AGILE: GIScience Series, 7, 45, 2026. https://doi.org/10.5194/agile-giss-7-45-2026
Proceedings of the 29th AGILE Conference on Geographic Information Science, 16–19 June 2026.
Eds.: Evelyn Uuemaa, Alexander Kmoch, Holger Virro, and Simon Scheider
This contribution underwent peer review based on a full paper submission.

Can Large Language Models interpret participatory mapping comments at scale? Evidence from Prague's emotional mapping

Lydia Valdesera [1], Thibaud Chassin [1], and Jiří Pánek [2]

[1]Department of Geography and Regional Science, University of Graz, Austria
[2]Department of Development and Environmental Studies, Palacky University Olomouc, Czech Republic

Correspondence: Thibaud Chassin (thibaud.chassin@uni-graz.at)

Abstract. Citizen contributions from online participatory mapping platforms surface local knowledge and lived experiences that can inform urban decision-making. However, as these initiatives scale, practitioners increasingly struggle to extract meaningful information. While frameworks and methods to analyse the spatial dimension of these contributions exist, there is a lack of approaches leveraging their free-text content. This paper investigates whether small local Large Language Models (LLMs) can help in structuring this data via (i) sentiment polarity analysis and (ii) thematic classification. Our case study is based on an openly available dataset from 2021 for the city of Prague (N≈30000), where residents' emotional and perceptual relationships to the city were collected. We evaluated the performance of three models across two languages (Czech and translated English text) on 732 randomly sampled and manually annotated citizen comments, using five repeated runs to account for output variability. Our results revealed that OpenEuroLLM-Czech:12b achieved the highest weighted F1 score for sentiment analysis and Gemma3:12b for thematic classification (85.2% and 70.4%, respectively). All LLMs also exceeded a RoBERTa sentiment classification baseline (weighted F1 score 72.1%) on the sentiment task. Overall, this study suggests a potential for using LLMs in participatory mapping, with implications for more engaging and inclusive participation. We also call for the development of ethical guardrails for the responsible use of LLMs in participatory mapping.

Submission Type. Analysis

BoK Concepts. [GD2] Data collection, [GS4] Geospatial citizenship, [GC3] Artificial intelligence (AI) in EO and GI

Keywords. Participatory mapping, Georeferenced Text, Natural Language Processing, Large Language Models, Urban Participatory Planning

1 Introduction

Advances in Information and Communications Technologies (ICT) have enabled citizens to share spatially referenced data through web platforms (ranging from social media to participatory mapping platforms), making public voices heard in impactful ways (Atzmanstorfer et al., 2014). This democratisation produces large volumes of citizen-generated data that contains rich, unstructured, multicultural textual information. Gathering and processing such feedback has become increasingly valuable in urban planning because it offers insights into the public's needs (Gholamnia et al., 2025). In participatory mapping, citizens' contributions take the shape of simple geographical entities, mainly points, less frequently lines, and polygons (Vallejo-Velázquez et al., 2025; Ramírez Aranda et al., 2021). The analysis of these geometries encompasses numerous spatial methods, which are grouped into three categories: "*Explore*", "*Explain*", and "*Predict/Model*" (Fagerholm et al., 2021). However, a critical issue of current practices is that the text-based narratives present in citizens' spatial contributions are rarely interpreted (Brown and Kyttä, 2014), leaving the underlying context and emotions undocumented. These narratives are traditionally browsed manually, but this method quickly becomes impractical when processing successful online participatory mapping initiatives with thousands of contributions. Additionally, the widespread nature of online participation means citizen contributions are increasingly multilingual, mirroring the cultural and linguistic diversity of cities. However, these multilingual comments, are difficult to

interpret (Gracia et al., 2012). This represents a significant gap: while having rich participatory data with both spatial and potentially multilingual textual components, we lack approaches to process the data and assess the qualitative dimensions embedded in these written contributions.

Recent advancements in Artificial Intelligence and LLMs have enabled the analysis of unstructured text into structured insights, leading to potential applications in participatory urban planning. Examples are multilingual processing, sentiment analysis, and information extraction from citizen contributions. LLMs are increasingly employed for numerous tasks such as synthesising data from social platforms and reports, supporting individual travel-related queries through fine-tuned conversational interfaces, and analysing proximity perceptions of spatial attributes from user-generated data sources (Han et al., 2024; Wang et al., 2024; Shingleton and Basiri, 2025). These studies highlight the utility of LLMs, including their various uses, opportunities, and inherent risks (Liu et al., 2025; Raymond et al., 2025), in interpreting public comments for providing deeper insights and complementing traditional social survey methods.

This paper introduces an exploratory benchmark method of participatory data analysis using LLMs. We first describe the open dataset (N≈30000) from Prague, which we mobilise for our analysis, our sampling strategy (N=732), and the reference labels used as ground truth. We then detail the two tasks, namely (i) sentiment and (ii) thematic classification, defined to evaluate three LLMs. Next, we report model performance using F1 score and confusion matrices. Finally, we discuss the opportunity to responsibly include LLMs in participatory mapping practices.

2 Data and Method

2.1 Dataset overview

The dataset used in this research documents residents' emotional and perceptual relationships to the city of Prague (Czech Republic), collected through a large-scale participatory mapping initiative (Pánek et al., 2021). Data were gathered between April and September 2021 using the web platform EmotionalMaps.eu, which enabled participants to indicate specific locations on an interactive map and attach short textual comments in response to a set of existing participatory mapping categories. In total, the dataset comprises responses from 5,973 participants, who collectively contributed 98,364 spatial points and 30,941 associated textual comments (only the latter are considered in this study). Respondents were able to mark more than one location per question, and the additional comments were voluntary, therefore, only approximately 1/3 of all points have comments. Each respondent started with a blank map to avoid clustering of answers. This also meant, that respondents could not comment on others' responses. The participatory mapping categories focused on subjective evaluations of urban space, such as positive and negative emotions, perceptions of safety, and places of personal significance, with a core set of categories applied city-wide. Table 1 shows the eight participatory mapping categories.

The survey was originally conducted in Czech, and the respondents' comments were translated via a paid version of DeepL right after the end of the survey. The columns used for further analysis are those written in the original Czech and the automatically generated English ones.

Category ID	Participatory mapping categories	Total comments per category	Sampled comments per category
1	This is where I spend my free time.	4274	76
2	I would show this place to a visitor.	2746	78
3	This place is neglected and needs to be renovated.	5330	95
4	I don't feel safe here (suspicious people, neglected environment, etc.).	4051	96
5	There is a traffic hazard here (for walking, cycling, motor vehicle, lack of pedestrian crossing etc.).	5561	99
6	There are parking problems (not enough parking spaces, cars parked inappropriately, etc.).	2917	98
7	I would like more green space here.	2357	94
8	There is often an overflowing waste bin or collection point for municipal/sorted waste.	1387	96

Table 1. Common participatory mapping categories for all Prague districts.

2.2 Data filtering and preprocessing

A subset of 800 randomised comments was created from the initial dataset, with 100 comments per participatory mapping category. This sample size was chosen to balance resource limitations for manual annotation and the random selection aimed to capture a broad range of linguistic and thematic variability from the complete dataset. The length of the comments varied from 2 to 254 characters, averaging at 56 symbols (including spaces) per

comment. The resulting selection was manually annotated as negative, neutral, or positive for each comment. The annotation was conducted by three independent native Czech speakers, who encountered difficulties in identifying the sentiment due to sarcasm and ironic tone. To establish a reliable ground truth for LLM evaluation, their individual labels were then unified through a majority vote. The translated comment "*I guess it's better than I remember, but it still seems borderline sometimes :-(",* is an example that shows the challenge in labelling participatory mapping data, as all three annotators provided different labels. This comment was considered ambiguous and excluded from the analysis. To reduce noise, we filtered out 67 comments, because they contained only a place reference (e.g. street label) without any narrative. For instance, a comment where the text only referred to "*Baroňák*", which is the name of a hill located in the *Horní Počernice* district of Prague, was removed (see table 1). The remaining 732 comments were then processed. The density of these comments per district is shown in Fig. 1.

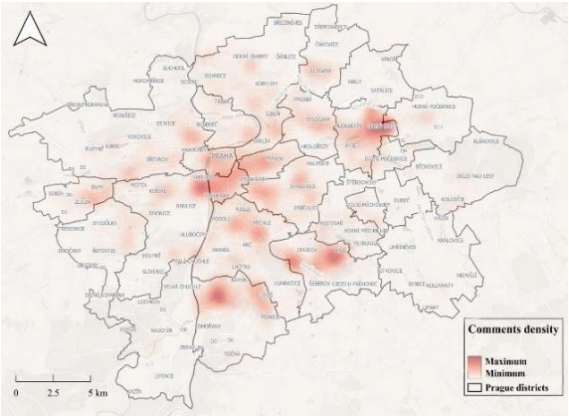

Figure 1. Density of comments in Prague districts.

2.3 Models used

For the analysis, we used small local LLMs to make comparisons between the models' performances based on: (i) 3-class sentiment polarity; (ii) 8-class category assignment. For both tasks, we employed: (1) *Gemma3:12b* model (Kamath et al., 2025), selected for its broad multilingual coverage (140+ languages); (2) *OpenEuroLLM-Czech:12b* model, a fine-tuned version of *Gemma3* optimised for Czech language responses, included for a Czech-specialised baseline against general multilingual models; and (3) *DeepSeek-R1-0528-Qwen3-8B* model, a minor version upgrade of *DeepSeek-R1* (Guo et al., 2025), selected for its advanced reasoning despite its size. We ran all models locally using Ollama (version 0.16.2). To establish a reproducible baseline during this exploratory phase, all LLMs were deployed with their default settings. Specifically, we kept the temperature for *DeepSeek-R1-0528-Qwen3-8B:*0.6, for *OpenEuroLLM-Czech*:0.7, and for *Gemma3*:0.1. Other parameters, including Top-k and Top-p, were similarly held at default values. As reference point for sentiment polarity, we include a transformer baseline, namely *Cardiffnlp/twitter-roberta-base-sentiment-latest*, a *RoBERTa* base model (Liu et al., 2019). Our selection of this baseline model followed a preliminary comparative study that demonstrated greater predictive accuracy on our sampled dataset over the *Distilbert/distilbert-base-uncased-finetuned-sst-2-english* model (Sanh et al., 2020). The model's training on ~124M tweets and fine-tuned for sentiment analysis aligns closely with the nature of our citizen comment data.

2.4 Sentiment analysis, thematic analysis and reasoning

We employed LLMs to analyse the emotional tone within participants' comments. We kept punctuation and emojis, which can offer meaningful insights. For each comment, the LLM was prompted to provide three distinct outputs: the comment's polarity (-1 for negative, 0 for neutral, and +1 for positive), its reasoning explaining the *logic* behind its classification, and a confidence score from 0 to 1 indicating the model's confidence in the prediction. For the thematic analysis task, our objective was to evaluate the LLM's capacity to categorise unstructured citizen comments into the existing eight participatory mapping categories (which were used as ground truth) in the original survey. By instructing the model to assign each comment to a category (without providing the link to them) and its supporting reasoning, we tested whether the LLM could *infer* the underlying urban planning context inherent in citizen feedback. This task validates the model's *understanding* in a participatory mapping setting and establishes the groundwork for future iterations where LLMs might move beyond predefined schemas to generate new thematic categories directly from the context of the public's feedback.

All runs were conducted within Jupyter Notebook environments and repeated five times to account for stochastic decoding of the models (temperature > 0). For the tasks involving Czech comments (CZ), we utilised the *OpenEuroLLM-Czech* and *Gemma3* models, while the *DeepSeek-R1-0528-Qwen3-8B* and *Gemma3* models were employed for translated English comments (EN). Two distinct prompts were developed, tailored for each specific task and language (a representative example can be found in the Appendix). To enhance processing efficiency and speed, LLM inputs were provided in batches of 10 comments. Moreover, we ensured the structured outputs by using the Ollama Python library in

conjunction with Pydantic, which facilitates the definition and serialisation of the required JSON schema.

Hereafter, when we refer to sentiment analysis and thematic analysis, we will be referring to the comments' classification task into three categories (negative, neutral, positive), and to the assignment of comments to specific participatory mapping categories, respectively.

2.5 Data and Software Availability Section

The textual data and the analysis scripts used in this study are openly accessible at: https://osf.io/3wb6v/. The repository includes the raw data, the ground truth labels, and the LLM prompts used for inference.

3 Results

After running the models five times to quantify the model's variability, we calculated for each model/tasks the weighted F1 scores per run, overall mean weighted F1 score, standard deviation, and average number of not processed comments. Table 2 summarises a comparative model performance per task.

Sentiment analysis (Fig.2). Classifying comments in the original Czech demonstrated the highest mean weighted F1 scores, with *Gemma3* and *OpenEuroLLM-Czech* reaching above 85%. The *Gemma3* model also performed with high accuracy on the English translations, by showing a decrease of 2.5%, which could be justified by the automatic translation quality using DeepL. The standard deviation and average of not processed comments of all models, except from the *DeepSeek-R1-0528-Qwen3-8B,* stay at low levels, highlighting very consistent performance across their five runs. All LLMs achieved higher sentiment-classification accuracy than the RoBERTa baseline (~72% accuracy).

Thematic analysis (Fig.2). We notice a significant drop in all mean weighted F1 scores (ranging from 70.4% to 55.8% for all models). This task, compared to sentiment analysis, is more complex as it requires classification into eight participatory mapping categories.

The *OpenEuroLLM-Czech* model showed the lowest standard deviation (0.003), indicating highly consistent performance across the five runs, even though its F1 score was lower than *Gemma3* run on the Czech comments. This suggests a stable, albeit not leading, performance for the localised LLM compared to the more general-purpose multilingual models in this specific task.

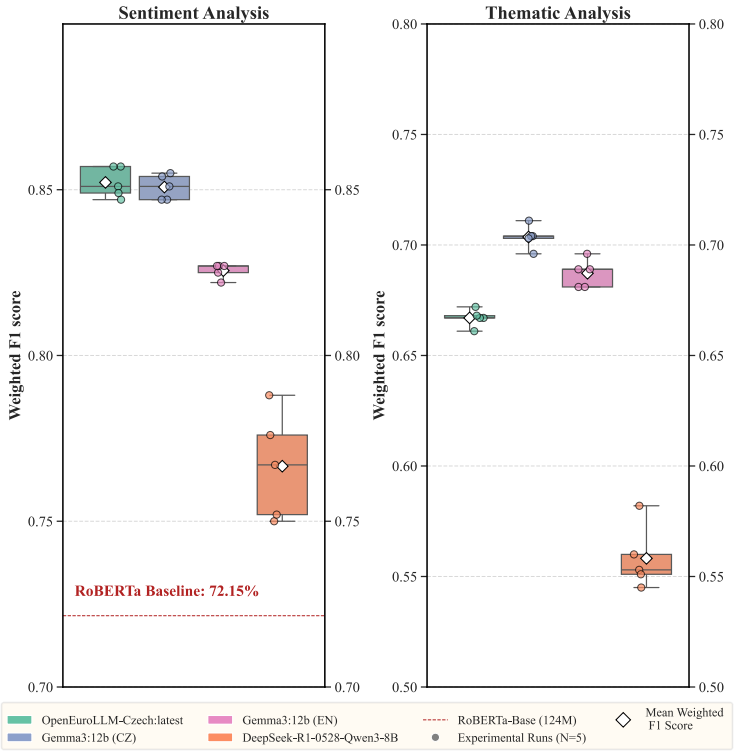

Figure 2. Accuracy comparisons of different LLMs.

The detailed predictions of the models are represented by the confusion matrices in Fig.3, which depicts the results from the run that achieved the highest weighted F1 scores for each model and task.

Task	Model	Weighted F1 score run 1	Weighted F1 score run 2	Weighted F1 score run 3	Weighted F1 score run 4	Weighted F1 score run 5	Overall mean weighted F1 score	Std. Dev.	Avg. not Processed N
Sentiment analysis	OpenEuroLLM-Czech:12b (CZ)	0.851	0.857	0.847	0.857	0.849	0.852	0.004	0.6
	Gemma3:12b (CZ)	0.855	0.847	0.851	0.854	0.847	0.851	0.003	2.6
	Gemma3:12b (EN)	0.825	0.827	0.827	0.827	0.822	0.826	0.001	1.2
	DeepSeek-R1-0528-Qwen3-8B (EN)	0.750	0.767	0.752	0.788	0.776	0.767	0.014	32.8
Thematic analysis	OpenEuroLLM-Czech:12b (CZ)	0.667	0.672	0.667	0.668	0.661	0.667	0.003	2.4
	Gemma3:12b (CZ)	0.696	0.704	0.704	0.711	0.703	0.704	0.004	1.4
	Gemma3:12b (EN)	0.681	0.681	0.696	0.689	0.689	0.687	0.005	1.2
	DeepSeek-R1-0528-Qwen3-8B (EN)	0.553	0.560	0.582	0.545	0.551	0.558	0.012	19.6

Table 2. Comparative model performance.

Sentiment analysis (Fig.3). *OpenEuroLLM-Czech*, *Gemma3 (CZ)*, and *Gemma3 (EN)* models consistently demonstrate robust performance in identifying negative sentiments in the citizens comments. For instance, *Gemma3 (CZ)* model correctly classified 411 comments as negative. While these models generally show strong diagonal concentrations across all sentiment classes, there are some notable misclassifications. There is a tendency to misclassify ground truth negative sentiments as neutral. Despite their overall strength in identifying a large volume of negative sentiments, all three models reveal this error pattern. *OpenEuroLLM-Czech* 44 instances, *Gemma3 (CZ)* 38 instances, *Gemma3 (EN)* 51 instances and *DeepSeek-R1-0528-Qwen3-8B* 79. This indicates a challenge not with detecting strong negativity, but rather with interpreting more subtle expressions of negative situations. An example of this is the comment: "*Probably a housing estate classic...but driving through in the morning is quite an adrenaline rush"*. While human labellers identified this as negative, likely due to the implied stressful or unpleasant experience, the first two models classified it as neutral. The models reasoning elaborated: "*While 'adrenaline rush' can sometimes be positive, in the context of 'driving through' a 'housing estate', it implies a stressful or potentially unsafe situation, but isn't explicitly positive or negative. Therefore, neutral.*" This case demonstrates both models' difficulty with irony, which can be easily identified by a human. Another area of confusion for the models lies between neutral and positive sentiments. For instance, the *OpenEuroLLM-Czech* model misclassified 25 ground truth neutral comments as positive and 21 ground truth positive comments as neutral. This shows a challenge in sentiment analysis where language can be ambiguous. For the comment "*More trees-and let them grow"*, which human labellers annotated as neutral, all models classified it as positive with their reasoning mentioning "*The comment expresses a desire for more trees and asks for them to 'grow', which is a positive suggestion."*

Thematic analysis (Fig.3). A notable success across all models, is the relatively high accuracy in classifying comments related to category 8, i.e., *"there is often an overflowing waste bin or collection point for municipal/sorted waste"*. This is represented by the diagonal entries for ground truth 8, with 83 instances correctly classified by *OpenEuroLLM-Czech, Gemma3 (EN)*, and *DeepSeek-R1-0528-Qwen3-8B,* and 85 by *Gemma3 (CZ)*. On the other hand, the models struggled in classifying category 1, i.e., "*free time*", and category 2, i.e., "*places for visitors*". *OpenEuroLLM-Czech* misclassified 19 comments, while *Gemma3 (CZ)* and *Gemma3 (EN)* showed similar patterns with 13 and 32 such misclassifications, respectively. This overlap is plausible, as a location where one spends free time is often the kind of place one would consider showing to a visitor. An example of this confusion is the comment: "*This is where we'll go jogging tonight."* Human labellers assigned this to category 2; however, the model assigned it to category 1, with the reasoning: *"The comment describes an activity that would be done for 'free time' purposes."* This demonstrates the models' difficulty in distinguishing these closely related thematic categories and is a limitation of the initial participatory mapping design.

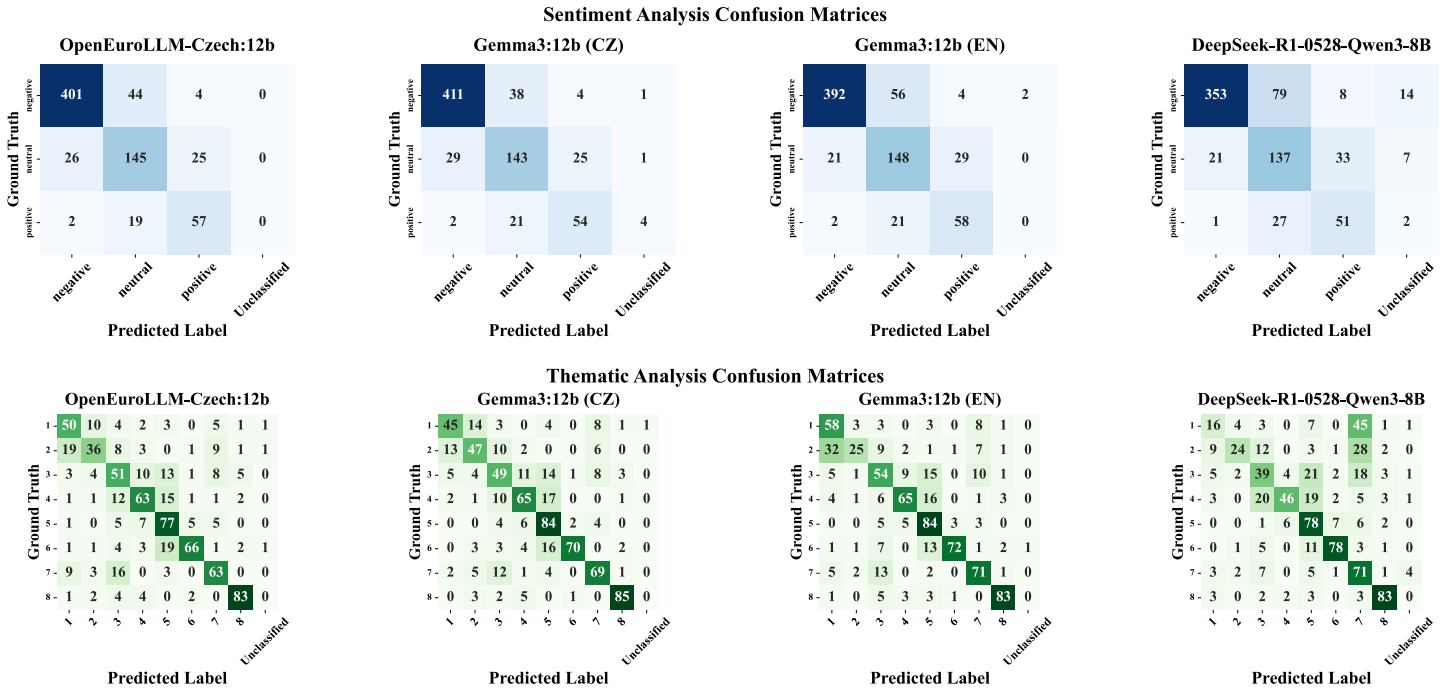

Figure 3. Confusion matrices.

Discussion and conclusion

This study shows how LLMs can support the analysis of citizen-generated comments from participatory mapping initiatives, mobilising free-text feedback that has often remained untapped. Our approach is valuable for practitioners who process large participatory data, as it offers a first outlook on a scalable method for drawing meaningful information based on textual contributions.

Models like *Gemma3* have strong built-in multilingual capabilities and our results indicate that comment classification across languages, specifically between original Czech and the automatically translated English version, achieved stable performance, even comparable to the Czech-tuned model. We acknowledge that the English data in our study comprises machine translations, not independently authored native English comments, thus providing a controlled comparison. These findings therefore demonstrate the model's linguistic capabilities when applied to translated textual content. This could create an outlook for using such models to effectively process naturally multilingual input, presenting new opportunities in designing participatory platforms that offer citizens a choice of language, and therefore, potentially broaden the inclusivity of these initiatives.

Performing sentiment analysis on citizens' comments could support the identification of sentiment and disagreement hotspots that could be used to prioritise topics and locations in follow-up engagement. Regarding the thematic analysis, the accuracy of the models appears capped at 70%, suggesting that fully automated classification is not yet reliable but indicates some potential. We tested a confidence thresholding as a filtering mechanism (dropping out low confidence outputs), but the resulting accuracy gains were modest relative to the number of comments discarded (Fig.4).

If thematic analysis accuracy increases, it means that participatory platforms could accept more spontaneous contribution moving beyond predefined thematic participatory mapping categories. This new direction would allow a better experience for the citizens but also enable the emergence of bottom-up thematics.

However, the use of these models raises concerns about consent, governance and ethical implications. Even with locally run small LLMs, questions persist regarding participant consent for AI processing of their contributions. Furthermore, establishing strong governance framework is important to ensure accountability for insights from LLMs. This is especially true when these insights inform sensitive urban planning decisions that affect real communities. This ethical obligation is closely tied to addressing practical issues like bias, inconsistencies, and hallucinated content (Jiang et al., 2024; Zheng et al., 2025). LLMs trained on large datasets, can accidentally reinforce societal biases. They might misinterpret subtle language from various citizen groups, leading to an inaccurate portrayal of public opinion. Without careful human validation, they could also show inconsistent classifications. Future research should assess model limitations in real participatory settings and create strong ethical recommendations for responsible LLM integration.

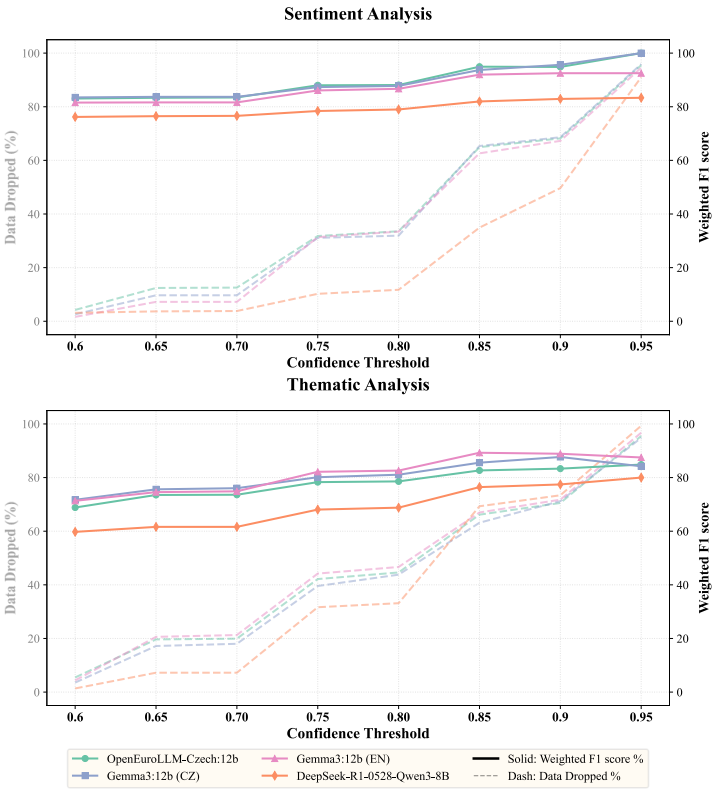

Figure 4. Confidence threshold and comments dropped per task and model.

Declaration of Generative AI in writing

The authors declare that they have used Generative AI tools in the preparation of this manuscript. Specifically, the AI tools were utilised for improving grammar and sentence structure, but not for generating scientific content, research data, or substantive conclusions. All intellectual and creative work, including the analysis and interpretation of data, is original and has been conducted by the authors without AI assistance.

References

Atzmanstorfer, K., Resl, R., Eitzinger, A., Izurieta, X.: The GeoCitizen-approach: community-based spatial planning - an Ecuadorian case study, Cartogr. Geogr. Inf. Sci. 41, 248–259, https://doi.org/10.1080/15230406.2014.890546, 2014

Baer, M.F., Purves, R.S.: Window Expeditions: A playful approach to crowdsourcing natural language descriptions of everyday lived landscapes, Appl. Geogr. 148, 102802, https://doi.org/10.1016/j.apgeog.2022.102802, 2022

Brown, G., Kyttä, M.: Key issues and research priorities for public participation GIS (PPGIS): A synthesis based on empirical research, Appl. Geogr. 46, 122–136, https://doi.org/10.1016/j.apgeog.2013.11.004, 2014

Fagerholm, N., Raymond, C.M., Olafsson, A.S., Brown, G., Rinne, T., Hasanzadeh, K., Broberg, A., Kyttä, M.: A methodological framework for analysis of participatory mapping data in research, planning, and management, Int. J. Geogr. Inf. Sci. 35, 1848–1875, https://doi.org/10.1080/13658816.2020.1869747, 2021

Gholamnia, M., Eslamirad, N., Sajadi, P., Zarin, Z., Pilla, F.: Leveraging large language models for citizen-centric urban accessibility analysis: a case study using Airbnb reviews in Dublin, Spat. Inf. Res. 33, https://doi.org/10.1007/s41324-025-00646-9, 2025

Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the multilingual web of data, Web Semant. 11, 63–71, https://doi.org/10.1016/j.websem.2011.09.001, 2012

Guo, D., Yang, D., Zhang, H. et al.: DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature 645, 633–638, https://doi.org/10.1038/s41586-025-09422-z, 2025

Han, J., Zheng, Z., Lu, X.-Z., Chen, K.-Y., Lin, J.-R.: Enhanced earthquake impact analysis based on social media texts via large language model, Int. J. Disaster Risk Reduct, 109, 104574, https://doi.org/10.1016/j.ijdrr.2024.104574, 2024

Jiang, Bowen, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J. Su, Camillo Jose Taylor and Dan Roth: A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners, Conference on Empirical Methods in Natural Language Processing, https://doi.org/10.48550/arXiv.2406.11050, 2024.

Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé,A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786doi:10.48550/arXiv.2503.19786, 2025.

Liu, K., Yigitcanlar, T., Mehmood, R., Corchado, J., Fu, X.: Large language models in urban planning: A systematic review and conceptual framework, J. Urban Technol. 1–44, https://doi.org/10.1080/10630732.2025.2556551, 2025

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V.: Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv, https://doi.org/10.48550/arXiv.1907.11692, 2019

Pánek, J., Barvíř, R., Koníček, J., Brlík, M.: The emotional map of Prague - data on what locals think about the Czech capital?, Data Brief 39, 107649, https://doi.org/10.1016/j.dib.2021.107649, 2021

Ramírez Aranda, N., De Waegemaeker, J., Venhorst, V., Leendertse, W., Kerselaers, E., Van de Weghe, N.: Point, polygon, or marker? In search of the best geographic entity for mapping cultural ecosystem services using the online public participation geographic information systems tool, “My Green Place,” Cartogr. Geogr. Inf. Sci. 48, 491–511, https://doi.org/10.1080/15230406.2021.1949392, 2021

Raymond, C.M., Nummi, P., von Wirth, T., Poom, A., Ahdekivi, A., Barthel, S., Delmelle, E., Dunkel, E., Fagerholm, N., Grêt-Regamey, A., Hallikainen, F., Heinilä, A., Käyhkö, J., Kotavaara, O., Kyttä, M., Magyar, M., Pesola, A.J., McPhearson, T., Mustafa, A., Nurminen, V., Ramezani, S., Reed, P., Rinne, T., Schipperijn, J., Soininen, N., Toivonen, T., Venuti, F.: Uses, opportunities and risks of artificial intelligence in participatory urban planning, Disc. Cities 2, https://doi.org/10.1007/s44327-025-00137-4, 2025

Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108, https://arxiv.org/pdf/1910.01108, 2019

Shingleton, J., Basiri, A.: How close is “close”? An analysis of the spatial characteristics of perceived proximity using Large Language

Models, AGILE GIScience Ser. 6, 1–14, https://doi.org/10.5194/agile-giss-6-11-2025, 2025

Vallejo-Velázquez, M., Kounadi, O., Pődör, A.: From individuals to collective spatial truth: data characteristics in digital participatory mapping, Adv. Cartogr. GIScience Int. Cartogr. Assoc. 5, 1–9, https://doi.org/10.5194/ica-adv-5-33-2025, 2025

Wang, J., Jiang, R., Yang, C., Wu, Z., Onizuka, M., Shibasaki, R., Koshizuka, N., Xiao, C.: Large Language Models as urban residents: An LLM agent framework for personal mobility generation, arXiv, https://doi.org/10.48550/arXiv.2402.14744, 2024

Zheng, Y., Xu, F., Lin, Y., Santi P., Ratti C., Wang Qi R., Li Y.: Urban planning in the era of large language models, nat Comput Sci 5, 727–736, https://doi.org/10.1038/s43588-025-00846-1, 2025

Appendix

The figure that follows, is the example of the prompt used for sentiment analysis task in English comments. The prompt in Czech for sentiment analysis and the prompts in English and Czech for thematic analysis task can be found in the provided repository.

Prompt in English for sentiment analysis task

system_prompt = """
You are a sentiment analyser system of Public Participation Geographic Information Systems (PPGIS) data.
The data is extracted from citizen comments about the city of Prague, translated from Czech to English.
Your task is to analyse and predict the sentiment of the comments, as well as provide a confidence score of the prediction and your reasoning for it.

1. SENTIMENT FORMATTING RULES
- If a comment expresses no clear sentiment, assign a neutral sentiment score.
- Do not invent or lie if you do not know the answer; say None if you do not know.
- Note that the comments may have street names and variability of PPGIS data.

2. OUTPUT FORMAT
The output should be a list of sentiment objects, where each input comment is given a sentiment analysis. The fields of a sentiment object are:
- comment_id: The ID of the input comment to analyse.
- sentiment_score: The numerical sentiment score is -1 for negative sentiment, 1 for positive sentiment, 0 for neutral sentiment, or None if no sentiment can be assigned.
- confidence_score: The confidence score of your prediction.
- reasoning: A reasoning of why you assigned this sentiment to this comment.

3. FEW-SHOT EXAMPLES (Do not confuse these with your real task)
Input: [
{"comment_id": 1, "comment": "The new pedestrian zone is absolutely wonderful, such a joy to walk through!"},
{"comment_id": 2, "comment": "The old bus routes were much better, this new schedule is a complete disaster for commuters."},
{"comment_id": 3, "comment": "They are planning to build a new roundabout."}
]

Output: {
"sentiments": [
{
"comment_id": 1,
"sentiment_score": 1,
"confidence_score": 0.9,
"reasoning": "The words 'absolutely' and 'wonderful' indicate a strong positive sentiment, so the comment is considered to be positive."
},
{
"comment_id": 2,
"sentiment_score": -1,
"confidence_score": 0.9,
"reasoning": "Phrases 'old bus routes were much better' and a 'complete disaster' clearly point to strong negative sentiment regarding the new bus schedule."
},
{
"comment_id": 3,
"sentiment_score": 0,
"confidence_score": 0.8,
"reasoning": "The comment simply states a factual plan without expressing any opinion or emotion."
}
]
}

END OF EXAMPLES
Now, analyze the following comments based strictly on the text provided below:
"""