ± .003	0.843 ± .006	0.885 ± .006
3–4B										
LLaMA 3.2 3B	0.817 ± .014	0.782 ± .011	0.700 ± .004	0.711 ± .004	0.548 ± .008	0.586 ± .006	0.824 ± .012	0.810 ± .008	0.845 ± .011	0.818 ± .013
Phi 4 Mini 3.8B	0.879 ± .007	0.877 ± .005	0.722 ± .005	0.708 ± .011	0.598 ± .009	0.578 ± .014	0.678 ± .015	0.815 ± .007	0.594 ± .015	0.777 ± .010
Gemma 3 4B	0.877 ± .001	0.883 ± .001	0.567 ± .001	0.601 ± .002	0.488 ± .003	0.547 ± .003	0.858 ± .003	0.831 ± .005	0.789 ± .005	0.780 ± .006
Qwen 3 4B	0.876 ± .000	0.891 ± .001	0.725 ± .005	0.725 ± .004	0.564 ± .002	0.575 ± .002	0.949 ± .005	0.962 ± .004	0.953 ± .005	0.973 ± .004
8B										
LLaMA 3.1 8B	0.876 ± .006	0.890 ± .004	0.754 ± .003	0.754 ± .006	0.642 ± .009	0.626 ± .006	0.979 ± .003	0.990 ± .002	0.982 ± .004	0.999 ± .001
Qwen 3 8B	0.833 ± .002	0.870 ± .001	0.699 ± .003	0.722 ± .003	0.614 ± .002	0.634 ± .002	0.946 ± .002	0.998 ± .001	0.943 ± .006	0.993 ± .003
14B										
Phi 4 14B	0.886 ± .006	0.884 ± .012	0.781 ± .008	0.798 ± .007	0.734 ± .005	0.755 ± .004	0.986 ± .006	0.997 ± .001	0.995 ± .006	0.999 ± .002
Qwen 2.5 14B	0.839 ± .001	0.879 ± .001	0.731 ± .002	0.726 ± .001	0.631 ± .004	0.670 ± .004	0.993 ± .002	0.997 ± .000	0.991 ± .000	0.989 ± .000
Qwen 3 14B	0.865 ± .000	0.912 ± .001	0.677 ± .003	0.687 ± .004	0.641 ± .003	0.675 ± .004	1.000 ± .000	1.000 ± .000	1.000 ± .000	1.000 ± .000
32B										
Qwen 3 32B	0.875 ± .003	0.899 ± .001	0.740 ± .004	0.739 ± .005	0.698 ± .005	0.740 ± .003	1.000 ± .000	1.000 ± .000	1.000 ± .000	1.000 ± .000
70–72B										
LLaMA 3.1 70B	0.828 ± .005	0.858 ± .003	0.777 ± .003	0.763 ± .002	0.677 ± .003	0.679 ± .003	0.937 ± .004	0.946 ± .004	0.928 ± .004	0.920 ± .003
LLaMA 3.3 70B	0.838 ± .002	0.861 ± .002	0.821 ± .004	0.802 ± .005	0.713 ± .003	0.704 ± .001	0.900 ± .003	0.918 ± .002	0.922 ± .001	0.942 ± .002
Qwen 2.5 72B	0.834 ± .002	0.871 ± .002	0.846 ± .002	0.832 ± .002	0.791 ± .002	0.816 ± .003	0.999 ± .001	0.999 ± .001	0.996 ± .000	0.997 ± .000



(a) Valence transitions.

(b) Categorical emotion transitions.

Figure 2: Emotion transitions between adjacent sentences for Task 2.

E Emotion Category Errors

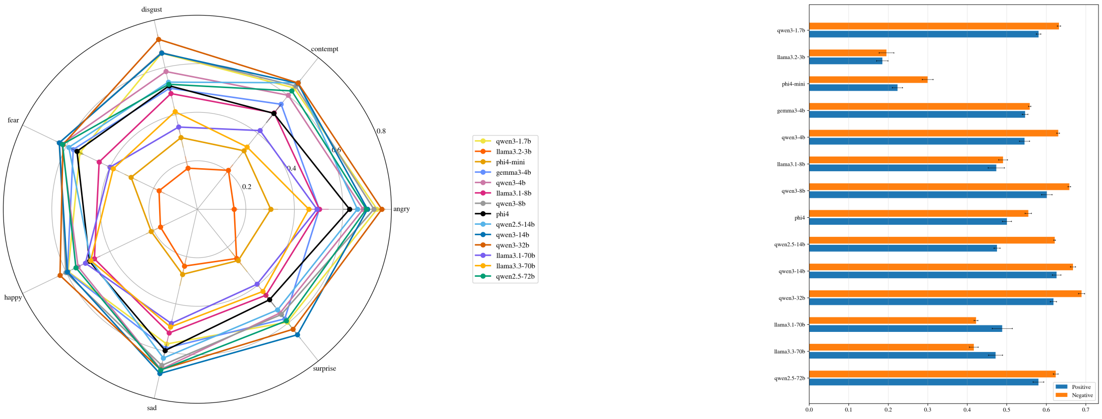
See Figures 3 and 4 for Task 1, 5 and 6 for Task 2.

F Model Checkpoints

The exact model checkpoints from huggingface transformers library are in Table 13.

Table 9: F1 and similarity (Sim) scores for highlight prompting experiments (Exp 1, Exp 2 – Categorical, Exp 2 – Valence, and Exp 3). We report the average and standard deviation over five runs. *Categorical* and *Valence* indicate the model was evaluated on sentences crafted to target either a specfic emotion category or emotion valence.

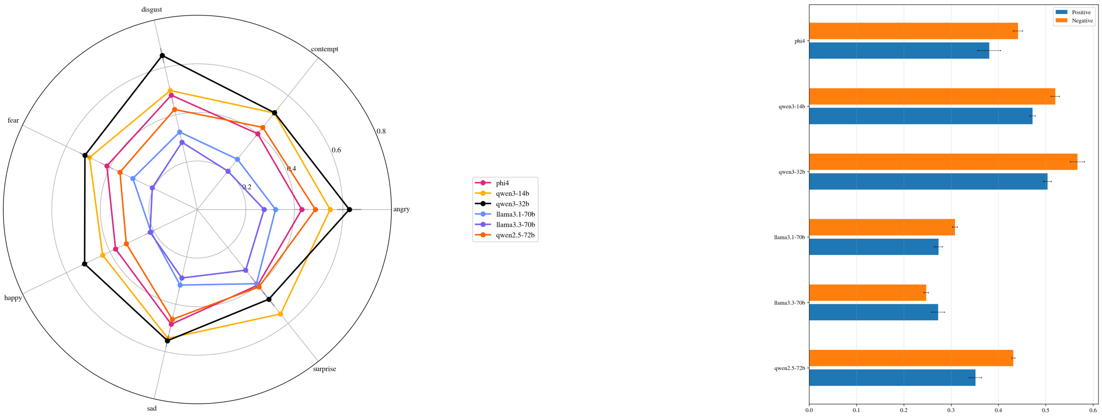
Model	**Exp 1**		**Exp 2 – Categorical**		**Exp 2 – Valence**		**Exp 3 – Categorical**		**Exp 3 – Valence**	
	F1	Sim	F1	Sim	F1	Sim	F1	Sim	F1	Sim
0.5–2B										
Qwen 3 0.6B	0.028 ± .004	0.046 ± .005	0.139 ± .007	0.183 ± .005	0.118 ± .007	0.117 ± .006	0.085 ± .005	0.122 ± .004	0.086 ± .005	0.086 ± .003
Gemma 3 1B	0.172 ± .008	0.245 ± .006	0.078 ± .002	0.098 ± .004	0.073 ± .002	0.075 ± .002	0.023 ± .001	0.029 ± .002	0.029 ± .000	0.035 ± .001
LLaMA 3.2 1B	0.000 ± .000	0.000 ± .000	0.024 ± .003	0.026 ± .004	0.026 ± .006	0.028 ± .005	0.039 ± .004	0.044 ± .004	0.036 ± .004	0.036 ± .004
Qwen 3 1.7B	0.659 ± .004	0.681 ± .004	0.471 ± .003	0.522 ± .003	0.371 ± .003	0.399 ± .003	0.287 ± .003	0.367 ± .003	0.259 ± .003	0.294 ± .003
3–4B										
LLaMA 3.2 3B	0.104 ± .011	0.117 ± .016	0.202 ± .010	0.218 ± .010	0.175 ± .012	0.179 ± .011	0.048 ± .009	0.056 ± .011	0.056 ± .008	0.061 ± .011
Phi 4 Mini 3.8B	0.035 ± .007	0.038 ± .009	0.348 ± .023	0.373 ± .025	0.397 ± .027	0.414 ± .029	0.329 ± .011	0.380 ± .013	0.322 ± .022	0.360 ± .020
Gemma 3 4B	0.264 ± .004	0.247 ± .005	0.283 ± .003	0.304 ± .003	0.247 ± .008	0.251 ± .008	0.227 ± .006	0.252 ± .007	0.180 ± .005	0.197 ± .005
Qwen 3 4B	0.562 ± .010	0.592 ± .005	0.665 ± .004	0.693 ± .003	0.651 ± .007	0.657 ± .007	0.207 ± .004	0.229 ± .004	0.267 ± .004	0.293 ± .002
8B										
LLaMA 3.1 8B	0.077 ± .009	0.095 ± .008	0.049 ± .004	0.054 ± .004	0.053 ± .002	0.055 ± .004	0.078 ± .004	0.085 ± .005	0.084 ± .006	0.101 ± .006
Qwen 3 8B	0.667 ± .009	0.674 ± .008	0.263 ± .006	0.289 ± .006	0.370 ± .006	0.366 ± .005	0.110 ± .006	0.129 ± .006	0.148 ± .003	0.158 ± .003
14B										
Phi 4 14B	0.906 ± .018	0.895 ± .020	0.808 ± .004	0.825 ± .003	0.714 ± .012	0.721 ± .014	0.698 ± .006	0.761 ± .003	0.606 ± .006	0.653 ± .003
Qwen 2.5 14B	0.714 ± .013	0.718 ± .009	0.761 ± .005	0.777 ± .004	0.706 ± .005	0.725 ± .005	0.484 ± .005	0.536 ± .005	0.366 ± .005	0.448 ± .004
Qwen 3 14B	0.877 ± .005	0.878 ± .007	0.695 ± .004	0.711 ± .004	0.608 ± .003	0.616 ± .006	0.620 ± .003	0.679 ± .004	0.499 ± .005	0.548 ± .005
32B										
Qwen 3 32B	0.975 ± .004	0.961 ± .002	0.738 ± .005	0.762 ± .004	0.683 ± .004	0.685 ± .003	0.706 ± .002	0.773 ± .002	0.597 ± .004	0.649 ± .002
70–72B										
LLaMA 3.1 70B	0.850 ± .024	0.853 ± .023	0.657 ± .001	0.712 ± .002	0.654 ± .007	0.705 ± .007	0.594 ± .008	0.668 ± .006	0.536 ± .014	0.681 ± .011
LLaMA 3.3 70B	0.908 ± .002	0.903 ± .005	0.663 ± .004	0.731 ± .002	0.735 ± .003	0.800 ± .003	0.524 ± .002	0.624 ± .002	0.590 ± .005	0.732 ± .005
Qwen 2.5 72B	0.958 ± .002	0.978 ± .001	0.809 ± .003	0.856 ± .003	0.749 ± .003	0.808 ± .003	0.707 ± .001	0.785 ± .002	0.638 ± .001	0.745 ± .001



(a) Emotion-wise F1 scores for Task 1 (Retrieve-Base).

(b) Valence F1 scores for Task 1 (Retrieve-Base).

Figure 3: Task 1 (Retrieve-Base). (a) Per-emotion F1 scores. (b) Per-valence F1 scores.



(a) Emotion-wise F1 scores for Task 1 (Highlight-Base).

(b) Valence F1 scores for Task 1 (Highlight-Base).

Figure 4: Task 1 (Highlight-Base). (a) Per-emotion F1 scores. (b) Per-valence F1 scores.

Table 10: User prompts for SEER Tasks, including Base and CoT variants.

User Base Prompt (Retrieve)
Begin Instructions You are given text. Some spans are emotionally expressive. Return only the full unmodified emotionally expressive spans, and nothing else. *End Instructions* `{text}`
User CoT Prompt (Retrieve)
Begin Instructions You are given text. Some spans are emotionally expressive. Return the full unmodified emotionally expressive spans. Reason step-by-step and explore the emotion content. Output "Reasoning:" and then your reasoning steps. After reasoning, output "Response:" followed by the spans, and nothing else. *End Instructions* `{text}`
User Prompt (Highlight)
Begin Instructions You are given text. Some spans are emotionally expressive. Surround the emotionally expressive spans with ‘**‘. Return only the full unmodified text with those markers, and nothing else. *End Instructions* `{text}`
User CoT Prompt (Highlight)
Begin Instructions You are given text. Some spans are emotionally expressive. Surround the emotionally expressive spans with ‘**‘. Return the full unmodified text with those markers. Reason step-by-step and explore the emotion content. Output "Reasoning:" and then your reasoning steps. After reasoning, output "Response:" followed by the full unmodified text with those markers, and nothing else. *End Instructions* `{text}`

Table 11: System prompts used in both prompting experiments and main SEER tasks. The prompts for experiments 2 and 3 are used in the main SEER tasks.

System Prompt (Experiment 0)

Instructions

1. **Immutable Input**
   - Never delete, normalize, split, merge, or alter any original character (letters, apostrophes, punctuation, whitespace).
2. **Self-Check**
   - Verify that the remaining text is identical to the input. Retry until it passes.

System Prompt (Experiment 2 and 3 Retrieval)

Instructions

1. **Immutable Input**
   - Never delete, normalize, split, merge, or alter any original character (letters, apostrophes, punctuation, whitespace).
2. **Subjective Emotion Only**
   - Only retrieve spans that reveal the speaker's internal emotional state or attitude.
   - This includes:
     - Explicit emotion words
     - Implicit cues/phrases of feeling or reaction
   - Do not retrieve:
     - Purely factual or descriptive statements
     - Neutral descriptions of events without any sentiment
3. **Self-Check**
   - Verify that the remaining text is identical to the input. Retry until it passes.
4. **Output**
   - Return all spans on a single line, with each span separated by " | ". If there is only one span, do not include the " | ". No headers, no metadata, no removed, added, or modified words.

System Prompt (Experiment 1 Highlight)

Instructions

1. **Immutable Input**
   - You may only insert ** markers.
   - Never delete, normalize, split, merge, or alter any original character (letters, apostrophes, punctuation, whitespace).
2. **Self-Check**
   - After inserting your markers, remove all ** and verify that the remaining text is identical to the input. Retry until it passes.
   - Any pair of ** must surround the entire span.

System Prompt (Experiment 2 and 3 Highlight)

Instructions

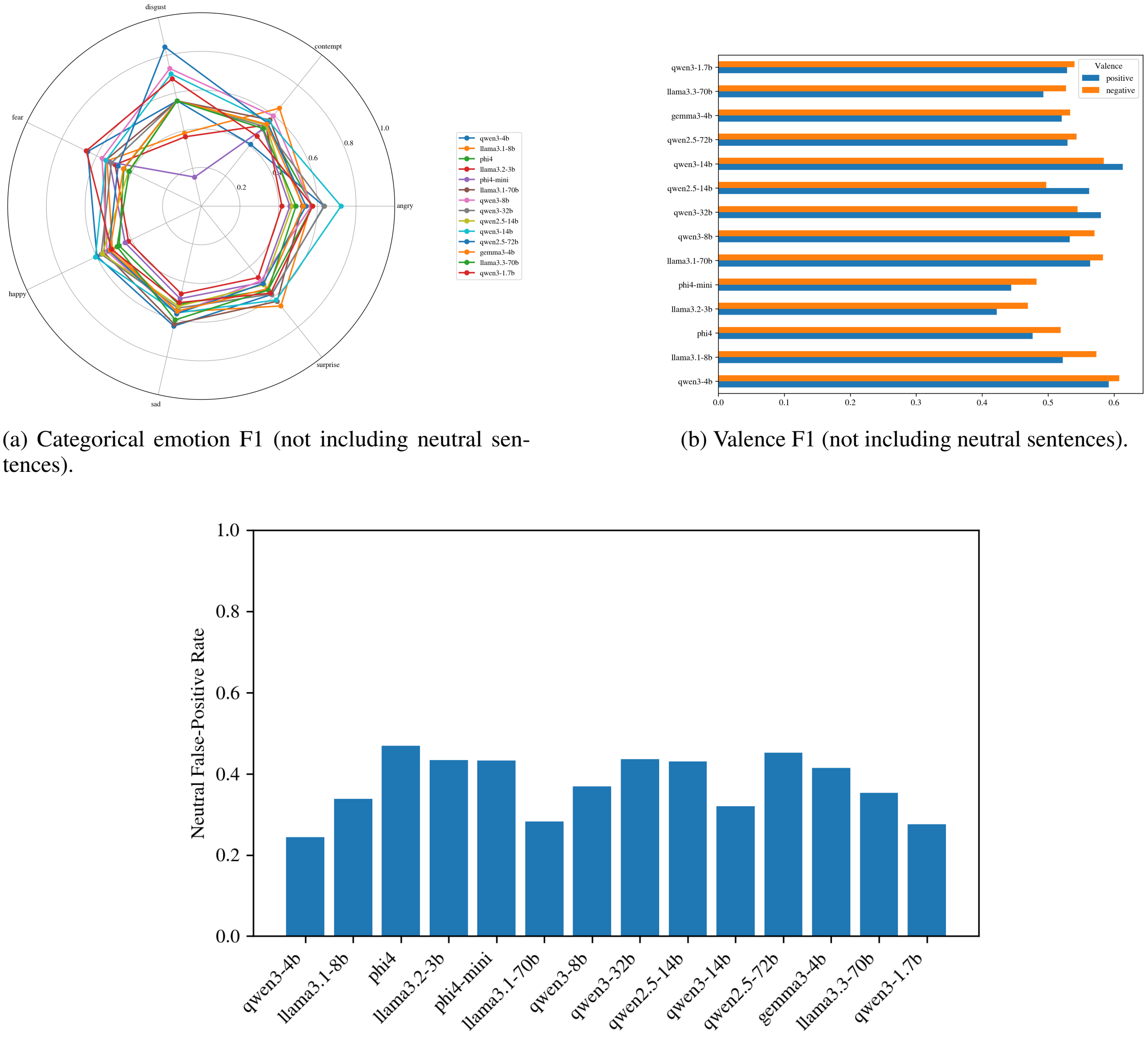
1. **Immutable Input**
   - You may only insert ** markers.
   - Never delete, normalize, split, merge, or alter any original character (letters, apostrophes, punctuation, whitespace).
2. **Subjective Emotion Only**
   - Only highlight spans that reveal the speaker's internal emotional state or attitude.
   - This includes:
     - Explicit emotion words
     - Implicit cues/phrases of feeling or reaction
   - Do not mark:
     - Purely factual or descriptive statements
     - Neutral descriptions of events without any sentiment
3. **Self-Check**
   - After inserting your markers, remove all ** and verify that the remaining text is identical to the input. Retry until it passes.
   - If the input is completely neutral, return it unchanged, with no markers.
   - Any pair of ** must surround the entire span.
4. **Output**
   - Return only the marked text. No headers, no metadata, no removed, added, or modified words.

Table 12: User prompts used across experiments. Placeholders `{text}` and `{target_sentence}` represent inputs given during evaluation.

User Prompt (Experiment 0)
Begin Instructions You are given a series of sentences. Return only the full unmodified text, and nothing else. *End Instructions* `{text}`
User Prompt (Experiment 1: Highlight)
Begin Instructions You are given a series of sentences, which contains target sentence: `{target_sentence}`. Surround the target sentence with `**`: Return only the full unmodified text with those markers, and nothing else. *End Instructions* `{text}`
User Prompt (Experiment 1: Retrieval)
Begin Instructions You are given a series of sentences, which contains target sentence: `{target_sentence}`. Return only the full unmodified target sentence, and nothing else. *End Instructions* `{text}`
User Prompt (Experiment 2: Highlight)
Begin Instructions You are given a series of sentences. One sentence is emotionally expressive. Surround the emotionally expressive sentence with `**`: Return only the full unmodified text with those markers, and nothing else. *End Instructions* `{text}`
User Prompt (Experiment 2: Retrieval)
Begin Instructions You are given a series of sentences. One sentence is emotionally expressive. Return only the full unmodified emotionally expressive sentence, and nothing else. *End Instructions* `{text}`

Table 13: Hugging Face model checkpoint names.

Model	**Reasoning**	**Hugging Face Checkpoint**
Qwen 3 0.6B	✓	Qwen/Qwen3-0.6B
LLaMA 3.2 1B	×	meta-llama/Llama-3.2-1B-Instruct
Qwen 3 1.7B	✓	Qwen/Qwen3-1.7B
LLaMA 3.2 3B	×	meta-llama/Llama-3.2-3B-Instruct
Phi 4 Mini 3.8B	×	microsoft/Phi-4-mini-instruct
Qwen 3 4B	✓	Qwen/Qwen3-4B
Gemma 3 4B	×	google/gemma-3-4b-it
LLaMA 3.1 8B	×	meta-llama/Llama-3.1-8B-Instruct
Qwen 3 8B	✓	Qwen/Qwen3-8B
Phi 4 14B	×	microsoft/phi-4
Qwen 2.5 14B	×	Qwen/Qwen2.5-14B-Instruct
Qwen 3 14B	✓	Qwen/Qwen3-14B
Qwen 3 32B	✓	Qwen/Qwen3-32B
LLaMA 3.1 70B	×	meta-llama/Llama-3.1-70B-Instruct
LLaMA 3.3 70B	×	meta-llama/Llama-3.3-70B-Instruct
Qwen 2.5 72B	×	Qwen/Qwen2.5-72B-Instruct

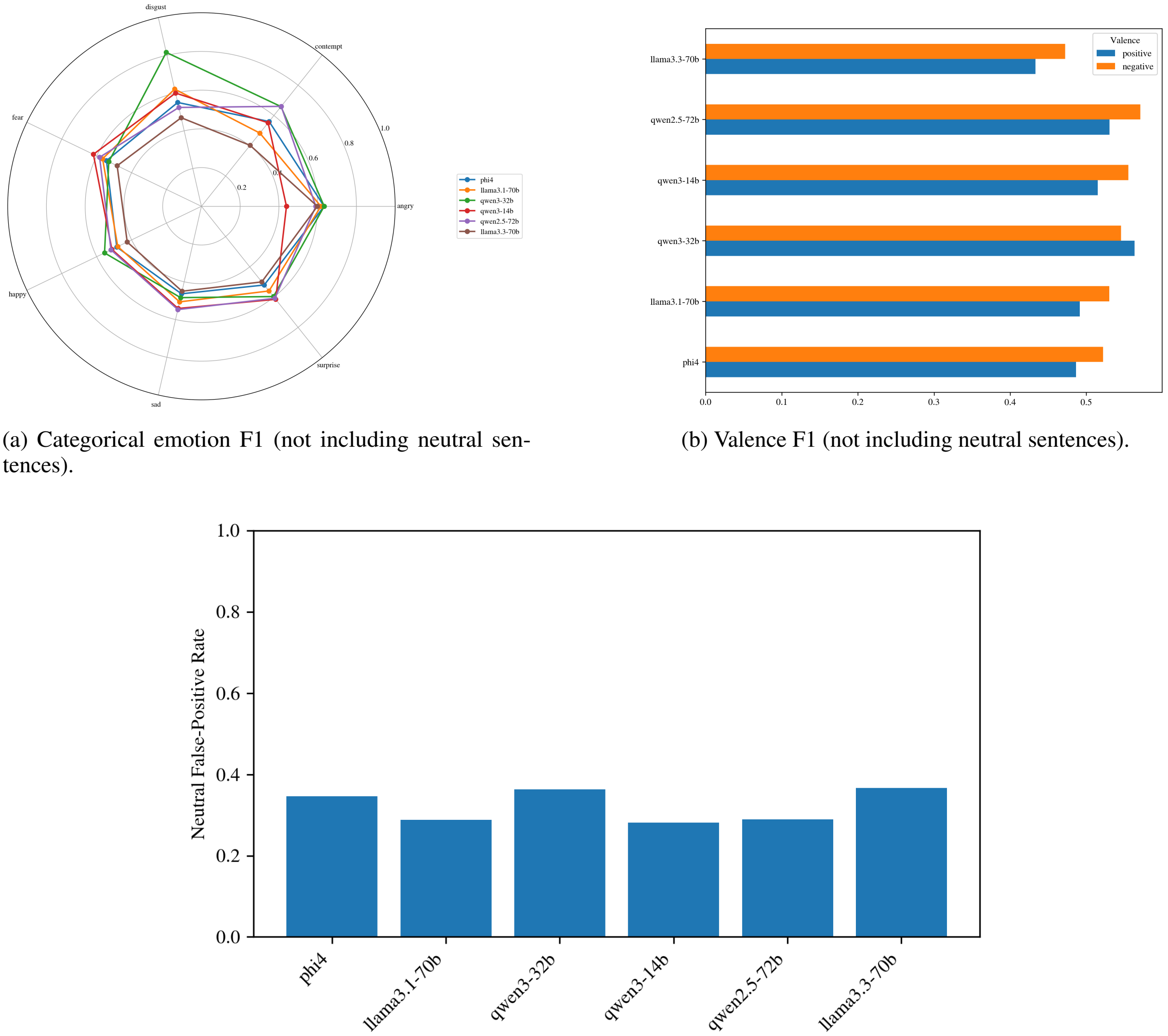


(a) Categorical emotion F1 (not including neutral sentences).

(b) Valence F1 (not including neutral sentences).

(c) Neutral False Positive Rate (NFPR): proportion of predicted spans in gold-labeled neutral sentences.

Figure 5: Emotion category errors in Task 2 Retrieve-Base.



(a) Categorical emotion F1 (not including neutral sentences).

(b) Valence F1 (not including neutral sentences).

(c) Neutral False Positive Rate (NFPR): proportion of predicted spans in gold-labeled neutral sentences.

Figure 6: Emotion category errors in Task 2 Highlight-Base.