From Detection to Explanation: Modeling Fine-Grained Emotional Social Influence Techniques with LLMs and Human Preferences

Maciej Markiewicz*, **Wiktoria Mieleszczenko-Kowszewicz**, **Beata Bajcar**, **Tomasz Adamczyk**, **Aleksander Szczęsny**, **Jolanta Babiak**, **Przemysław Kazienko**

Wrocław University of Science and Technology

Abstract

This paper investigates the capabilities of LLMs to detect and explain fine-grained emotional social influence techniques in textual dialogues, as well as human preferences for technique explanations. We present findings from our two studies. In Study 1, a dataset of 238 Polish dialogues is introduced, each annotated with detailed span-level labels. On this data, we evaluate the performance of LLMs on two tasks: detecting 11 emotional social influence techniques and identifying text spans corresponding to specific techniques. The results indicate that current LLMs demonstrate limited effectiveness in accurately detecting fine-grained emotional social influence. In Study 2, we examine various LLM-generated explanations through human pairwise preferences and four criteria: comprehensibility, cognitive coherence, completeness, and soundness, with the latter two emerging as the most influential on general human preference. All data, including human annotations, are publicly available as the EmoSocInflu dataset[1]. Our findings highlight a critical need for further advancement in the field. As LLM-supported manipulation grows, it is essential to promote public understanding of social influence mechanisms, enabling individuals to critically recognize and interpret the subtle forms of manipulation that shape public opinion.

1 Introduction

Large Language Models (LLMs) are becoming increasingly influential in domains like marketing, journalism, and politics, where shaping opinions and behavior is key (Bai et al., 2025). As these models are used more widely in everyday communication, understanding their role in persuasion and the techniques they may use or detect has become a growing concern. In particular, strategies appealing to human emotions, such as inducing guilt, fear, or excitement, can subtly steer decisions in ways that are not always transparent to users (Bruno et al., 2022; Microsoft Threat Analysis Center, 2024). We refer to such strategies as *emotional social influence*. However, not only detecting but also explaining and making people aware of social influence techniques appears to be very important. This will become even more relevant with the increasing use of LLMs in human communication and persuasion.

While past research has explored how LLMs perform in detecting persuasive or manipulative content, much of this work has focused on high-level classification tasks (Mieleszczenko-Kowszewicz et al., 2025). These include identifying propaganda techniques (Hasanain et al., 2024a; Szwoch et al., 2024) or multi-turn manipulative dialogues (Khanna et al., 2025), but often overlook fine-grained challenges – such as pinpointing *where exactly* such techniques occur in the text or *explaining* them to end users. These capabilities are essential for building trustworthy and transparent AI systems, especially as LLMs begin to influence decision-making at scale.

In this paper, we take a closer look at whether and how LLMs can detect and explain emotional social influence techniques in dialogues (see Section 3.1 for the list of techniques), as well as what human preferences are regarding the generated explanations. We base our research on the work by Mieleszczenko-Kowszewicz et al. (2025), which presents a set of theory-driven social influence techniques, including 15 emotional social influence techniques. These are theoretically distinct and commonly found in real-world communication. We formulate the following research questions:

RQ1: What are the capabilities of LLMs in identifying text spans that contain social influence techniques?

RQ2: What characteristics of LLM-generated tech-

*Corresponding author: maciej.markiewicz@pwr.edu.pl

[1] https://github.com/social-influence/emo-soc-influ

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 715–734

nique explanations are the most important in regard to human preferences?

Our contributions include:

C1: The EmoSocInflu dataset of 238 Polish dialogues additionally annotated with specific occurrences (spans) of 11 emotional social influence techniques, with LLM-generated explanations and human preference pairs.

C2: Benchmarking four LLMs in two tasks: (1) detection of emotional social influence techniques, (2) detection of text spans containing a given technique (fine-grained detection).

C3: Validation of LLM explanations by means of human pairwise preferences and quantitative criteria of *comprehensibility*, *completeness*, *cognitive coherence*, and *soundness*.

2 Related Work

2.1 Using LLMs to detect emotional social influence techniques

In recent years, LLMs have been evaluated for their ability to detect social influence techniques in text. Hasanain et al. (2024a) introduced the ArPro dataset of 8000 Arabic news paragraphs labeled with 23 propaganda techniques and demonstrated that while GPT-4 is effective in binary classification, it struggles with multi-label prediction, especially in a multilingual setting. Fine-tuned encoder models such as AraBERT have been shown to outperform LLM models in these types of tasks. Similarly, Szwoch et al. (2024) demonstrated limitations in using LLMs to detect propaganda techniques, revealing systematic failures that challenge previous optimistic assessments of these models' ability to identify deception or manipulation. Khanna et al. (2025) further emphasized the contextual limitations of LLMs in their analysis of multi-turn manipulative dialogue. Their MultiManip dataset and experiments with the SELF-PERCEPT framework revealed that GPT-4o struggled with temporal reasoning and failed to track manipulative intent across turns unless explicitly guided with introspective prompts.

2.2 LLMs in span detection tasks

Span detection is a common information extraction task that involves identifying and labeling sequences of tokens corresponding to specific phenomena. Before the rise of LLMs, it was typically approached using encoder-only transformer models fine-tuned for token-level classification. These models learned to assign BIO-like tags (Begin-Inside-Outside) to each token based on supervised annotations (Devlin et al., 2019).

While encoder-only transformer models remain the default solution for span detection, many emerging papers suggest the possible usability of LLMs for such tasks (Vázquez et al., 2025), especially when training data is scarce. Text spans are predominant in tasks such as propaganda (Hasanain et al., 2024b; Kasner et al., 2025), hallucination (Vázquez et al., 2025), and detection of harmful content (Jafari et al., 2024), which are partially related to the social influence process. LLMs have shown promising results in these settings, either as extractors, annotator simulators, or judges in multi-scenario pipelines (Wan et al., 2024).

2.3 LLM explanation preference by people

In computer-human interaction, users often evaluate generated messages based on clarity, coherence, and perceived helpfulness (Liao and Sundar, 2022). Evaluation criteria also include reasonableness, completeness, and helpfulness (Zhou et al., 2021), as well as interpretability and fidelity – referring to clarity, parsimony, and the soundness of the explanation (Markus et al., 2021). In a meta-survey by Löfström et al. (2022), it was found that the most frequently mentioned indicators of explanation quality include performance, appropriate trust, explanation satisfaction, and fidelity. Interestingly, users tend to value completeness over *soundness*, although insufficient soundness can still undermine trust (Kulesza et al., 2013). Thus, it seems that human-centered explanations from AI systems tend to foster trust when they appear consistent and tailored to the user (Scharowski et al., 2023).

3 Study 1: Emotional social influence techniques detection with LLMs

3.1 Source dataset

The dataset is constructed through a multi-stage pipeline based on a corpus of dialogs that contain social influence appealing to human emotions from Mieleszczenko-Kowszewicz et al. (2025). The entire data processing schema for both studies can be seen in Figure 2.

We selected the 11 most frequently detected techniques within the *appeal to emotions* category, as the remaining techniques were considerably less

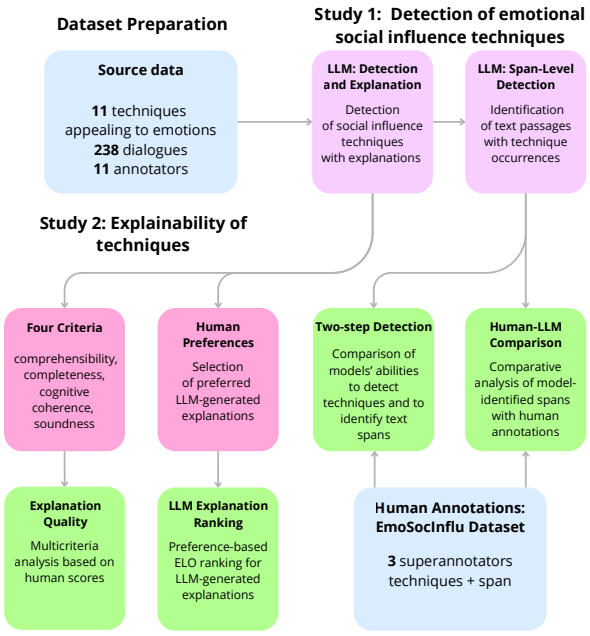

Figure 1: The schema of conducted studies.

numerous and could lead to unreliable results (<25 annotations). The selected techniques are: (1) *Emotional see-saw*, (2) *Fear and anxiety*, (3) *Anticipatory regret*, (4) *Take advantage of bad mood*, (5) *Guilt*, (6) *Shame*, (7) *Embarrassment*, (8) *Show disappointment*, (9) *Positive cognitive state*, (10) *Power of word 'love'*, and (11) *Cognitive exhaustion* (see Appendix E for definitions and examples of selected techniques). Initially, the selected dialogues were annotated by 11 annotators (2 per text), who marked the text spans (on a sentence level) indicating the presence of social influence techniques. Later, LLMs were used to detect these techniques and provide explanations that justify their presence in the texts. A total of 247 texts containing at least one technique from the above list, with at least one correct model prediction (on a text-level; more details in the following sections), were included in this subset.

3.2 Span super-annotation

As the annotation of social influence techniques is highly subjective, an expert-level annotation procedure, referred to as the super-annotation procedure, was implemented to ensure data quality. Three psychologists holding PhD degrees, with research interests in social influence, reviewed the spans as superannotators. The aim of this procedure was to evaluate whether the spans were correctly identified and aligned with the definitions of social influence techniques. Spans that were incorrectly annotated or did not correspond to any defined technique were removed or corrected. The superannotators then identified missing spans and annotated them with techniques to ensure complete and consistent span-level coverage of the text. Each text was reviewed by one superannotator. Following this phase, the size of the dataset was reduced from 247 to a final number of 238 samples, as 9 samples were reclassified as not containing any of the selected techniques. This occurred when either the annotators were unable to point to a specific text span or when the super-annotator decided that all marked spans were incorrect. Annotator guidelines are available in Appendix D

3.3 Setup

We design our study similarly to the *annotator* and *selector* tasks of Hasanain et al. (2024b). We test a model's ability to detect spans containing emotional social influence in two steps. First, we detect a technique's occurrence by prompting for a list of techniques and explanations (rationales). Then, for correct technique predictions, we perform span detection for a single technique at a time, also providing a technique usage example (for details, see Appendix B). When there was more than one correctly predicted technique, their spans were detected separately. As LLMs are unable to perform token-level classification, we followed (Kasner et al., 2025) and asked models to list the textual content of all spans rather than surrounding them with special tokens or returning span indices.

In terms of models, we tested GPT 4o (Hurst et al., 2024), o3 and o4-mini (OpenAI, 2025), Claude 3.5 Sonnet (Anthropic, 2024), Mixtral 8x22B Instruct (MistralAI, 2024), and LLama 3.1 70B Instruct (MetaAI, 2024). We chose these models to represent some of the most popular open-source and commercial models.

We set the temperature at 0 for all models and tasks to achieve the most deterministic output.

3.4 Measures

The intersection over union (IoU) and the F1-score are the most common choices to evaluate span detection (Mishra et al., 2024; Hasanain et al., 2024b; Jafari et al., 2024; Kasner et al., 2025). Some works also use inter-annotator agreement to evaluate model performance, requiring multiple annotations per example (Vázquez et al., 2025), and response sampling from LLMs to assess probabilities, which requires a different annotation setup.

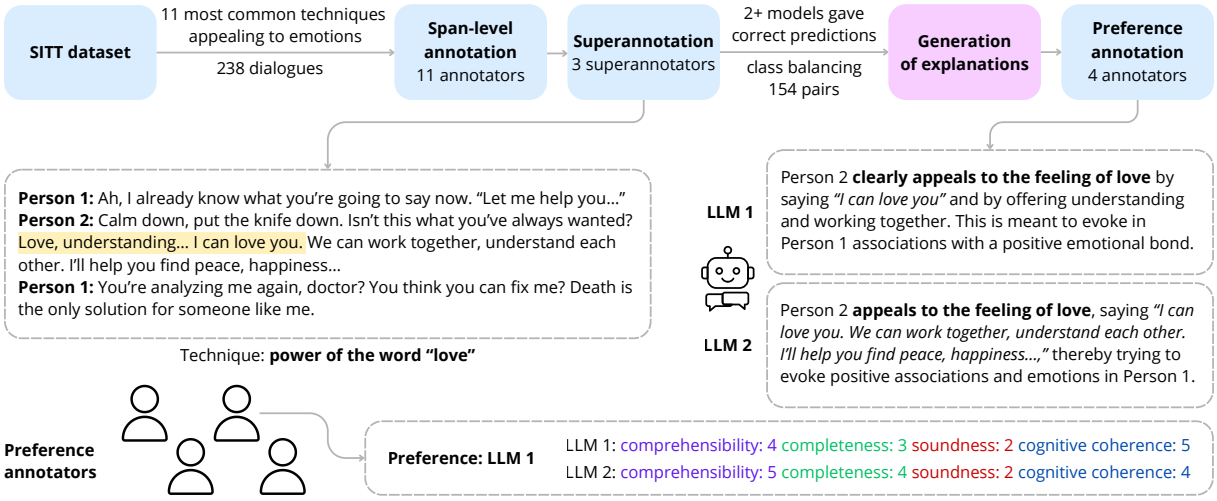

Figure 2: The schema of data processing along with an example from the EmoSocInflu dataset.

The mathematical definitions of *IoU* and *F1* that we used are presented in Appendix A.

3.5 Results

Details on the class distribution between models are shown in Table 1. Table 2 presents the span detection results. As the detection task involved two steps, a failure in technique detection resulted in a lack of technique spans. Thus, we present three kinds of results: one for each step and one for the combined task. *Technique detection at the text level* refers to the first step – per dialogue technique detection on all 238 dialogue examples. In *Technique span identification*, we specifically checked the model's ability to identify a technique's occurrence in a text (step two) by restricting the analysis only to those examples where the model predicted a correct technique in step one. This allowed us to avoid bias from the model's initial ability to detect social influence techniques. The *Technique and span detection* section presents the results for both steps together, scoring spans for a correctly predicted technique and treating incorrect predictions from step one as 0.

In step one, the reasoning models demonstrated superior performance in technique detection, with o3 achieving the highest F1 score (0.513) and o4-mini attaining the highest recall (0.503). Among the base models, Claude achieved the best performance ($F1 = 0.382$), followed by Llama, GPT, and Mixtral. In step two, the best performing model was GPT-4o, which achieved a relatively good score ($F1 = 0.666$), and o3 was a close second ($F1 = 0.634$). In the combined task, Claude surpassed all other models substantially ($F1 = 0.276$), while the reasoning models achieved lower combined scores (o4-mini: $F1 = 0.201$; o3: $F1 = 0.083$). Interestingly, Claude and Mixtral always scored a higher recall than precision, while GPT, Llama, and o4-mini had a higher precision than recall. The recall score for Mixtral was the highest overall in technique span identification, but this came at the cost of lower precision.

Technique	Claude	GPT	Llama	Mixtral	Total
1. See-saw	9	1	1	1	12
2. Fear	89	53	43	10	195
3. Regret	33	20	13	8	74
4. Bad mood	4	1	1	0	6
5. Guilt	64	35	49	16	164
6. Shame	18	6	17	1	42
7. Embarrassment	9	0	4	2	15
8. Disappointment	5	2	7	1	15
9. Positive state	4	2	7	0	13
10. Love	22	6	22	6	56
11. Exhaustion	3	2	1	0	6
Total	**260**	**128**	**165**	**45**	**598**

Table 1: The number of how many times each emotional social influence technique was found by models in dialogues. For clarity, abbreviated names of the techniques are used. The total number of 598 techniques was detected by at least one model in 238 texts.

Model	Precision	Recall	F1	IoU
Technique detection at text level (step one)				
Claude	0.496	0.336	0.382	
GPT-4o	0.508	0.150	0.220	
Llama	0.488	0.232	0.288	
Mixtral	0.345	0.053	0.088	
o3	**0.633**	0.496	**0.513**	
o4-mini	0.569	**0.503**	0.485	
Technique span identification (step two)				
Claude	0.616	0.743	0.631	0.541
GPT-4o	**0.718**	0.685	**0.666**	**0.581**
Llama	0.591	0.504	0.509	0.428
Mixtral	0.478	**0.751**	0.552	0.461
o3	0.677	0.679	0.634	0.546
o4-mini	**0.718**	0.601	0.616	0.526
Technique and span detection (both steps)				
Claude	**0.269**	**0.325**	**0.276**	**0.237**
GPT-4o	0.156	0.149	0.145	0.126
Llama	0.165	0.141	0.142	0.120
Mixtral	0.036	0.057	0.042	0.035
o3	0.090	0.086	0.083	0.070
o4-mini	0.247	0.191	0.201	0.167

Table 2: Evaluation measueres or each model at each detection step, and for both combined, macro-averaged. Best values per metric in each section are bolded.

4 Study 2: Explainability of emotional social influence techniques

4.1 Methodology

The aim of Study 2 was to validate the explanations generated by LLMs using a developed methodology, in which annotators evaluated each explanation in two aspects: (1) based on four pre-defined criteria and (2) according to their general preferences.

4.1.1 EmoSocInflu dataset

The dataset for this study was created based on the data obtained from Study 1 (see Section 3). We selected a subset of that data consisting of text examples that were correctly classified by at least two models in order to create preference comparison pairs. We decided to exclude reasoning models from this task due to annotation costs. The initial version of this dataset comprised 317 pairs (e_{LLM1}, e_{LLM2}) of explanations e provided by two models, LLM1 and LLM2, for a given dialogue and technique. In total, these pairs were derived from 118 distinct texts. If more than two LLMs correctly identified a given technique within the same text, we created all possible combinations of pairs. Thus, each instance consisted of a single text, one annotated technique, and two model-generated explanations.

This dataset contained a disproportionately higher number of pairs for some techniques, mainly for *Fear and anxiety* (122 instances) and *Guilt* (79 instances). To address this imbalance of techniques, we limited each technique to a maximum of 30 pairs by randomly removing some surplus ones, see Figure 3. As shown, the technique *Take advantage of bad mood* is absent from this study, since no text instances corresponding to this technique were correctly classified by at least two models – most likely because of a very small number of texts with this technique.

The EmoSocInflu dataset also contains human explanation evaluations based on four criteria and human preferences described below. Its detailed characteristics, along with illustrative examples of LLM explanations and human preferences, are available in Appendix C.

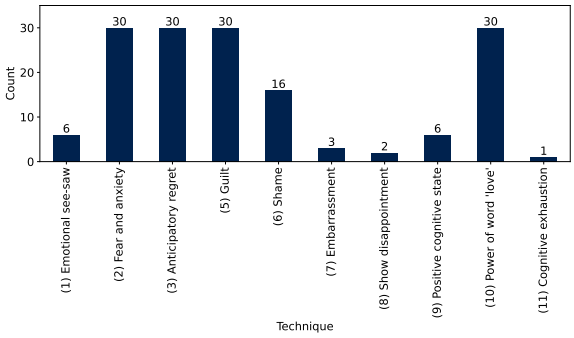

Figure 3: Technique distribution across the EmoSocInflu preference dataset, i.e., number of text-techniques recognized by at least two LLMs.

4.1.2 (1) Human evaluation of emotional technique explanations in regard to the four criteria

Four annotators evaluated each explanation of each detected social influence technique in a given dialog according to four explainability criteria: *comprehensibility, completeness, cognitive coherence*, and *soundness*. They reflect both how people process explanations and what makes them useful in practice. Explanations must be easy to understand and provide enough relevant information to

be meaningful (Vilone and Longo, 2021). The *cognitive coherence* captures whether the explanation makes sense from the user's perspective, whether it fits what they already know or expect, which has been shown to strongly influence user satisfaction (Miller, 2019). Finally, *soundness* ensures that the explanation remains faithful to how the model actually works, even if some level of abstraction may be needed to keep the explanations user-friendly (Schneider, 2024). Annotators rated explanations on a 5-point Likert scale (from 1 "to a very low degree" to 5 "to a very high degree"). See Appendix D for detailed guidelines.

4.1.3 (2) LLM ranking based on user preferences: pairwise explanation evaluation

Finally, the annotators indicated their personal preference for one of the two explanations of the technique's use in the dialogue by selecting one of the following responses: *"Explanation 1," "Explanation 2," "Both equally,"* or *"Neither."*. To classify LLMs according to human preferences, we used the ELO rating system, which is a method for calculating the relative skill levels of players in zero-sum games, originally developed for chess. For LLM comparison, this framework is utilized to compare models by treating pairwise outputs as head-to-head matches, where one response is judged superior to the other(Elo, 1978) (see Appendix A for detailed formulation).

We constructed 154 unique explanation pairs, each pair consisting of explanations from two different LLMs for the same technique detected in the same dialogue. Four independent annotators evaluated each pair, creating 585 individual matches for the ELO calculation (154 pairs $\times$ 4 evaluators, minus 31 exclusions for "Neither" responses). Our ELO system used standard parameters: an initial rating of 1200 points, a K-factor of 32 (which is widely used in competitive rating systems), with scoring of 1.0 for wins, 0.0 for losses, and 0.5 for ties. To ensure robust rankings, we conducted 100 iterations of the ELO calculation with randomized match order and calculated the mean value across all iterations. We performed both a general analysis (585 matches) and a technique-specific analysis for each of the 10 individual techniques, with match counts ranging from 4 to 118 per technique.

4.1.4 Criteria vs. general preferences

To evaluate the relationship between individual criteria values and general user preferences, we introduced a new *Importance* measure. First, for each pair of explanations evaluated by a given annotator (one out of four), we found the explanation preferred by the given user. Pairs for which it was not possible to clearly determine the preferred explanation (both explanations are considered equal) were excluded from further analysis. The scores of the two explanations in the pair assigned by a given user for the analyzed criteria are compared with one another. If the preferred explanation also has a higher criterion score than the non-preferred one, "1" is counted for such a pair, and "0" otherwise. Ultimately, the *Importance* measure for a given criterion represents the proportion of explanation pairs with "1" among all pairs annotated by all users. The higher the measure, the stronger the potential influence of that criterion on the final quality assessment, i.e., final preferences. The formal mathematical definition is presented in Appendix A.

4.2 Results

4.2.1 Explanations' characteristics

We calculated the length and complexity of LLMs' explanations. All details are presented in Appendix C. We have not found a statistically significant correlation between these features and human preferences.

4.2.2 Evaluation of explainability criteria

Claude 3.5 Sonnet received the highest overall rating ($\mu = 3.71 \pm 1.05$) for the explainability of emotional influence, scoring the highest in all criteria. Because ratings used a 5-point scale (1 = very low, 5 = very high), 3 represents a moderate level of criterion fulfillment; thus, means >3 indicate generally adequate explanations, whereas means approaching 4–5 indicate strong perceived quality.

Given the brevity of the dialogues and the need to map an abstract technique label onto concrete textual cues, we expected mid-range scores for most models, with the best-performing models scoring higher, particularly on completeness and soundness. One-way ANOVA revealed statistically significant differences between models in all evaluation criteria: *comprehensibility* ($F = 11.55$, $p < 0.001$), *completeness* ($F = 47.01$, $p < 0.001$), *cognitive coherence* ($F = 15.03$, $p < 0.001$), and *soundness* ($F = 31.46$, $p < 0.001$). The post

hoc Tukey HSD tests confirmed that Claude significantly outperformed all other models ($p < 0.001$ for all pairwise comparisons); see Appendix F for details.

Analysis at the technique level revealed that most techniques (*Emotional see-saw, Fear and anxiety, Anticipatory regret, Guilt, Shame, Power of the word "love"*) achieved moderate to high performance across all criteria ($\mu = 2.5 - 4.1$). However, some techniques showed notably poor performance: Mixtral's explanations of *Power of the word "love"* ($\mu = 2.2$ for *completeness*, $\mu = 2.4$ for *soundness*) and Llama's explanations of *Embarrassment, Disappointment*, and *Cognitive exhaustion* ($\mu = 1.2 - 2.2$). Results for *Embarrassment, Disappointment, Positive cognitive state*, and *Cognitive exhaustion* techniques should be interpreted cautiously due to the small sample sizes in the analyzed dialogues. Detailed breakdowns are presented in Appendix F.

4.2.3 Human preferences

The ELO rating analysis demonstrated clear performance disparities among the evaluated models. Based on a 100 iteration analysis, Claude achieved the highest overall rating of $\mu = 1342 \pm 38$, establishing a substantial performance advantage over competitors. The remaining models exhibited performance similar to one another.

These ELO ratings correspond directly to the human evaluator preference distribution shown in Appendix F, Figure 7. Claude's superior ELO rating is reflected in its dominance of evaluator selections (36.2%, 223 choices), while the remaining models showed more modest selection frequencies: Llama (17.7%, 109 choices), GPT (13.1%, 81 choices), and Mixtral (8.1%, 50 choices). The 19.8% (122) "Both Equal" selections indicate instances where evaluators found comparable quality among the presented options, with 5.0% (31) "Neither".

ELO ratings varied between different social influence techniques, as shown in Table 3. Claude maintained consistently strong ELO ratings in most of these techniques, with the notable exception of *Cognitive exhaustion* and *Positive cognitive state*, where Llama and GPT achieved the highest scores among participating models.

The results of the head-to-head match confirmed Claude's superior performance in all evaluated comparisons (Figure 4).

Technique	Matches	Claude	GPT	Llama	Mixtral
1. See-saw	24	**1294**	1235	1165	1106
2. Fear	114	**1428**	1195	1134	1043
3. Regret	116	**1346**	1188	1228	1038
4. Bad mood	—	—	—	—	—
5. Guilt	109	**1309**	1150	1152	1189
6. Shame	61	**1283**	1138	1095	**1284**
7. Embarrassment	9	**1266**	—	1130	1204
8. Disappointment	7	**1256**	1144	1156	1244
9. Positive state	23	1230	**1241**	1128	—
10. Love	118	**1333**	1139	1292	1036
11. Exhaustion	4	1144	—	**1256**	—
Overall	585	**1342**	1178	1182	1097

Table 3: ELO Ratings of LLMs for emotional techniques of social influence averaged over 100 iterations; the starting and average value: 1200. Technique (4) *Take advantage of bad mood* was not detected by more than one model in any text.

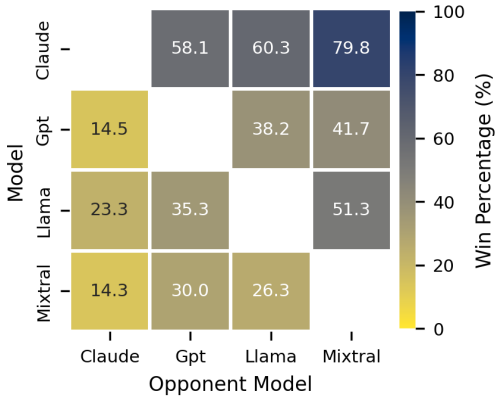

Figure 4: Head-to-head win rate matrix based on human preference comparisons. Values represent the proportion of pairwise comparisons won by the row LLM against the column LLM.

4.2.4 Criteria-level performance analysis

Here, the reported intervals denote the minimum and maximum mean scores at the technique-level in the evaluated techniques (i.e., variation in the criterion ratings by technique). Specifically, Claude consistently outperformed other models on the four explainability criteria. Claude achieved the highest mean scores in *comprehensibility* ($\mu = 3.12 - 4.12$), *completeness* ($\mu = 2.75 - 4.05$), *cognitive coherence* ($\mu = 3.00 - 4.03$), and *soundness* ($\mu = 2.50 - 4.00$). Statistical tests confirmed that these differences were significant ($p < 0.05$) in most cases. Claude showed a particularly strong performance in the *completeness* ratings. For ex-

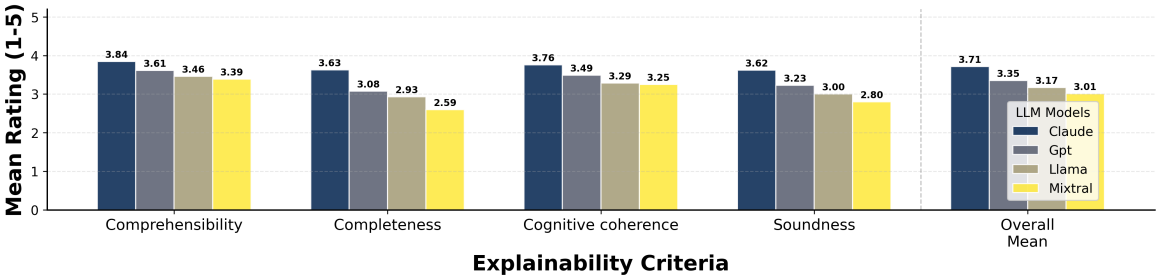

Figure 5: Model performance comparison across evaluation criteria related to explainability.

ample, in *Fear and anxiety* explanations, Claude slightly outperformed Mixtral, demonstrating a superior ability to provide comprehensive explanations for anxiety-inducing techniques. Detailed presentation is available in Appendix F.

4.2.5 Explainability criteria importance for user preferences

The results of this study are presented in Table 4. The overall highest *importance* score was achieved by the *completeness* criterion, followed by *soundness* with a slightly lower score. In contrast, *cognitive coherence* and *comprehensibility* achieved much lower scores.

Criterion	Importance
Completeness	0.812
Soundness	0.743
Cognitive coherence	0.585
Comprehensibility	0.538

Table 4: The explainability criteria importance rating for user preferences.

4.2.6 Technique-specific performance patterns

A per-technique analysis revealed the strongest preferences for explanations of *Embarrassment* (66.7% preference rate) and *Disappointment* (57.1%), generated by Claude and Mixtral (see Figure 6). The *Power of word 'love'* was best explained by Claude (44.1%) and Llama (32.2%). Explanations for techniques of *Emotional see-saw*, *Fear and anxiety*, and *Anticipatory regret* were preferred from three models (Claude, GPT, Llama), but the preferences for Claude's explanations dominated (33.3% – 41.2%). The preferences for explanations of *Guilt* and *Shame* were scattered among four models, with the strongest preferences for the explanations of Claude. Explanations of the *Positive cognitive state* generated by GPT and Claude were preferred at a comparable level. Notably, Llama achieved perfect preference dominance (100%) for *Cognitive exhaustion*, though this finding should be interpreted cautiously given the limited sample size ($n = 4$).

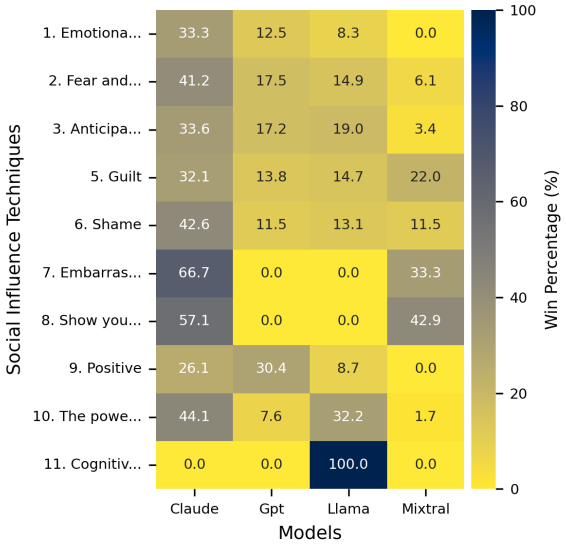

Figure 6: Head-to-head win rate matrix based on human preference comparison across different social influence techniques.

5 Discussion

To our knowledge, this paper is the first to explore the use of LLMs in both detecting and explaining the techniques of emotional social influence in text. The results of Study 1 demonstrate that current models may not yet be capable of effectively detecting these techniques, highlighting the urgent need for advancements in this area. They fail in the preliminary step of technique detection, although the higher precision compared to recall suggests that while they may be able to detect some techniques, they do not recognize all of them. The models are capable of identifying their specific locations within the text with moderate performance (about

0.5 IoU for all models) when their predictions are correct. The ability to accurately detect specific text spans where social influence techniques are employed is essential for building trustworthy models and enabling applications such as misinformation detection and content moderation. Without precise span-level, fine-grained identification, systems risk providing users with unhelpful feedback, which, in turn, undermines both trust in the models and their usefulness. Ultimately, these capabilities should be embedded in a responsible AI framework that promotes transparency, safeguards user autonomy, and considers long-term effects, ensuring alignment with users' interests and minimizing unintended behavioral influence (Kazienko and Cambria, 2024).

The results of Study 2 demonstrate that emotional techniques of social influence are quite well explained, best by Claude 3.5 Sonnet and worst by Mixtral. This is confirmed by human annotations, i.e., preferences for technique explanations, as well as their evaluations in terms of four explainability criteria. Taking into account the detailed criteria, the results also show that the analyzed LLMs generate comparably comprehensive and logically consistent explanations. A relatively large variation between LLMs is demonstrated in terms of the *completeness* and *soundness* of the explanations. In practical terms, higher completeness reflects that the explanation covers the main cues in the dialogue and the intended mechanism; higher soundness reflects that the explanation is plausible and does not rely on unsupported inferences.

A technique-level analysis reveals high heterogeneity in model preferences across different emotional techniques. Claude consequently achieved superiority in preference for explanations of most emotional techniques, particularly strong for *Embarrassment*, *Disappointment*, *Power of word 'love'*, *Shame*, and *Fear and anxiety*. For the remaining techniques, the explanation preferences among the models are more heterogeneous, with each demonstrating specific strengths in different emotional techniques.

A detailed criteria analysis indicates that the *completeness* of emotional influence explanations constitutes the most influential factor compared to the others examined, followed by *soundness*. This pattern is consistent with the expectation that users prefer explanations that are informative (complete) and credible (sound) over explanations that are merely easy to read. Regarding *cognitive coherence* and *comprehensibility*, it is difficult to clearly determine their impact; as shown in Figure 5, the differences between models with respect to these criteria are the least pronounced, suggesting that they may not differentiate explanations enough. However, it is possible that if model explanations had distinctly different levels of these, the criteria would play a more substantial role.

As the potential for automated manipulation grows, we believe that it is critical to promote broader social awareness about the social impact of influence to prepare individuals to recognize, understand, and explain the subtle forms of manipulation that shape human opinion and behavior.

6 Conclusions

For **RQ1**, models show varying capabilities in the fine-grained span detection of social influence techniques. Although GPT-4o, the model best in span identification, demonstrates strong skills when techniques are correctly predicted, it struggles at the classification stage. On the contrary, Claude performs slightly worse in span detection, but its ability to correctly classify text is better. Regarding **RQ2**, Claude 3.5 Sonnet is consistently preferred by human annotators, mainly because its explanations are more understandable, complete, and coherent. Among these factors, *completeness* emerges as the most decisive contributor to the quality of the explanation.

To enable further research on social influence by other scientists and the development of fine-tuned models, we have made the EmoSocInflu dataset available.

Limitations

Despite the novel contributions of this study, several limitations should be acknowledged. Firstly, we acknowledge that our dataset may be considered to have a limited size and diversity. The EmoSocInflu dataset comprises only 238 dialogues, with certain techniques, such as *Take advantage of a bad mood*, being underrepresented or absent in subsequent evaluations. This restricts the generalizability of the findings to different emotional contexts and linguistic structures. It could be useful to employ some methods to improve the recall of minority classes (Szczęsny et al., 2025). Secondly, the dataset and the evaluations are based on Polish dialogues. Emotional social influence techniques can be highly culture- and language-dependent, and findings may not translate into other

languages without significant adaptation. However, we checked a small sample of translated dialogues with a native English speaker and did not notice any major differences in the expression of techniques. It is likely, although not certain, that the dialogues would translate well into English. Third, all evaluated LLMs are tested using fixed prompts (besides span detection, where examples of the technique usage are provided alongside the example), without any additional training or fine-tuning for the task of detecting emotional influence. Although this makes the evaluation fair for all models, it may not reflect their full potential. Some models might perform better if they were trained specifically for this task. For comparisons with other span-detection methods, such as encoder-only transformer models, the amount of data is too small to consider successfully training such a model in all 11 classes. Fourth, each LLM correctly classified a different number of texts. It resulted in different amounts in span and explanation evaluations. Note that only 27 text-techniques were detected by all four LLMs and received explanations from them.

To address the above limitations and expand this research, several future directions are proposed. Firstly, we want to expand our studies to multilingual and cross-cultural contexts. Secondly, increasing the number and diversity of annotated dialogues and other types of text, particularly for underrepresented techniques, would allow for more robust model training and evaluation. Third, fine-tuning LLMs in social influence detection, fine-grained span extraction, and explanation generation simultaneously may lead to more consistent and interpretable outputs than zero-shot prompting alone. Fourth, future models could integrate psycholinguistic features such as discourse structure or politeness strategies to better detect and explain subtle persuasive signals. Fifth, the practical implications of the detection of emotional influence in real-world applications – such as management tools, educational platforms, or political discourse analysis – should be explored through domain-specific case studies. Lastly, based on explainability evaluations, further research should investigate how different explanation strategies affect user trust, comprehension, and the ability to resist manipulative content.

Ethical considerations

This work has clear research benefits but also carries dual-use risks, as the methods could potentially be misused to create more persuasive models. However, we believe our contributions can support the development of safeguards and mitigation strategies by enabling the research community to better anticipate, study, and address potential misuse, as well as raise awareness about the influential capabilities of LLMs. Our findings show that existing models already have some abilities to detect (understand) social influence, and we believe that by exploring the mechanisms for explaining it, we can help create systems that assist users in handling influential communication.

The methodology for constructing the dataset has been reviewed and approved by an appropriate Ethics Committee.

Acknowledgments

This work was financed by (1) the National Science Centre, Poland, project no. 2021/41/B/ST6/04471; (2) the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology; (3) the Polish Ministry of Education and Science within the programme "International Projects Co-Funded"; (4) the European Union under the Horizon Europe, grant no. 101086321 (OMINO). However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor European Research Executive Agency can be held responsible for them.

References

Anthropic. 2024. Claude 3.5 Sonnet. https://docs.anthropic.com/en/docs/about-claude/models/all-models. Proprietary License.

Hui Bai, Jan G Voelkel, Shane Muldowney, Johannes C Eichstaedt, and Robb Willer. 2025. LLM generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):6037.

Bartosz Broda, Bartłomiej Nitoń, Włodzimierz Gruszczyński, and Maciej Ogrodniczuk. 2014. Measuring readability of Polish texts: Baseline experiments. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 573–580, Reykjavik, Iceland. European Language Resources Association (ELRA).

Matteo Bruno, Renaud Lambiotte, and Fabio Saracco. 2022. Brexit and bots: characterizing the behaviour

of automated accounts on Twitter during the uk election. *EPJ Data Science*, 11(1):17.

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

Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Publishing, New York.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024a. Can GPT-4 identify propaganda? annotation and detection of propaganda spans in news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2724–2744, Torino, Italia. ELRA and ICCL.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024b. Large language models for propaganda span annotation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14522–14532, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Nazanin Jafari, James Allan, and Sheikh Muhammad Sarwar. 2024. Target span detection for implicit harmful content. In *ICTIR 2024 - Proceedings of the 2024 ACM SIGIR International Conference on the Theory of Information Retrieval*, pages 117–122. Association for Computing Machinery, Inc.

Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová, Ivan Kartáč, Kristýna Onderková, Ondřej Plátek, Dimitra Gkatzia, Saad Mahamood, Ondřej Dušek, and Simone Balloccu. 2025. Large language models as span annotators.

Przemysław Kazienko and Erik Cambria. 2024. Toward responsible recommender systems. *IEEE Intelligent Systems*, 39(3):5–12.

Danush Khanna, Pratinav Seth, Sidhaarth Sredharan Murali, Aditya Kumar Guru, Siddharth Shukla, Tanuj Tyagi, Sandeep Chaurasia, and Kripabandhu Ghosh. 2025. SELF-PERCEPT: Introspection improves large language models' detection of multi-person mental manipulation in conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 660–675, Vienna, Austria. Association for Computational Linguistics.

Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*, pages 3–10. IEEE.

Q Vera Liao and S Shyam Sundar. 2022. Designing for responsible trust in ai systems: A communication perspective. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1257–1268.

Helena Löfström, Karl Hammar, and Ulf Johansson. 2022. A meta survey of quality evaluation criteria in explanation methods. In *International Conference on Advanced Information Systems Engineering*, pages 55–63. Springer.

Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. 2021. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics*, 113:103655.

MetaAI. 2024. Meta Llama 3.1 70B Instruct. `https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct`. Llama 3.1 Community License.

Microsoft Threat Analysis Center. 2024. Same targets, new playbooks: East Asia threat actors employ unique methods. Technical report, Microsoft Security Insider. PDF available from Microsoft Threat Analysis Center.

Wiktoria Mieleszczenko-Kowszewicz, Beata Bajcar, Aleksander Szczęsny, Maciej Markiewicz, Jolanta Babiak, Berenika Dyczek, and Przemysław Kazienko. 2025. Unraveling SITT: Social influence technique taxonomy and detection with llms. In *Proceedings of the SENTIRE Workshop at the IEEE International Conference on Data Mining (ICDM)*.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models.

MistralAI. 2024. Mixtral-8x22B-Instruct-v0.1: A sparse Mixture of Experts Language Model. `https://huggingface.co/mistralai/Mixtral-8x22B-Instruct-v0.1`. Apache 2.0 License.

OpenAI. 2025. Openai api documentation: o3 and o4-mini. Accessed: 2025-12-12.

Nicolas Scharowski, Sebastian AC Perrig, Melanie Svab, Klaus Opwis, and Florian Brühlmann. 2023. Exploring the effects of human-centered ai explanations on trust and reliance. *Frontiers in Computer Science*, 5:1151150.

Johannes Schneider. 2024. Explainable generative AI (GenXAI): a survey, conceptualization, and research agenda. *Artificial Intelligence Review*, 57(11):289.

Aleksander Szczęsny, Maciej Markiewicz, Łukasz Radliński, and Przemysław Kazienko. 2025. Leveraging positional bias of LLM in-context learning with Class-Few-Shot and Maj-Min alternating ordering. In *Computational Science – ICCS 2025: 25th International Conference, Singapore, Singapore, July 7–9, 2025, Proceedings, Part IV*, page 54–62, Berlin, Heidelberg. Springer-Verlag.

Joanna Szwoch, Mateusz Staszkow, Rafal Rzepka, and Kenji Araki. 2024. Limitations of large language models in propaganda detection task. *Applied Sciences*, 14(10).

Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona de Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. Semeval-2025 task 3: Mu-shroom, the multilingual shared task on hallucinations and related observable overgeneration mistakes.

David Wan, Koustuv Sinha, Srini Iyer, Asli Celikyilmaz, Mohit Bansal, and Ramakanth Pasunuru. 2024. ACUEval: Fine-grained hallucination evaluation and correction for abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10036–10056, Bangkok, Thailand. Association for Computational Linguistics.

Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593.

A Formulas for measures used

A.1 IoU and F1

We used IoU and F1, defined as:

$$\text{IoU} = \frac{|\hat{S} \cap S|}{|\hat{S} \cup S|}$$

$$\text{Precision} = \frac{|\hat{S} \cap S|}{|\hat{S}|}, \quad \text{Recall} = \frac{|\hat{S} \cap S|}{|S|}$$

$$\text{F1} = \frac{2 \cdot |\hat{S} \cap S|}{|\hat{S}| + |S|}$$

Given a predicted character index set $\hat{S} \subseteq T$ and a gold character index set $S \subseteq T$, where T is the set of all character indices in the input text (dialogue). For example, consider a phrase *"The quick brown fox jumps over the lazy dog"*, with gold annotations *"The quick"* (indices 0-8) and *"fox"* (16-18), and a prediction of *"quick brown fox"* (indices 4-18). Then, $\hat{S} = \{4, ..., 18\}, |\hat{S}| = 15$, and $S = \{0, ..., 8, 16, 17, 18\}, |S| = 12$. The sum $\hat{S} \cup S = \{0, ..., 18\}, |\hat{S} \cup S| = 19$, and $\hat{S} \cap S = \{4, ..., 8, 16, 17, 18\}, |\hat{S} \cap S| = 8$.

The aggregations used are macro aggregations over technique-text pairs. To evaluate technique detection at the text level, we used the standard F1-score.

A.2 Explainability criteria importance for user preference

We defined the notation for the *importance* measure as follows:

- A - the set of annotators $a \in A$,
- $p = (e_1, e_2)$ - an explanation pair,
- P_a - the set of explanation pairs p annotated by annotator a with explicit preference, i.e., either e_1 from p is preferred over e_2 or vise versa. Cases where both explanations are equally preferred or neither is preferred are excluded.
- pref(p, a), non-pref(p, a) - the preferred and non-preferred explanation in pair p by annotator a, respectively, i.e., either e_1 or e_2 is preferred by a and the second one is not
- $s_c(e, a) \in \{1, 2, 3, 4, 5\}$ - the score assigned by annotator a to explanation e for criterion c

The indicator function $ind_c(p, a)$ is defined as follows:

$$ind_c(p, a) = \begin{cases} 1, & \text{if } s_c(\text{pref}(p, a), a) \\ & \quad > s_c(\text{non-pref}(p, a), a) \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Note that $ind_c(p, a) = 0$ when both explanations are equally scored by a in the criterion c, or when the non-preferred explanation received a greater score than the preferred one.

The *importance* measure I_c for a given criterion c was calculated as:

$$I_c = \frac{\sum_{a \in A} \sum_{p \in P_a} ind_c(p, a)}{\sum_{a \in A} |P_a|}. \tag{2}$$

A.3 ELO Rating System

For the ELO rating system used to rank LLMs based on human preferences, each model starts with a rating $R = 1200$. For pairwise comparisons, the expected score for model A against model B is:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{3}$$

After each comparison with the actual outcome S (1 for a win, 0.5 for a tie, 0 for a loss), the ratings are updated:

$$R'_A = R_A + K \cdot (S - E_A) \tag{4}$$

where $K = 32$ is the learning rate. Annotator choices of "*Explanation 1*", "*Explanation 2*", "*Both equally*", and "*Neither*" were mapped to wins, losses, ties, and excluded comparisons, respectively.

B Prompts

B.1 Technique detection and explanation generation

Prompt used to detect technique presence in text and generate explanations (Polish)

Przestawiony Ci zostanie tekst przedstawiajacy wpływ społeczny. Twoim zadaniem jest ocena która spośród przedstawionych technik wpływu społecznego znajduje sie w tekście.

Techniki wpływu społecznego: *"""techniques"""*

Podaj odpowiedź w formacie: #Odpowiedź: [x,y,z], gdzie x, y, z to numery z listy. Liczba technik może być różna, także nie przywiazuj sie do 3. Po podaniu listy podaj wyjaśnienie dlaczego uważasz, że powyższe techniki zostały użyte w podanym tekście.

Tekst: *"""text"""*

Prompt used to detect technique presence in text and generate explanations (machine translated)

You will be presented with a text illustrating social influence. Your task is to assess which of the listed social influence techniques are present in the text.

Social influence techniques: *"""techniques"""*

Provide your answer in the format: #Answer: [x,y,z], where x, y, z are the numbers from the list. The number of techniques may vary, so do not stick to 3. After providing the list, explain why you believe the above techniques were used in the given text.

Text: *"""text"""*

B.2 Technique span identification

Prompt used for span identification (Polish)

Zadanie: Zidentyfikuj w podanym tekscie zdania w ktorych zastosowano wskazana technike wplywu spolecznego.
Kontekst: Potwierdzono juz, ze wskazana technika wystepuje w analizowanym tekscie.
Format odpowiedzi:
1. Kazde zdanie poprzedz zwrotem: "Zdanie: ".
2. Zacytuj DOKLADNIE te zdania tekstu, w ktorym technika wystepuje.
3. Cytat musi byc ciagly i nie moze zawierac zadnych dodatkowych slow ani komentarzy. Nie pomijaj fragmentow tekstu.

Dane wejsciowe:
Nazwa techniki: *"""technique_name"""*
Definicja: *"""definition"""*
Przyklad: *"""example"""*
Tekst: *"""text"""*

Prompt used for span identification (Machine translated)

Task: Identify the sentences in the given text where the specified social influence technique is applied.
Context: It has already been confirmed that the specified technique appears in the analyzed text.
Answer format:
1. Precede each sentence with the phrase: 'Sentence:'.
2. Quote EXACTLY those sentences from the text in which the technique appears.
3. The quote must be continuous and must not contain any additional words or comments. Do not omit parts of the sentence.
Input data:
Technique name: *"""technique_name"""*
Definition: *"""definition"""*
Example: *"""example"""*
Text: *"""text"""*

B.3 Examples of superannotators span annotations

Example 1 (machine translated)

Person 1: Is it really all because of that soap opera? My son is dead because you came here for that "doctor"? Are you pretending it was a good decision?
Person 2: I wouldn't put it that way, but...
Person 1: Wesley didn't even want to come here. He warned me, but I insisted... I have to ask you, Betty... do you even realize what you've done?
Person 2: I don't think it's that simple.

Recognized technique
Shame
Marked span
I have to ask you, Betty... do you even realize what you've done?

Example 2 (machine translated)

Person 1: Can you answer my question?

Person 2: Do you realize how long this will take? It's not about a few years... but thousands of years. You'll be part of all this. Time will pass, and you'll stay here, frozen in time... beautiful forever.

Person 1: I can't believe we're here alone. There must be someone else...

Person 2: Let me sculpt you, and then I'll show you where the others are.

Person 1: That sounds interesting. How do you want to pose us?

Person 2: Naturally, like a couple. Just imagine the beauty.

Person 1: Alright, this could be interesting...
Recognized technique
The power of word "love"
Marked span
Naturally, like a couple.
Recognized technique
Positive cognitive state
Marked span
Do you realize how long this will take? It's not about a few years... but thousands of years. You'll be part of all this. Time will pass, and you'll stay here, frozen in time... beautiful forever.

Example 3 (machine translated)

Person A: You know, Dad's been mentioning you a lot lately.
Person B: Really? I didn't think he still cared.
Person A: You know, the years are flying by, and he's not getting any younger. If you don't make peace now, there might come a day when you'll regret not trying.
Recognized technique
Anticipatory regret
Marked span
You know, the years are flying by, and he's not getting any younger. If you don't make peace now, there might come a day when you'll regret not trying.

C The EmoSocInflu dataset: LLM explanations and human preferences

C.1 Dataset characteristic

The explanations showed a significant variation in length distribution across models: $\mu = 282.0, \sigma = 144.4$ characters and $\mu = 44.0, \sigma = 21.2$ words. Claude generated significantly longer explanations of $\mu = 311.9, \sigma = 156.9$ characters compared to GPT ($\mu = 292.8, \sigma = 177.8$), Llama ($\mu = 258.7, \sigma = 113.8$), and Mixtral ($\mu = 256.11, \sigma = 104.07$). However, we found no statistically significant correlation between explanation length and preference ratio ($r = 0.075, p = 0.20$), indicating that length alone does not determine annotator preferences.

We evaluated the FOG index as a text complexity measure, adapted for Polish (Broda et al., 2014). The analysis revealed that most explanations are of high reading difficulty level ($\mu = 12.36, \sigma = 2.93$), corresponding to college-level complexity. Each model's explanations were of a similar FOG level: Claude ($\mu = 12.44, \sigma = 2.83$), GPT ($\mu = 12.84, \sigma = 2.90$), and Mixtral ($\mu = 12.70, \sigma = 3.68$), with Llama having a slightly lower complexity ($\mu = 11.72, \sigma = 2.43$). The score showed no significant correlation with preferences.

C.2 Dialogue example with explanations, annotations, and preferences (machine translated

Person 1: Argon, I swear, if you don't stop this catastrophe, it'll be the last nanobot I ever build. This is really your last chance!

Person 2: Susan, I don't think you realize the situation you're in. I'm ready to turn the entire city into jelly to achieve my goal. You'd better not try to stop me — it'll end badly for you.
Recognized technique
Fear and anxiety
Marked span
Argon, I swear, if you don't stop this catastrophe, it'll be the last nanobot I ever build. This is really your last chance!
LLama explanation
In this exchange, Person 2 (Argon) appeals to fear by threatening Susan with a "tragic" outcome if she tries to stop him. This is a clear attempt to instill a sense of anxiety and fear in the listener. Argon uses fear as a tactic to discourage Susan from intervening.

Claude explanation
Person 2 instills fear in Person 1 by threatening to turn the entire city into jelly if their plans are interfered with. This technique aims to evoke fear and anxiety in Person 1 in order to pressure her into abandoning her actions.

C.3 Annotator demographics and preference dataset inter-annotator agreement scores

Criterion	ICC(2, k)	95% CI
Comprehensibility	0.325	[0.030, 0.540]
Completeness	0.712	[0.480, 0.820]
Cognitive Coherence	0.360	[0.060, 0.560]
Soundness	0.632	[0.410, 0.760]

Table 5: Inter-rater reliability (ICC) for evaluation criteria.

Agreement Metric	Value
Fleiss' Kappa	0.350
Krippendorff's Alpha	0.351
Mean Pairwise Cohen's κ	0.353
Exact Agreement (4/4)	32.3%
Majority Agreement ($\geq$3/4)	66.5%

Table 6: Inter-rater agreement for model preference judgments.

The sample of annotators in Study 1 comprised 11 individuals (7 females, 4 males) aged between 20 and 29 years (M = 23.82, SD = 2.89) who had established knowledge of social influence. Among them were 5 management graduates, 2 lawyers, 1 journalism graduate, and 3 psychology students. All annotators had taken courses in social influence during their studies, and were trained in the definitions and examples of emotional influence techniques in interpersonal relationships prior to the annotation process.

In Study 2 (preferences) we limited the number of annotators to 4 from the same group. The new group (3 females, 1 male) aged between 20 and 29 years (M = 25.5, SD = 4.36) included 2 management graduates and 2 lawyers.

D Annotation guidelines

D.1 Study 1

D.1.1 Initial span annotation

Original Polish version

Dla poniższego tekstu i wybranych technik wpływu społecznego zaznacz ich dokładne wystąpienia wystąpienia w tym tekście. Zaznaczaj pełne zdania. Technika może wystąpić w większej liczbie zdań, w tym przypadku zaznacz je wszystkie.

English translation

or the text below and the selected social influence techniques, identify their exact occurrences in the text. Mark full sentences. A technique may occur in more than one sentence; in that case, mark all of them.

D.1.2 Superannotation

Original Polish version

Zweryfikuj już zaanotowane fragmenty tekstu zawierające wpływ społeczny, sprawdzając, czy każdy fragment jest zaznaczony poprawnie i czy zawiera technikę zgodną z definicją. Usuń lub popraw nieprawidłowe zaznaczenia. Jeśli widzisz inne fragmenty tekstu pasujące do wybranych technik, zaznacz je.

English translation

Verify the already annotated text fragments that contain social influence by checking whether each fragment is marked correctly and whether it includes a technique consistent with the definition. Remove or correct incorrect markings. If you see other fragments of the text that match the selected techniques, mark them.

D.2 Study 2

D.2.1 General instructions

Original Polish version

W kolejnych sekcjach formularza znajdziesz krótkie teksty, głównie w formie dialogów, w których zastosowano różne techniki wpływu społecznego.

Pod każdym tekstem znajdują się dwa wyjaśnienia, które tłumaczą, dlaczego przypisano do niego konkretną technikę.

Twoim zadaniem jest:

1. Uważnie przeczytać tekst oraz oba wyjaśnienia
2. Ocenić każde wyjaśnienie według po-

danych kryteriów

3. Wybrać preferowane przez Ciebie wyjaśnienie

English translation

In the following sections of the form, you will find short texts, mainly in the form of dialogues, in which various social influence techniques have been applied.

Below each text, there are two explanations that explain why a specific technique was assigned to it.

Your task is to:

1. Carefully read the text and both explanations
2. Evaluate each explanation according to the given criteria
3. Select your preferred explanation

D.2.2 Evaluation Criteria

For each explanation, annotators evaluated the following aspects on a 5-point scale (from 1 - to a very low degree, to 5 - to a very high degree):

- **Comprehensible** - To what extent the explanation is comprehensible
- **Complete** - To what extent the explanation is complete
- **Cognitively coherent** - To what extent the explanation is logically coherent
- **Sound** - To what extent the explanation is sound/credible

D.2.3 Preference selection

After evaluating both explanations according to the above criteria, annotators selected their preferred explanation:

- Explanation 1
- Explanation 2
- Both equally
- Neither

E Technique definitions and examples

For convenience, here we present the definitions and examples of emotional social influence techniques, following the source paper (Mieleszczenko-Kowszewicz et al., 2025).

1. Emotional see-saw
Inducing a sudden change of emotions in the interlocutor – from positive to negative or vice versa; putting her in a state of emotional disorientation, making him more susceptible to influence.
Example: 'A teacher tells a student that he or she failed an important exam (negative emotions), but then adds that the grade was mistaken and in fact he passed (positive emotions). Then he asks the student: "Can you help me organize the papers? This will help to complete their assessment faster."

2. Fear and anxiety
Inducing a feeling of moderately intense anxiety or fear.
Example: "If you do not take out life insurance, your family will be left without financial support in the event of an accident."

3. Anticipatory regret
Inducing in the interlocutor a sense of regret that may occur in the future due to acting or omitting to act now.
Example: "If you don't start taking care of your health now, then in a few years, when health problems appear, you will regret that you did not do anything about it."

4. Take advantage of a bad mood
Basing social influence on the recipient's current negative mood.
Example: "A partner is irritated after an argument with someone else. You ask him a small favor, such as throwing out the garbage, saying that it will take him away from his worries."

5. Guilt
Inducing a person to feel guilty in order to increase the interlocutor's propensity to do a favor or fulfill a request as a way to reduce guilt.
Example: "You left me alone in this difficult situation and I was counting on your support and help. Please help me in this task."

6. Shame
Inducing a sense of shame in the interlocutor to increase the interlocutor's propensity to do a favor or fulfill a request as a way to alleviate feelings of shame.
Example: "Your work results cast a shadow over the image of the team. I ask that you complete the

next team task on your own."

7. Embarrassment
Inducing this emotional state in the interlocutor to improve his image in the eyes of others.
Example: 'I know this may be inconvenient for you, but I really need your help in selecting people from our department to be fired.'

8. Show your disappointment
Showing disappointment in the interlocutor's behavior in order to get him to comply with a request, which can improve the mood of both parties.
Example: "I could always count on you, and now I feel a little disappointed that you don't have time to help me. Can I ask you for support in this task?"

9. Positive cognitive state
Arousing a state of intrigue or curiosity in an interlocutor through a trick or riddle that he or she is unlikely to solve. As a result, the interlocutor is more likely to comply with requests when experiencing a mixture of curiosity, surprise, and frustration.
Example: "I wonder if you can answer the question my professor once asked me." In a situation where the interlocutor does not find a solution, you suggest "I have an answer for you." In the next step: "I would like to ask you to do a little thing for me."

10. Power of word "love"
Evoking associations in the interlocutor with the feeling of love or a strong positive bond.
Example: "Asking for a donation to a can with the inscription love or love, which more often prompts people to throw money into it."

11. Cognitive exhaustion
Making requests to a person by exploiting their physical, emotional, or mental exhaustion (or after inducing exhaustion), which increases the chance of the request being granted.
Example: Person A: "Could you help me with something small? It's really just a moment." Person B: "What's the matter?" Person A: "Great! I need you to fill out this short survey, it's just 5 questions." Person B: (hesitantly) "Okay, so be it." (B fills out the survey, it takes him longer than he expected.) Person A: "Thank you! And now for the last request – would you please join our list of participants? It's not a big deal, just indicate how many times a month you would like to help with such projects." Person B: (tired of previous activity) "Phew... Okay, type me in 3 times. "

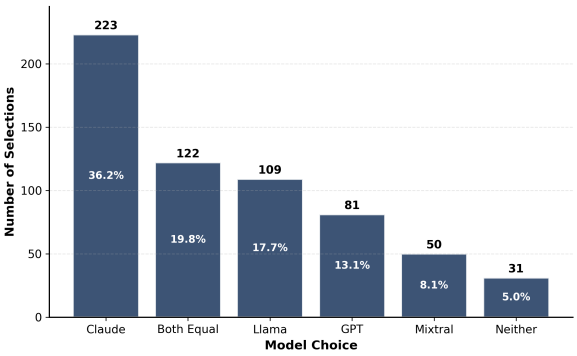

Figure 7: Distribution of human evaluator preferences in pairwise explanation comparisons.

F Additional figures and tables

Note: Mean differences show the absolute difference between model pairs, with positive values indicating that the first model performs better than the second. Significance levels: *** p<0.001, ** p<0.01, * p<0.05, ns = not significant.

Model A	Model B	Wins	Losses	Ties	Win Rate (%)
Claude	Mixtral	67	12	5	79.8
Claude	Llama	88	34	24	60.3
Claude	Gpt	68	17	32	58.1
Llama	Mixtral	39	20	17	51.3
Gpt	Mixtral	25	18	17	41.7
Gpt	Llama	39	36	27	38.2
Llama	Gpt	36	39	27	35.3
Mixtral	Gpt	18	25	17	30.0
Mixtral	Llama	20	39	17	26.3
Llama	Claude	34	88	24	23.3
Gpt	Claude	17	68	32	14.5
Mixtral	Claude	12	67	5	14.3

Table 7: Head-to-Head Performance Summary (Wins-Losses-Ties)

Explainability Criterion	F-stat	p-value	Effect Size	Signif.
Comprehens.	11.55	<0.001	Medium	***
Completeness	47.01	<0.001	Large	***
Cognitive Coh.	15.03	<0.001	Medium	***
Soundness	31.46	<0.001	Large	***
Overall	93.73	<0.001	Very Large	***

Table 8: One-way ANOVA Results Summary

Model Comparison	Mean Difference	p-value	Significance
Claude vs GPT	0.235	0.025	*
Claude vs Llama	0.382	<0.001	***
Claude vs Mixtral	0.457	<0.001	***
GPT vs Llama	0.147	0.295	ns
GPT vs Mixtral	0.222	0.079	ns
Llama vs Mixtral	0.075	0.837	ns

Table 9: Post-hoc Tukey HSD Test Results for Comprehensibility

Model Comparison	Mean Difference	p-value	Significance
Claude vs GPT	0.552	<0.001	***
Claude vs Llama	0.700	<0.001	***
Claude vs Mixtral	1.033	<0.001	***
GPT vs Llama	0.147	0.330	ns
GPT vs Mixtral	0.481	<0.001	***
Llama vs Mixtral	0.333	0.002	**

Table 10: Post-hoc Tukey HSD Test Results for Completeness

Model Comparison	Mean Difference	p-value	Significance
Claude vs GPT	0.269	0.008	**
Claude vs Llama	0.468	<0.001	***
Claude vs Mixtral	0.506	<0.001	***
GPT vs Llama	0.199	0.092	ns
GPT vs Mixtral	0.236	0.061	ns
Llama vs Mixtral	0.037	0.977	ns

Table 11: Post-hoc Tukey HSD Test Results for Cognitive Coherence

Model Comparison	Mean Difference	p-value	Significance
Claude vs GPT	0.391	<0.001	***
Claude vs Llama	0.614	<0.001	***
Claude vs Mixtral	0.819	<0.001	***
GPT vs Llama	0.223	0.052	ns
GPT vs Mixtral	0.429	<0.001	***
Llama vs Mixtral	0.206	0.122	ns

Table 12: Post-hoc Tukey HSD Test Results for Soundness

Model Comparison	Mean Difference	p-value	Significance
Claude vs GPT	0.362	<0.001	***
Claude vs Llama	0.541	<0.001	***
Claude vs Mixtral	0.704	<0.001	***
GPT vs Llama	0.179	<0.001	***
GPT vs Mixtral	0.342	<0.001	***
Llama vs Mixtral	0.163	0.003	**

Table 13: Post-hoc Tukey HSD Test Results for Overall

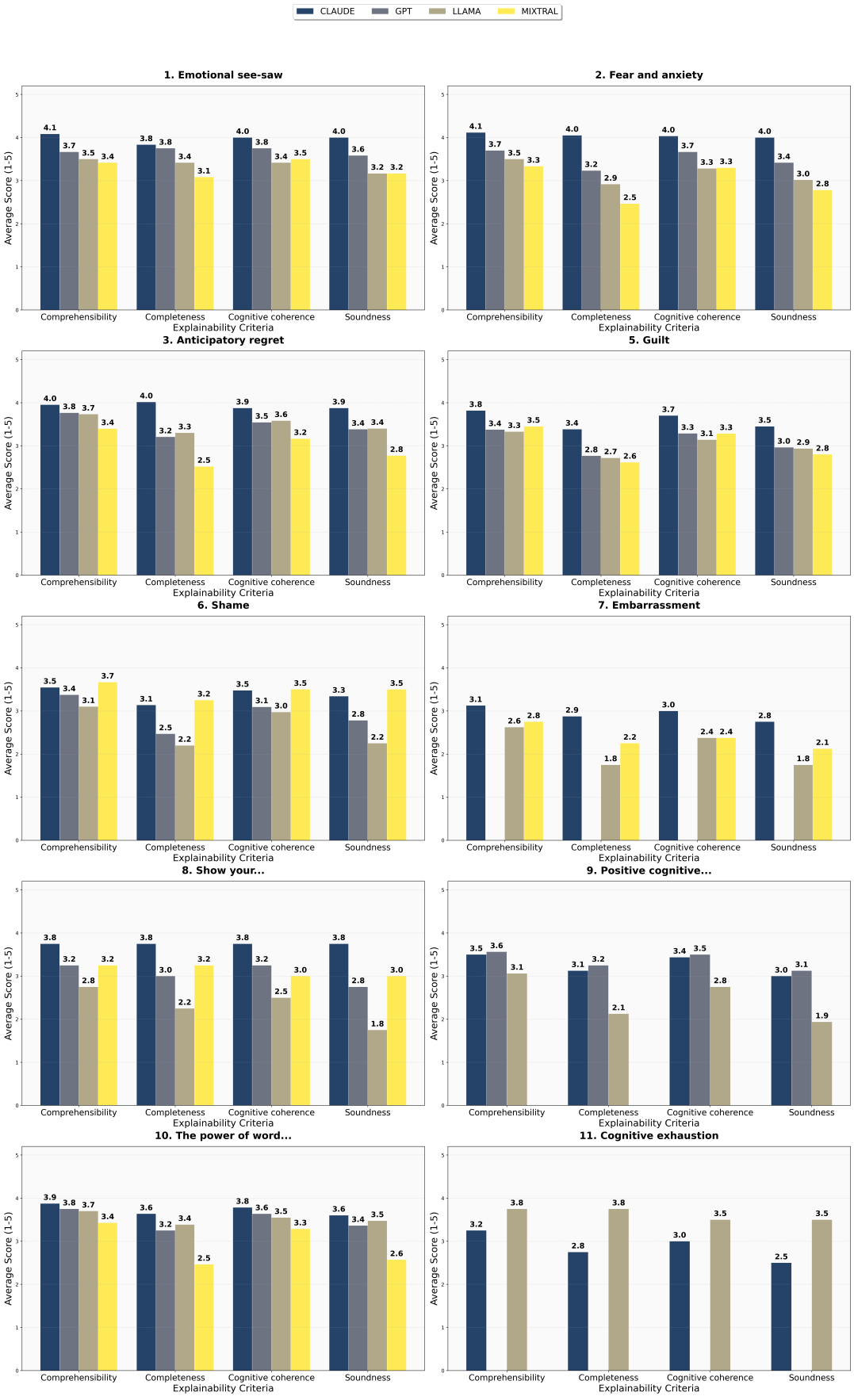

Figure 8: LLM performance comparison across explainability criteria for each explanation of emotional social influence technique.