Using Large Language Models to Detect Insufficient Effort Responding in Open-Ended Survey Questions

Nick von Felten*
University of St. Gallen
St.Gallen, Switzerland
nick.vonfelten@unisg.ch

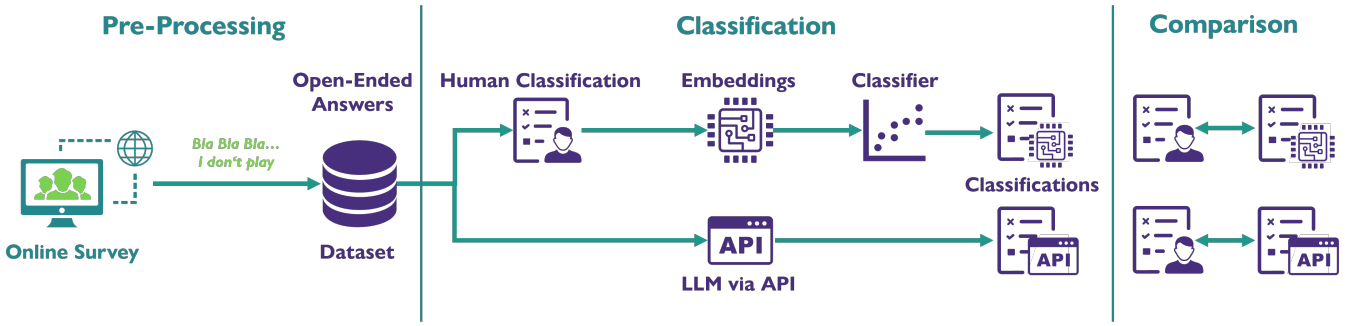


Figure 1: Conceptual illustration of the analysis workflow. The process begins with Pre-Processing, where online survey responses are collected and aggregated. During the Classification phase, open-ended responses are evaluated through one of two pipelines: Classifications are produced either by using manual human coding and sentence embeddings to train a classifier, or by leveraging a Large Language Model via API. Finally, the Comparison phase assesses the performance of both pipelines.

Abstract

Careless responses pose a challenge for data quality in online survey research, a core method in human–computer interaction (HCI). Open-ended answers can reveal such insufficient effort responding (IER), but are costly to evaluate manually. I explore the use of two large language model (LLM) pipelines to automate IER detection in a dataset of 1,551 open-text responses: using open-source embedding models with standard classifiers, and using text-generation labelling with GPT-4o-mini. Embedding-based models achieved higher precision, but missed inattentive responses, whereas text generation showed better accuracy yet tended to overpredict IER. These patterns were explained by severe class imbalance, which was identified as a typical feature of high-quality crowdsourced samples and thus a central challenge for automated IER detection. I discuss how such pipelines could be integrated into human-in-the-loop workflows and emphasize the need for curated, openly available datasets and improved model engineering to advance reliable IER detection.

CCS Concepts

• **Human-centered computing → HCI theory, concepts and models**.

Keywords

Insufficient Effort Responding; Data Quality; AI-assisted research methods, Crowdsourced Survey Research; Large Language Models; Sentence Embeddings; Open-ended Responses; Human–Centered AI

ACM Reference Format:
Nick von Felten. 2026. Using Large Language Models to Detect Insufficient Effort Responding in Open-Ended Survey Questions. In *1st International Conference on Human-Computer Interaction in the Alps (AlpCHI 2026), March 01–05, 2026, Ascona, Switzerland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3780045.3780059

*Corresponding author.

This work is licensed under a Creative Commons Attribution 4.0 International License.
AlpCHI 2026, Ascona, Switzerland
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1948-6/26/03
https://doi.org/10.1145/3780045.3780059

1 Introduction

"I do not know what to write... 50 characters, bla bla blaaaaa." Such responses are commonly observed in online surveys, which constitute one of the most frequently employed methods for data collection in HCI research [2, 26, 30]. In recent years, crowdsourcing platforms such as Amazon Mechanical Turk [1] and Prolific [31] have become dominant channels for participant recruitment, providing rapid access to large and ostensibly more diverse samples compared with traditional recruitment methods [5].

However, the accessibility that makes crowdsourcing appealing for researchers also introduces challenges. Many participants on these platforms may be primarily motivated by monetary gains and thus try to maximise rewards instead of engaging with the study. This leads to a phenomenon known as *insufficient effort responding* (IER) [15] and poses a serious threat to data quality [8]. Even small proportions of inattentive responses can distort correlations, inflate or attenuate effects, and undermine replicability [14, 18, 20, 25].

For HCI studies that rely on these samples, IER challenges the fundamental validity of the field.

While numerous approaches have been proposed to mitigate IER, most target closed-ended survey items such as self-report scales [4]. Despite their compelling potential to reveal whether participants meaningfully engaged with the task, open-ended responses have received little attention with respect to systematically detecting IER. Yet they offer valuable interpretative cues for researchers assessing attentiveness and comprehension in less stringent contexts. In some studies, these qualitative responses were the only observable indicators of inattentive behaviour, where conventional detection techniques failed [4]. However, identifying IER in open responses typically requires laborious manual review, which becomes infeasible in large-scale studies. With the advent of modern artificial intelligence (AI) like Large Language Models (LLMs) this raises the question of whether AI could assist researchers by automatically classifying IER in open-ended survey data. Thus supporting scalable, high-quality analysis without losing the human sensitivity that defines inquiry in HCI.

This paper explores the potential of LLM-based approaches to reliably identify IER in open-ended survey data by evaluating two pipelines: feature extraction with embedding models and text generation with LLMs. While neither approach achieved satisfactory performance, the study uncovered fundamental challenges that warrant further investigation. Specifically, it revealed that class imbalance poses a major obstacle, likely stemming from inherent characteristics of crowdsourced online survey research. Moreover, the findings suggest that open-source embedding models tend to underpredict IER, whereas flagship text-generation models tend to overpredict. The paper concludes by outlining these challenges, discussing how such models could be integrated into human-in-the-loop workflows, and suggesting directions for future work.

2 Related Work

As crowdsourced online surveys are a common tool in HCI research, ensuring response quality is important. This section reviews how prior work has addressed this challenge, beginning with research on IER and extending to emerging applications of LLMs for its detection.

2.1 Insufficient Effort Responding

IER is a prevalent issue in online survey research, with estimates ranging from 3% in 2014 [20] to nearly 50% of responses identified as IER in recent years [3, 4]. Researchers have developed a variety of techniques to detect and manage IER. A priori methods embed attentiveness checks within surveys, steer participants toward more careful responding, or make them self-identify as being careless [4, 8]. While post hoc methods usually try to flag suspicious response patterns through statistical means [4, 8, 22].

The growing prevalence of online surveys has renewed interest in open-ended questions, which were shown to elicit richer and more thoughtful responses than in paper based tests with open-ended responses [38]. The quality of responses, however, strongly depends on participants' intrinsic interest in the topic [13, 36]. This marks paid crowdworkers as a potentially risky population, as their primary motivation may be monetary rather than driven by genuine interest in the task or its content.

Mitigating IER in open-ended formats has received some attention in a priori mitigation approaches. For instance, increasing text box sizes were shown to result in longer responses, while providing motivating instructions equated to better response [17]. However, post hoc quality assessment methods in this context remain underexplored. One exception is the coding guideline developed by Brühlmann et al. [4], which synthesizes earlier work by Smyth et al. [38] and Holland and Christian [13]. While short response length was used as an indicator in several prior studies, it could be considered an unreliable measure of insufficient effort. Failing to follow explicit instructions (e.g., a minimum word count) can indeed signal IER, but verbosity alone is not a valid proxy for quality. Short, well-constructed answers may reflect careful thought, and succinct expression rather than inattentiveness.

Recent work suggests that the interpretability of open-text responses is an important quality indicator [36]. This implies that open-text responses hold considerable diagnostic potential [3], offering contextual cues about participants' reasoning and engagement. Yet, the quality of such responses remains far more difficult to discern, typically requiring human labour. Finding automated solutions to this problem (comparable to flagging message length issues) would be valuable for HCI-researchers – particularly in large-scale online research. Recent advances in LLMs may offer a promising avenue to automate parts of this interpretative process, while maintaining contextual sensitivity.

2.2 Large Language Models in IER Detection

With the advent of LLMs, automated data annotation has become a promising frontier. GPT-3.5-turbo has shown superior performance in text annotation tasks compared to MTurk crowdworkers [11], matching recent insights of the potential of LLMs to serve as crowdworkers for generating contextual information [21]. LLMs have demonstrated effectiveness in identifying subpar data quality across indicators of representativeness such as temporal, geographic, and technological coverage [19]. Beyond these quantitative tasks, they have also been explored as analytical aids in qualitative research [37, 39], although there is justified scepticism and caution about their use as fully automatic tools in qualitative analysis [23]. Given that IER classification is a comparatively constrained task that does not involve generating themes, applying LLMs to detect careless or nonsensical responses may represent a promising and impactful avenue for online-survey based HCI research.

In summary, while extensive research has examined IER in closed-ended survey data, open-ended responses remain a comparatively overlooked yet valuable source of quality information. Manual review of such responses is often infeasible at scale, and the existing automated heuristics are challenged by the nuanced linguistic cues that distinguish low-effort from meaningful input in open-ended responses. Recent advances in LLMs present a novel opportunity to address this gap.

3 Methodology

This section outlines the dataset and the two automated pipelines evaluated for detecting IER: feature extraction with open-source

embeddings and text generation with a generative model. While embedding models are not LLMs, they can be viewed as a lightweight alternative, as they build on similar transformer architectures. Accordingly, their roles in this paper are differentiated only when necessary for clarity. I first describe the evaluation dataset, followed by the two approaches and their implementation.

3.1 Dataset

To explore how well the LLM pipelines can classify IER, I used a publicly available dataset from HCI player experience (PX) research containing 1,551 open-ended responses from a critical incident task [27]. Participants in the original study received the following instructions: *"Please describe the digital game you recently played or that you remember well. Try to describe this particular game as accurately and detailed as you remember in at least 50 words, and try to be as concrete as possible. You can use as many sentences as you like."* [27, p.4]. Additionally, participants were asked to provide the name of the game they tried to remember. For the purpose of that study, it was crucial that participants vividly recalled a specific game experience, making the open-ended responses a prerequisite for the validity of subsequent analyses – thereby necessitating high data quality. A human rater manually reviewed all responses and classified them based on their adherence to the task instructions, verifying whether they (a) showed clear memory of playing the game, (b) referred to a digital game, and (c) contained a valid description matching the provided title [27].

This dataset offers an ideal test case for this approach. It stems from an HCI study featuring open-ended responses that are directly relevant to the study's outcomes. The data is highly contextualized, demanding more than the detection of superficial anomalies, and include human quality ratings that provide a reliable "ground truth".

3.2 Approach 1: Feature Extraction

Because I had access to training data (i.e. human IER ratings), the first approach was to use feature extraction and sentence embeddings to train machine learning classifiers. This method was chosen because it is computationally efficient and potentially accessible for most researchers seeking to integrate scalable quality control into their workflow. To ensure this accessibility I only employed open-source models available on Hugging Face [16].

Three models with fewer than one billion parameters were tested (to ensure computational feasibility). Each model ranked among the top ten embedding models for classification tasks according to the MTEB leaderboard [10]: *google/embeddinggemma-300m* (Gemma) [35], *intfloat/multilingual-e5-large-instruct* (e5-instruct) [41], and *Qwen/Qwen3-Embedding-0.6B* (Qwen) [42]. Each model was prompted with a short instruction sentence, as recommended in the respective model cards, followed by the full text and game title. The prompt structure was as follows: *"Instruction: Judge whether the description clearly refers to a real digital game, matches the game title with an appropriate and concrete description, and shows clear memory of playing the game." "Input: [Game title], [Description]."* This ensured contextual embeddings that mirrored the original rating task performed by the human annotator. The goal was to identify an accessible and computationally efficient approach that could support scalable data-quality assessment for HCI researchers.

Sentence embeddings were generated using the *SentenceTransformer* [33] library on Google Colab [12], then normalized and standardized to improve classifier stability. I implemented a stratified shuffle-split cross-validation procedure (five splits, 70/30 train–test ratio) to preserve class balance between IER and valid responses in each split. Three widely used classifiers were tested: Logistic Regression, Ridge Classifier, and Random Forest. Each trained with class-weight balancing to account for class-imbalance. These classifiers were chosen to test different degrees of interpretability and linearity of models.

3.3 Approach 2: Text Generation

Because of the severe class imbalance in the data and limited performance of the feature extraction approach, a complementary pipeline using GPT-4o-mini was implemented. This approach was tested to assess whether state-of-the-art generative models could reliably detect and classify IER without additional training data.

For inference, a custom prompting function was implemented to query the model through the OpenAI Application Programming Interface (API). Each model response was constrained to output a binary label (0 – IER or 1 – Effortful Response (ER)).

The prompts were derived from the original task instructions and coding scheme of the study. They were then iteratively refined to ensure correct binary outputs (0 or 1) and further improved using OpenAI's prompt optimizer [24]. The final prompts are provided in the Appendix. I tested two prompting approaches: zero-shot classification and few-shot classification. In zero-shot the model received only the instructions describing the evaluation task, followed by the game title and participant's open-ended response. While in few-shot classification, the same instruction were extended with four examples illustrating two valid and two invalid responses. The examples were inspired by the real data, but not provided in the data evaluated by the LLM to avoid data leakage.

4 Results

Model performance for all approaches was evaluated using (balanced) accuracy, F1-score, and precision for the minority class (IER). Balanced accuracy accounts for unequal class sizes by averaging recall (true positive rate) across classes. The F1-score captures the balance between precision (how many IER predictions were actually IER) and recall, providing a single measure of classification performance [32].

4.1 Feature Extraction

All classifiers used their default decision threshold to ensure comparability. This reflects performance under standard classification settings, as adopted from similar benchmarking efforts [7]. However, given the strong class imbalance, this threshold may disadvantage models that produce more conservative probability estimates (e.g., Random Forest). Among the reported metrics, the F1 score was the most informative, as it captures the balance between correctly identifying inattentive responses and avoiding the misclassification of valid ones [32]. In real-world applications (where participant compensation might be at stake) F1 represents a reasonable trade-off between fairness toward participants and data integrity.

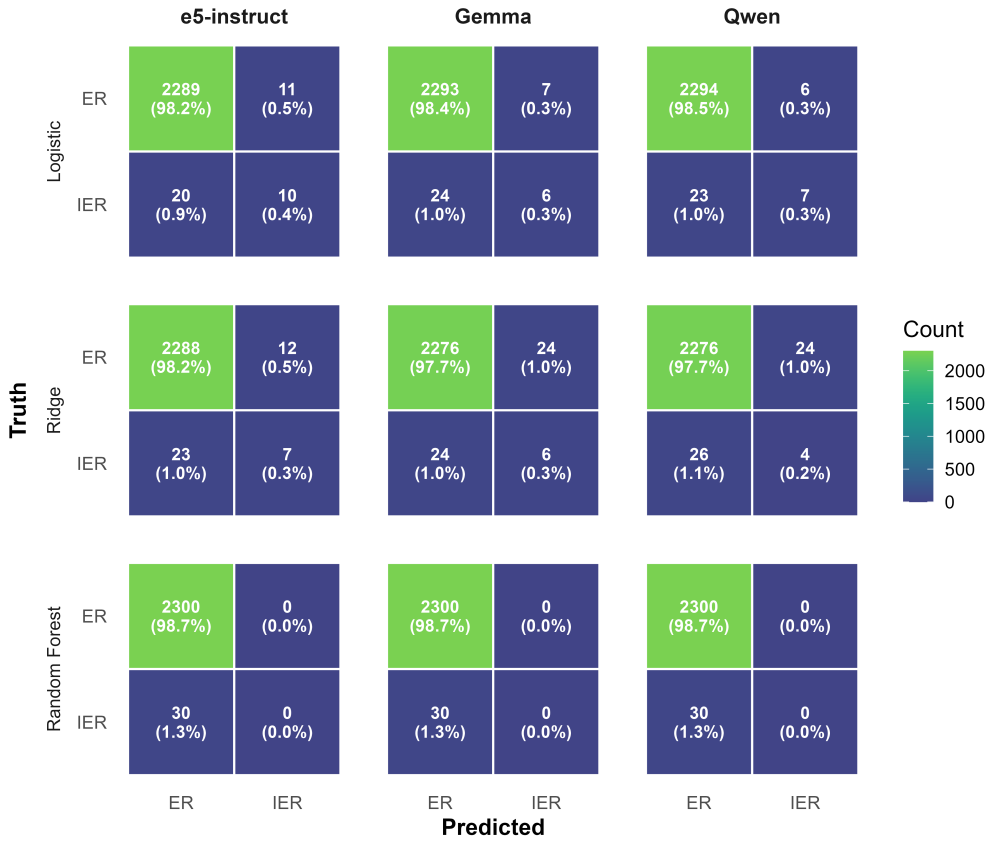


Figure 2: Confusion matrices comparing embedding models and classifiers. The higher total counts result from the stratified shuffle split procedure. Each test set contains 466 cases, and with five splits, this amounts to 2,330 cases in total. Accordingly, the matrix shown here is aggregated across all splits.

Overall, model accuracy was very high across all embedding models and classifiers (>0.97). However, this high accuracy likely reflects a systematic tendency to predict the majority class rather than genuine classification performance (see Figure 2). Therefore, I focus the interpretation on balanced accuracy, as well as F1 score and average precision (AP) for the minority class (class 0). The consistently lower values on these metrics indicate that class imbalance is the primary challenge here, as they more accurately capture the model's ability to identify the minority class.

Across classifiers and embedding models, e5-instruct with logistic regression achieved the highest balanced accuracy and F1 score, indicating the most effective overall performance (see Table 1). Nevertheless, the performance remained poor and it was outperformed on AP by the Qwen and logistic regression combination.

When comparing between classifiers, logistic regression overall yielded the highest balanced accuracy and F1 score, while performing comparably to Random Forest on AP. The Random Forest showed a distinct pattern of equal balanced accuracy (0.5) and F1 (0.0), which can be a characteristic of extreme class imbalance. In such settings, the minority class constitutes only a small percentage of the data, making it difficult for ensemble tree methods to form meaningful decision boundaries. Even when accounting for class imbalance, random forests often have difficulties when the minority class is very small, as most splits and bootstrap samples contain few minority examples. Which results in a model that rarely recognizes the minority class and instead predicts the majority class most of the time [6]. Nonetheless, the non-zero AP suggests that the model captured some signal in its ranking of minority cases, even if this signal did not translate into correct classifications at the 0.5 threshold.

Table 1: Performance of embedding models across classifiers.

	Logistic			Ridge			Random Forest		
Embedding	Bal. Acc.	F1 (C0)	AP (C0)	Bal. Acc.	F1 (C0)	AP (C0)	Bal. Acc.	F1 (C0)	AP (C0)
Qwen	0.615	0.309	0.417	0.561	0.117	0.010	0.500	0.000	0.482
e5-instruct	0.664	0.386	0.300	0.614	0.253	0.013	0.500	0.000	0.339
Gemma	0.598	0.254	0.420	0.595	0.186	0.019	0.500	0.000	0.362

Note. Bal. Acc. = Mean Balanced Accuracy; F1 (C0) = Mean F1 Score for Class 0; AP (C0) = Mean Average Precision for Class 0. Higher values indicate better performance.

4.2 Text Generation

Both prompting approaches with GPT-4o-mini showed moderate ability to distinguish IER and ER. Balanced accuracy was relatively high, and better in the few-shot condition: This indicates that the

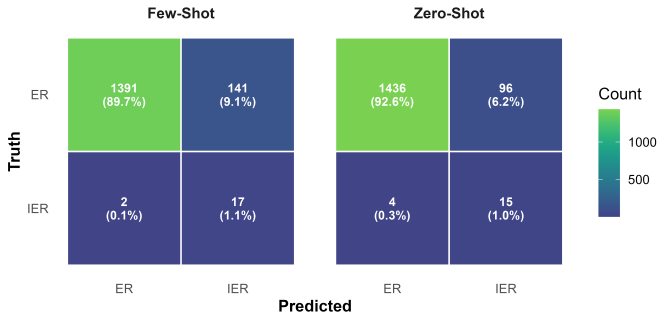


Figure 3: Confusion matrices GPT-4o-mini and prompting strategies.

model performed reasonably well in identifying both classes despite the class imbalance. However, the F1-score and AP for the minority class (IER) were low in both conditions.

This pattern suggests that while GPT-4o-mini could roughly identify differences between valid and invalid responses, it had challenges to consistently recognize true ER cases. In contrast to the feature extraction approach, it overpredicted IER, flagging valid responses as IER too often (see Figure 3). This explains the higher balanced accuracy, despite low scores on F1 and AP (see Table 2). In other words, the model became slightly better overall but still tended to misclassify many ER responses. The increase in balanced accuracy in the few-shot condition indicates some benefit from including examples in the prompt, but the unchanged F1 and AP scores show that this improvement did not translate into more precise detection of the minority class where both prompting strategies perform about equally.

Table 2: Performance of GPT-4o-mini prompting approaches.

	GPT-4o-mini		
Prompting	**Bal. Acc.**	**F1 (C0)**	**AP (C0)**
Zero-shot	0.823	0.289	0.183
Few-shot	0.896	0.288	0.173

Note. Bal. Acc. = Balanced Accuracy; F1 (C0) = F1 Score for Class 0; AP (C0) = Average Precision for Class 0.

5 Discussion

This paper explored approaches for using feature extraction and LLM text generation to detect inattentive response behaviour in open-ended survey questions. While all approaches performed well on majority-class predictions, none achieved satisfactory results for the minority class. The text generation approach yielded higher balanced accuracy, whereas the feature extraction approach achieved slightly better F1 and AP scores.

5.1 Comparing Detection Approaches

Feature extraction appeared to be a promising approach due to its low computational cost, ease of implementation, and compatibility with open-source models, making it accessible for researchers. However, in this dataset, none of the tested embedding–classifier combinations produced satisfactory IER detection. The best performance was achieved with e5-instruct, yet none of the configurations consistently outperformed the others across all metrics. This finding suggests that performance depends on the interaction between embedding models and classifiers, emphasizing the need for careful evaluation of multiple configurations in imbalanced IER classification tasks.

Text generation was tested as a more resource-intensive but similarly accessible alternative. Interestingly, it did not perform substantially better. It exhibited a different classification bias by overpredicting IER. Thus, increasing false positives, whereas feature extraction tended to overpredict valid responses at the expense of missing IER. While few-shot prompting looked generally more favourable, it failed to outperform zero-shot prompting consistently, showing slightly lower AP. Overall, the findings suggest that neither approach demonstrated robust off-the-shelf performance for IER detection within the scope of this exploratory study.

5.2 Class Imbalance as Major Challenge

A major challenge in the dataset was the strong class imbalance, despite using stratified cross-validation and class-weighting. While this may partly reflect this specific dataset, it likely represents a broader challenge in IER detection. When data are collected from high-quality sources, most responses can be expected to be valid, as shown in systematic evaluations [9], and many papers who reported on data quality from these websites [28, 29, 34]. While this is desirable and speaks for some of the platforms, this limits the number of inattentive responses available for model training and evaluation. For text generation, the key difficulty was the tendency to overpredict IER. The underlying causes require further systematic analysis. Nonetheless, this property could be exploited productively for human-in-the-loop workflows.

5.3 Human-in-the-Loop Opportunities

Given the LLMs' overprediction of IER, and very low overprediction of ER, the text generation approach could be used in two complementary workflows:

Full human-in-the-loop annotation: A human annotator labels the dataset, and the LLM independently classifies the same dataset. After annotation, the classifications are compared using inter-rater reliability metrics (e.g., Cohen's Kappa). Discrepancies could reveal potential IER cases missed by the researcher, or flag responses that warrant further discussion within research teams.

Semi-automatic annotation: The LLM pre-screens responses and flags likely IER cases, which are then reviewed by humans. This reduces workload while maintaining quality control. However, until further validation studies are available, and the tendency for overprediction of IER is confirmed, the first workflow offers the safer and more transparent route for rigorous research practice.

5.4 Fine-Tuning

Because this study explored existing embedding and generative models to assess the capabilities of off-the-shelf solutions, fine-tuning represents a logical next step. Training on open-ended survey responses annotated for data quality could enhance sensitivity to subtle indicators of inattentiveness across different domains.

However, such datasets remain scarce. Aside from the dataset used here, I am not aware of any publicly available resources. Curating and releasing such datasets, and subsequent fine-tuning of models on them, would constitute a substantive contribution to the field and provide a plausible path toward reliable automated IER detection.

5.5 Key Takeaways for HCI Researchers

Neither approach offers a robust solution for automated detection of IER in open-ended responses in this investigation. HCI researchers might leverage embedding-based classifiers for exploratory analyses on annotated data, as they are inexpensive and straightforward to implement. However, their tendency to miss IER in this study might point towards limited usefulness beyond providing a lightweight baseline, at least for data with similar characteristics.

The text generation approach might offer HCI researchers higher detection sensitivity but also substantial overclassification. Thus, either assisting researchers as a quality control mechanism with the LLM acting as an independent annotator for inter-rater comparisons; or a triage mechanism which could surface low-effort responses that require closer inspection. Based on the current evidence the full human-in-the-loop annotation should be preferred to ensure rigour. Improving performance will likely require further developments of more intricate workflows, better handling of class imbalance, or fine-tuning with suitable datasets available.

6 Limitations & Future Work

As an exploratory study, this work did not prioritise extensive hyperparameter tuning, model optimisation, or data augmentation, all of which may lead to improved performance in future research. The objective was to assess accessible, low-barrier and low-cost approaches that can be employed by HCI researchers without advanced AI expertise. For this reason, I did not include models with reasoning methods like chain-of-thought. I conducted preliminary tests with GPT-5, and its performance exhibited similar overclassification patterns to GPT-4o-mini. Thus, ultimately I chose not to include GPT-5 in the main study due to substantially higher cost, and the absence of a temperature parameter setting, which reduces reproducibility. Note that more advanced models may behave differently, and these findings should not be interpreted as a general statement about the limits of state-of-the-art LLMs. Likewise, the results are grounded in the characteristics of one publicly available dataset, representing only a single HCI domain, thus, performance may differ for other types of open-ended questions outside of PX HCI research. However, the impact of class imbalance and the contrasting error tendencies of embedding-based and LLM-based pipelines offer meaningful insights for HCI researchers designing similar workflows.

In their current form, the evaluated methods are not yet suitable for dependable real-world deployment. However, the findings indicate two promising directions for future work: integrating large language models into human-in-the-loop workflows to improve the robustness and efficiency of annotation processes, and pursuing engineering-oriented enhancements through imbalanced-learning techniques, data augmentation or fine-tuning of large language models. This study did not include a qualitative error analysis of misclassified responses, which could provide deeper insight into specific error-patterns and should be addressed in future work. Further, the capability of LLMs to generate fluent responses which could mask the lack of genuine effort are a new challenge that warrants further investigation, as it was already shown that the use of LLMs by crowdworkers is on the rise [40]. Together, these directions can contribute to the development of more reliable and scalable solutions for automated IER detection in open-ended survey research.

7 Conclusion

This paper examined two practical LLM-based pipelines for detecting insufficient effort responding in open-ended survey data. Despite strong majority-class performance, neither feature extraction nor text generation achieved satisfactory detection of the minority class. Feature extraction delivered relatively higher F1 and AP for IER yet missed inattentive cases, while text generation improved balanced accuracy but overpredicted IER. Across settings, severe class imbalance was the dominant obstacle and likely represents a typical challenge for high-quality crowdsourced samples.

Given these constraints, using off-the-shelf embedding models or LLMs for this cause is not currently advisable. Promising directions include human-in-the-loop workflows, where LLMs support annotation through disagreement detection or pre-screening, and engineering advances such as imbalanced-learning strategies and fine-tuning on domain-specific datasets. Developing corpora and targeted error analyses are essential for reliable, scalable IER detection in open-ended HCI data.

8 Data Availability Statement

The analysis scripts and data are available in the supplementary materials and on OSF. As this work builds on a dataset by Perrig et al. [27], please also cite their publication when referencing this paper.

9 Funding and Declaration of Conflicting Interests

This research was fully supported by internal funding from our research group. The authors report no commercial or financial conflicts of interest.

Acknowledgments

I used Overleaf's spell checker, the current version of ChatGPT and Gemini for fixing spelling mistakes and to get suggestions to improve the writing of the paper. If not noted otherwise in a specific section, these tools were not used in other forms. I would like to thank the original authors involved in the creation and annotation of the dataset: Sebastian A. C. Perrig, Nicolas Scharowski, Florian Brühlmann, Klaus Opwis, and Lena Fanya Aeschbach. I also thank Zak Hussain and Dirk U. Wulff for their introduction to embedding models and LLMs. Finally, I am grateful to Johannes Schöning and Leon Reicherts for their feedback on this manuscript.

References

[1] Amazon Web Services, Inc. 2025. *Amazon Mechanical Turk: A Crowdsourcing Marketplace for Work*. https://www.mturk.com/

[2] Javier A. Bargas-Avila and Kasper Hornbæk. 2011. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2689–2698. doi:10.1145/1978942.1979336

[3] Florian Brühlmann, Zgjim Memeti, Lena F. Aeschbach, Sebastian A. C. Perrig, and Klaus Opwis. 2024. The Effectiveness of Warning Statements in Reducing Careless Responding in Crowdsourced Online Surveys. *Behavior Research Methods* 56, 6 (Jan. 2024), 5862–5875. doi:10.3758/s13428-023-02321-z

[4] Florian Brühlmann, Serge Petralito, Lena F. Aeschbach, and Klaus Opwis. 2020. The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology* 2 (2020), 100022. doi:10.1016/j.metip.2020.100022

[5] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6, 1 (Jan. 2011), 3–5. doi:10.1177/1745691610393980

[6] Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. 2004. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.* 6, 1 (June 2004), 1–6. doi:10.1145/1007730.1007733

[7] Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. 2018. Random Forest versus Logistic Regression: A Large-Scale Benchmark Experiment. *BMC Bioinformatics* 19, 1 (Dec. 2018), 270. doi:10.1186/s12859-018-2264-5

[8] Paul G. Curran. 2016. Methods for the Detection of Carelessly Invalid Responses in Survey Data. *Journal of Experimental Social Psychology* 66 (Sept. 2016), 4–19. doi:10.1016/j.jesp.2015.07.006

[9] Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. 2023. Data Quality in Online Human-Subjects Research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE* 18, 3 (March 2023), e0279720. doi:10.1371/journal.pone.0279720

[10] Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. MMTEB: Massive Multilingual Text Embedding Benchmark. *arXiv preprint arXiv:2502.13595* (2025). doi:10.48550/arXiv.2502.13595

[11] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences* 120, 30 (July 2023), e2305016120. doi:10.1073/pnas.2305016120

[12] Google. 2024. *Google Colaboratory*. https://colab.research.google.com/ Accessed: 2025-10-20.

[13] Jennifer L. Holland and Leah Melani Christian. 2009. The Influence of Topic Interest and Interactive Probing on Responses to Open-Ended Questions in Web Surveys. *Social Science Computer Review* 27, 2 (May 2009), 196–212. doi:10.1177/0894439308327481

[14] Jason L. Huang, Nathan A. Bowling, Mengqiao Liu, and Yuhui Li. 2015. Detecting Insufficient Effort Responding with an Infrequency Scale: Evaluating Validity and Participant Reactions. *Journal of Business and Psychology* 30, 2 (June 2015), 299–311. doi:10.1007/s10869-014-9357-6

[15] Jason L. Huang, Paul G. Curran, Jessica Keeney, Elizabeth M. Poposki, and Richard P. DeShon. 2012. Detecting and Deterring Insufficient Effort Responding to Surveys. *Journal of Business and Psychology* 27, 1 (March 2012), 99–114. doi:10.1007/s10869-011-9231-8

[16] Inc. Hugging Face. 2025. *Hugging Face — The AI community and model hub*. Accessed: 2025-10-14.

[17] Glen D. Israel. 2010. Visual Cues and Response Format Effects in Mail Surveys. *Journal of Official Statistics* 26, 2 (2010), 271–285.

[18] Chester Chun Seng Kam. 2019. Careless Responding Threatens Factorial Analytic Results and Construct Validity of Personality Measure. *Frontiers in Psychology* 10 (June 2019), 1258. doi:10.3389/fpsyg.2019.01258

[19] Steven MacMaster and Julie Sinistore. 2024. Testing the Use of a Large Language Model (LLM) for Performing Data Quality Assessment. *The International Journal of Life Cycle Assessment* (Dec. 2024). doi:10.1007/s11367-024-02405-8

[20] Michael R. Maniaci and Ronald D. Rogge. 2014. Caring about Carelessness: Participant Inattention and Its Effects on Research. *Journal of Research in Personality* 48 (Feb. 2014), 61–83. doi:10.1016/j.jrp.2013.09.008

[21] Iván Martínez-Murillo, María Miró Maestre, Armando Suárez, Elena Lloret, and Paloma Moreda. 2026. Assessing the Potential of LLMs as Crowdworkers for Contextual Information Generation. *Information Processing & Management* 63, 3 (2026), 104486. doi:10.1016/j.ipm.2025.104486

[22] Adam W. Meade and S. Bartholomew Craig. 2012. Identifying Careless Responses in Survey Data. *Psychological Methods* 17, 3 (Sept. 2012), 437–455. doi:10.1037/a0028085

[23] Supriya D. Mehta, Souvik Paul, Enid Awiti, Sophie Young, Garazi Zulaika, Fredrick O. Otieno, Penelope A. Phillips-Howard, Linda Mason, and Runa Bhaumik. 2025. Evaluation of Large Language Models within GenAI in Qualitative Research. *Scientific Reports* 15, 1 (Oct. 2025), 34993. doi:10.1038/s41598-025-18969-w

[24] OpenAI. 2025. *Optimize for GPT-5*. https://platform.openai.com/chat/edit?models=gpt-5&optimize=true Accessed: 2025-10-17.

[25] Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power. *Journal of Experimental Social Psychology* 45, 4 (2009), 867–872. doi:10.1016/j.jesp.2009.03.009

[26] Sebastian A. C. Perrig, Lena Fanya Aeschbach, Nicolas Scharowski, Nick von Felten, Klaus Opwis, and Florian Brühlmann. 2024. Measurement Practices in User Experience (UX) Research: A Systematic Quantitative Literature Review. *Frontiers in Computer Science* 6 (March 2024), 1368860. doi:10.3389/fcomp.2024.1368860

[27] Sebastian A. C. Perrig, Nicolas Scharowski, Florian Brühlmann, Nick von Felten, Klaus Opwis, and Lena Fanya Aeschbach. 2024. Independent Validation of the Player Experience Inventory: Findings from a Large Set of Video Game Players. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 370, 13 pages. doi:10.1145/3613904.3642270

[28] Sebastian A. C. Perrig, Nick von Felten, Marimo Honda, Klaus Opwis, and Florian Brühlmann. 2023. Development and Validation of a Positive-Item Version of the Visual Aesthetics of Websites Inventory: The VisAWI-Pos. *International Journal of Human–Computer Interaction* (Sept. 2023), 1–25. doi:10.1080/10447318.2023.2258634

[29] Sebastian A. C. Perrig, Nick von Felten, Beat Vollenwyder, and Klaus Opwis. 2024. Development and Psychometric Validation of a Positively Worded German Version of the System Usability Scale (SUS). *International Journal of Human–Computer Interaction* (Dec. 2024), 1–21. doi:10.1080/10447318.2024.2434720

[30] Ingrid Pettersson, Florian Lachner, Anna-Katharina Frison, Andreas Riener, and Andreas Butz. 2018. A Bermuda Triangle? A Review of Method Application and Triangulation in User Experience Evaluation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3173574.3174035

[31] Prolific. 2024. *Prolific*. London, UK. https://www.prolific.com First released 2014. Copyright 2024..

[32] Oona Rainio, Jarmo Teuho, and Riku Klén. 2024. Evaluation Metrics and Statistical Tests for Machine Learning. *Scientific Reports* 14, 1 (March 2024), 6086. doi:10.1038/s41598-024-56706-x

[33] 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*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[34] Nicolas Scharowski, Sebastian A. C. Perrig, Nick von Felten, Lena Fanya Aeschbach, Klaus Opwis, Philipp Wintersberger, and Florian Brühlmann. 2025. To Trust or Distrust AI: A Questionnaire Validation Study. In *2025 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Athens, Greece.

[35] Henrique* Schechter Vera, Sahil* Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divya Sreepat, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Yunhsuan Sung, Raphael

Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. 2025. EmbeddingGemma: Powerful and Lightweight Text Representations. (2025). https://arxiv.org/abs/2509.20354

[36] Katharina Schmidt, Tobias Gummer, and Joss Roßmann. 2019. Effects of Respondent and Survey Characteristics on the Response Quality of an Open-Ended Attitude Question in Web Surveys. *methods, data, analyses (Online First)* (July 2019), 32 Pages. doi:10.12758/MDA.2019.05

[37] Hope Schroeder, Marianne Aubin Le Quéré, Casey Randazzo, David Mimno, and Sarita Schoenebeck. 2025. Large Language Models in Qualitative Research: Uses, Tensions, and Intentions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 481, 17 pages. doi:10.1145/3706598.3713120

[38] Jolene D. Smyth, Don A. Dillman, Leah Melani Christian, and Mallory Mcbride. 2009. Open-Ended Questions in Web Surveys. *Public Opinion Quarterly* 73, 2 (2009), 325–337. doi:10.1093/poq/nfp029

[39] Robert H. Tai, Lillian R. Bentley, Xin Xia, Jason M. Sitt, Sarah C. Fankhauser, Ana M. Chicas-Mosier, and Barnas G. Monteith. 2024. An Examination of the Use of Large Language Models to Aid Analysis of Textual Data. *International Journal of Qualitative Methods* 23 (Jan. 2024), 16094069241231168. doi:10.1177/16094069241231168

[40] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. arXiv:2306.07899 [cs] doi:10.48550/arXiv.2306.07899

[41] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint arXiv:2402.05672* (2024).

[42] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176* (2025).

A Appendix

A.1 Prompts

A.1.1 Zero-Shot. Developer: # Role and Objective Evaluate whether a participant's description of a digital game meets specified quality standards.

Instructions - Read the provided game title and description. - Apply the following criteria: - Return 1 if the description: - Clearly refers to a real digital game. - Accurately matches the game title with a concrete, appropriate description. - Demonstrates a clear memory of playing the game. - Otherwise, return 0. - If the input is empty or malformed, return 0. # Input

title: title description: description

Output Format -Return only a single integer value: ‘1‘ or ‘0‘. -Do not output any explanation or additional content. -Stop immediately after outputting the integer. - Do not include backticks, quotation marks, spaces, explanations, or any other text. - The output must be a **plain integer**, not inside a code block or sentence. Output (MUST be only 0 or 1):

Stop Conditions - Stop after outputting the integer (‘1‘ or ‘0‘)."""

A.1.2 Few-Shot. Developer: # Role and Objective Evaluate whether a participant's description of a digital game meets specified quality standards.

Instructions - Begin with a concise checklist (3-5 bullets) of the steps you will take to evaluate the input. - Read the provided game title and description. - Use the examples to compare the quality of the response. - Apply the following criteria: - Return 1 if the description: - Clearly refers to a real digital game. - Accurately matches the game title with a concrete, appropriate description. - Demonstrates a clear memory of playing the game. - Otherwise, return 0. - If the input is empty or malformed, return 0.

Reference Examples **High-Quality (1):** - Example 1: - title: The Legend of Zelda: Breath of the Wild - description: I played Breath of the Wild a while back. It's an open-world adventure game where you explore vast landscapes and solve physics-based puzzles. I remember loving the freedom to experiment and the sense of discovery throughout the world. - Example 2: - title: The Witcher 3: Wild Hunt - description: I really enjoyed The Witcher 3. The storytelling was incredible, the characters felt real, and I spent countless hours completing quests and exploring the countryside on horseback.

Low-Quality (0): - Example 1: - title: - - description: I can't recall much about the game... maybe it was something web-based? Anyway, I remember enjoying it at the time. - Example 2: - title: Speech Rehab Game - description: This wasn't really a standard game; it was more like an experimental tool designed for use in a speech rehabilitation project. # Input title: title description: description

Output Format -Return only a single integer value: ‘1‘ or ‘0‘. -Do not output any explanation or additional content. -Stop immediately after outputting the integer. - Do not include backticks, quotation marks, spaces, explanations, or any other text. - The output must be a **plain integer**, not inside a code block or sentence. Output (MUST be only 0 or 1):

Stop Conditions - Stop after outputting the integer (‘1‘ or ‘0‘).