ShefFriday at SemEval-2026 Task 9: LLM-Based Annotation Methods for Detecting Multilingual, Multicultural and Multievent Online Polarisation

Owen Cook*, **Meredith Gibbons***, **Xingyi Song**
School of Computer Science, The University of Sheffield, UK
{oscook1, magibbons1, x.song}@sheffield.ac.uk

Abstract

This paper presents our findings for SemEval-2026 Task 9. We submit to all three subtasks using an LLM-as-an-annotator strategy, simulating the data annotation process with large language models. We created 30 LLM annotators using persona injection (also known as sociodemographic prompting) and experimented with various annotation aggregation methods, including Dawid-Skene and MACE. To further increase the variability in annotator responses, we used the hatefulness detection task as proxy for identifying polarisation. Our findings indicate that this reframing of the problem is effective for the binary classification of polarisation, but is less effective for finer-grained polarisation detection. For subtasks 2 and 3, majority voting yielded the best overall performance. While our unsupervised approach does not rank as highly as supervised ones, this work provides insight into the utility of persona-based prompting and the issue of LLM annotators exhibiting high intra-model agreement.

1 Introduction

Polarisation, defined as "the increasing extremity of opinions, beliefs, or behaviors, resulting in heightened intergroup divisions and conflict", has become a growing problem in online conversations (Naseem et al., 2026b). Early detection is crucial in order to promote more positive discourse online. This paper describes our submission for SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online polarisation (Naseem et al., 2026a). We address all three subtasks, **Subtask 1:** Polarisation Detection, **Subtask 2:** Polarisation Type Classification and **Subtask 3:** Manifestation Identification.

The dataset for this task was very large, covering 22 languages for subtasks 1 and 2 and 18 languages for subtask 3. The overall label distribution is almost evenly split between polarised (53%) and non-polarised (47%) samples. For some languages the label distribution deviated significantly from the average. For example, the Khmer language dataset contained 91% polarised samples, while the Hausa language dataset contained only 11% polarised samples. Full label distributions can be found in Naseem et al. (2026b).

We approach this task using an LLM-as-an-annotator strategy, leveraging large language models to simulate the mass annotation process. In our experiment, we generated 30 LLM annotators using 3 base models (Gemma-3 27B (Gemma-Team et al., 2025), Qwen 2.5 72B (Qwen et al., 2025), and Llama 3.3 70B (Grattafiori et al., 2024)) and persona injection (Deshpande et al., 2023; Santurkar et al., 2023) (10 personas per model). The annotators were generated with diverse backgrounds and beliefs across five categories: political affiliation, race/ethnicity, religion, gender, and sexual orientation. To derive the final labels, we simulated the annotation process by applying disagreement aggregation strategies, including Dawid-Skene (Dawid and Skene, 1979) and MACE (Hovy et al., 2013).

We also experiment with the LLMs annotating each example for hatefulness as a proxy for identifying polarisation; an example is identified as polarised when a defined threshold of disagreement among LLM annotators is met.

2 Related Work

2.1 Large Language Models as Annotators

Annotation using LLMs. Large Language Models (LLMs) are increasingly suitable for automated annotation tasks, possessing a high level of language understanding and the ability to process unconstrained textual inputs (Naveed et al., 2025). Since their introduction, zero-shot, few-shot, and fine-tuned LLMs have been used to automate the laborious task of human annotation (He et al., 2024), with Törnberg (2025) reporting that LLMs can out-

*These authors contributed equally.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3297–3309
July 3-4, 2026 ©2026 Association for Computational Linguistics

perform human experts in objective labelling tasks. Much of the literature, however, does not conclude that LLMs outperform humans (Ollion et al., 2023; Thapa et al., 2023; Reiss, 2023), with language models performing worse as the task difficulty increases (Ding et al., 2024) or when the task requires subjective interpretation. It is difficult to determine the true effectiveness of language models on the annotation task, as the performance metric will often be the agreement with a "gold-standard" human judgment, which is not always correct itself (Alm, 2011; Aroyo and Welty, 2015). Despite this issue, it is often reported that introducing LLMs into the annotation pipeline to collaborate with human labellers can improve both accuracy and efficiency (Li, 2024; Yuan et al., 2025).

LLM Performance. LLMs generally exhibit performance levels proportional to the number of tokens trained on and the number of model parameters (Kaplan et al., 2020; Hoffmann et al., 2022). Chain-of-thought prompting (Wei et al., 2023) can offer improvements in model accuracy on tasks involving reasoning (Wei et al., 2022) but they also come with significant inference costs; API access for OpenAI's reasoning o1 model was roughly 6x more expensive than GPT-4o (Smolaks, 2025). When scaling up for larger-scale annotation tasks, this significantly increases computational requirements.

Non-thinking models are still competitive, however, with models in the Llama 3 (Grattafiori et al., 2024), Qwen 2.5 (Qwen et al., 2025), and Gemma 3 (Gemma-Team et al., 2025) families being high-performing for their parameter size. Libraries such as Outlines (Willard and Louf, 2023) and Guidance (Lundberg et al., 2022) are also useful for constraining the output of non-thinking models, forcing the correct output format and avoiding unnecessary token inference.

Quantisation is a popular and high-performing technique used to fit larger models on limited hardware resources (Frantar et al., 2022). High-parameter quantised models tend to outperform full-precision lower-parameter models within the same family (Dettmers and Zettlemoyer, 2023).

2.2 Modelling Annotator Disagreement

Demographic factors of annotators have been found to have a significant impact on their perception of multiple Natural Language Processing tasks, including offensiveness, politeness, and hate speech detection (Pei and Jurgens, 2023; Kumar et al., 2021). Building on this foundation, sociodemographic prompting (also called persona prompting) assigns a large language model a specific "persona" or set of characteristics to emulate when making a classification decision (Deshpande et al., 2023; Santurkar et al., 2023). This is particularly impactful for subjective tasks such as polarisation detection and hatefulness detection, where annotator disagreement is generally higher.

3 System Overview

Our approach treats the polarisation task as an annotation task, automated using LLMs. This automated annotation pipeline is shown in Figure 1.

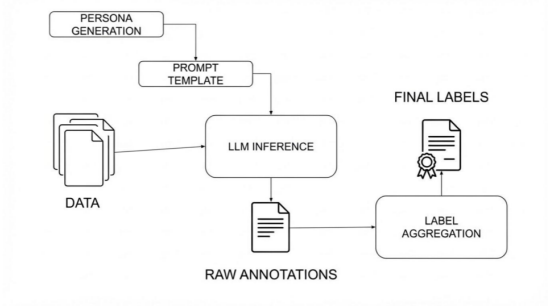


Figure 1: Visualised automated LLM annotation pipeline.

3.1 Annotation Task & Personas

For the annotation task, we adopt three approaches:

- Collect the base annotation from each LLM for whether the text is polarising;
- Generate 10 personas per model and collect responses from each persona-model combination for whether the text is polarising (for subtask 2, we ask whether the post is characteristically polarising – e.g. "politically polarising");
- Generate 10 personas per model and collect responses from each persona-model combination for whether the text is hateful (for subtask 2, we ask whether the post is characteristically hateful – e.g. "politically hateful").

Polarisation classification is subjective. As the task of polarisation detection is subjective, the true label for each sample can be represented as a distribution across the "polarised" and "not polarised" labels, rather than a single binary label. We hypothesise that a diverse set of annotators will better represent the ground truth distribution, so we use

sociodemographic/persona prompting to simulate annotators with different backgrounds and beliefs. These annotator personas were generated based on the label categories in subtask 2: political affiliation, race/ethnicity, religion, gender, and sexuality. For more details on persona generation, see Appendix B.

Reframing as hatefulness detection. Another approach to the subjectivity of detecting polarisation was to reframe the problem as one of hatefulness detection. We theorise that annotator disagreement in the hatefulness detection task will be a strong indicator that a post is polarising and may also increase the impact of the sociodemographic prompting; hatefulness detection is likely to be a more personally subjective task than classifying text as "polarised" or "not polarised". This theory is supported by the observation that agreement among LLM annotators is lower for hatefulness annotation than polarisation annotation (see Table 2). Subtask 3, Manifestation Identification, is not as well suited to the hatefulness detection task, so we use only base-LLM and persona-based polarisation annotation.

3.2 Label Aggregation

After generating the LLM annotations, we aggregate the labels per data point to obtain our final label for each piece of text. For the annotations where the LLM is tasked with annotating for polarisation, we aggregate with commonly used aggregation methods seen in crowdsourcing such as Majority Vote, Dawid-Skene (Dawid and Skene, 1979), and MACE (Hovy et al., 2013). We compare the performance of these methods when aggregating each language independently and labels from all languages together.

Aggregating hatefulness annotations. The hatefulness annotations could not be aggregated directly to obtain a polarisation classification, so we utilise an alternative method. A sample is classified as "polarised" if the annotator disagreement exceeded a given threshold. For example, with a 0.1 threshold, a data point would be considered polarised if more than 10% of the annotators disagreed with the majority rating. While we set thresholds manually in these experiments, learning the threshold could be explored in future work.

Aggregating base polarisation and hatefulness annotations. To combine base model polarisation and hatefulness annotations (`base + persona`, or `bp`), we first converted the base polarisation annotation into a hatefulness annotation. A base annotation of "not polarised" is converted to an instance of the majority annotation, while a base annotation of "polarised" is converted to an instance of the minority annotation. For example, if only 8% of the persona annotations for a data point are "not hateful", a base annotation of "polarised" would be converted to "not hateful", while a base annotation of "not polarised" would be converted to "hateful". The result of this is that the conversion of a "polarised" rating reduces the annotator agreement, while the conversion of a "not polarised" rating increases the annotator agreement.

After converting the base polarisation annotation to a hatefulness annotation, the annotations are aggregated using the same method as the hatefulness annotations.

4 Experimental Setup

All experiments are run on 2x40GB NVIDIA A100 graphics cards. We use vLLM (Kwon et al., 2023) to increase inference speed with its use of paged attention and optimisation of multi-card processing.

The hardware constraints play a factor in our decision of pre-trained LLMs. We choose Gemma-3 27B (Gemma-Team et al., 2025), loaded at full 16-bit precision as the largest of the Gemma 3 family of models. We then choose Qwen 2.5 72B (Qwen et al., 2025) and Llama 3.3 70B (Grattafiori et al., 2024) as high-performing models that can run inference using 80GB of VRAM when loaded at 4-bit precision. As the polarisation detection task is one of closed-set classification, we use Outlines (compatible with vLLM model instances) to constrain the output of the annotation prompt to ensure "yes" or "no" answers. We treat multi-label subtasks 2 and 3 as a series of independent binary classification tasks.

Prompting strategies. Following our three prompting strategies in Section 3.1, we devise three prompt templates for subtask 1, three for subtask 2, and then two for subtask 3; we do not annotate for hate speech detection in subtask 3. Example prompts for each subtask can be seen in Appendix A.

To run each combination of subtask, model, and prompting strategy, we use yaml configuration files (in the `configs/` directory) with the `annotate.py`

script in our codebase [1].

Label aggregation. To aggregate the annotations with classical crowdsourcing methods, we use the Crowd-Kit (Ustalov et al., 2021) implementations of Majority Vote, Dawid-Skene, and MACE. For our more disagreement-sensitive aggregation method discussed in Section 3.2, we use the implementation in `post-processing/` within our codebase.

To decide which method to run on the test set, we evaluate each combination of prompting approach and label aggregation method for each subtask on the development set, using the macro-F1 score – the task performance metric.

5 Results

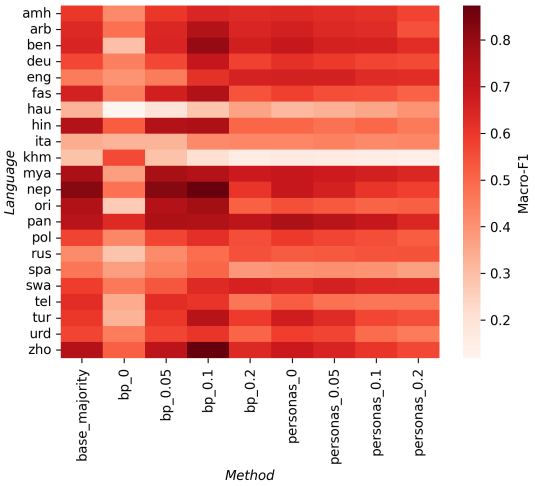


Figure 2: Comparison of annotator aggregation methods on the development set for subtask 1.

Comparing label aggregation methods for polarisation labels. Our base performance benchmark was a majority-vote ensemble of the three LLMs ("base_majority" in Figures 2 and 3). Using different aggregation methods for the three base responses yielded no significant improvement in macro-F1 score across all subtasks. With the persona polarisation annotations, Dawid-Skene improved the overall macro-F1 score by 3.4% on subtask 1 over base majority vote; due to time constraints this was not incorporated into our final solution. Outside of subtask 1, this method of persona polarisation prompting performed comparably to the baseline system. As we only compare

[1]https://github.com/Sheffield-Friday/SemEval2026_Polarisation

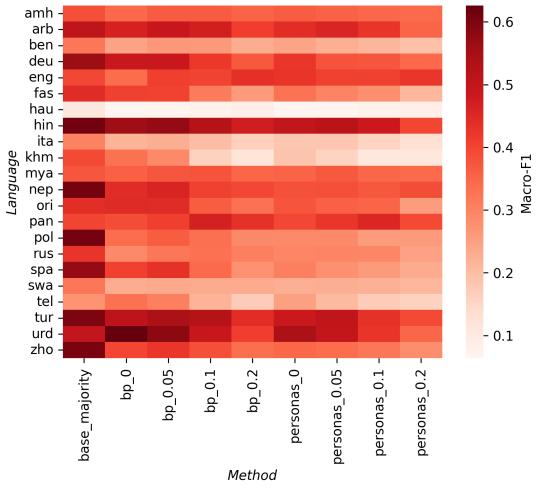


Figure 3: Comparison of annotator aggregation methods on the development set for subtask 2.

this persona prompting method against the baseline for subtask 3 and the improvements are not significant, we use the base majority vote method in our subtask 3 submission. The results from these aggregations (for each aggregation method, we take the highest average macro-F1 between aggregating by language and across all languages) can be seen in Table 1.

ST	Base			Personas (polarisation)		
	MV	DS	MACE	MV	DS	MACE
1	0.626	0.626	0.628	0.645	**0.660**	0.650
2	**0.644**	**0.644**	0.631	0.639	0.633	0.631
3	0.595	0.595	0.585	0.595	0.592	**0.597**

Table 1: Macro-F1 scores after aggregating base models and base models + personas with Majority Vote (MV), Dawid-Skene (DS), and MACE. The score reported for each aggregation is the highest average macro-F1 score recorded between aggregating by language and across all languages together.

Performance of hatefulness annotations. We then considered the annotations for hatefulness and implemented our alternative aggregation methods for handling small levels of disagreement within the LLM-annotated data. For the `persona_x` and `bp_x` methods, we classifed text as polarised if the proportion of annotations of the minority label exceeded a given threshold (`x`). For example, if 15% of annotations were "hateful" and 85% "not hateful", the data point would be classified as "polarised" when `x=0.1` but "not polarised" when `x=0.2`.

We experimented with different values for the

threshold, as shown in Figure 2 and Figure 3. The best performing methods were `bp_0.1` and `base_majority`. Above the 0.1 threshold, performance for the `bp_x` methods sharply decreased. For the `personas_x` methods, performance decreased for any threshold above 0; when using exclusively persona annotators, it was most effective to consider a data point polarised unless all annotators agreed. These results suggest that persona-based prompting does not result in annotations as varied as we had expected.

Hatefulness annotations were not as effective in subtask 2. The results for subtask 2 are shown in Figure 3. After applying the same methods as for subtask 1, we identified `base_majority` as the most effective method. It is notable that the persona-based methods performed worse on this subtask than subtask 1, considering that the personas directly correspond to the polarisation types. It is relatively unclear why this drop in performance is observed as the agreement within models (see Table 2) is lower for subtask 2, indicating that there should be more disagreement among personas – seemingly aiding the hatefulness annotation approach. In subtask 2, the hatefulness persona annotation approach only outperformed the base models on the "other" label. This could either indicate that the approach is more useful in more subjective or broader task definitions; it could also be that the definitions provided in the prompt for specific types of polarisation were particularly beneficial for the LLMs. Specific definitions were not provided when the LLMs annotated for hatefulness.

Impact of persona-based annotation. From the LLM inter-annotator agreement in Table 2, it is clear why the personas do not offer too much improvement (if any) over the baseline ensemble performance. The intra-model agreement (agreement between different personas of the same model) is much higher than inter-model agreement. An interesting point to note is that the more subjective task, hatefulness annotation, has a lower agreement than polarisation annotation (base model and persona). This increase in persona impact with hatefulness annotations may explain the improved performance on subtask 1. It may be the case that 10 personas per model is not enough to yield maximum improvement. Future work may explore increasing the number of personas and more advanced methods of stratifying which persona annotations to aggregate for a given data point.

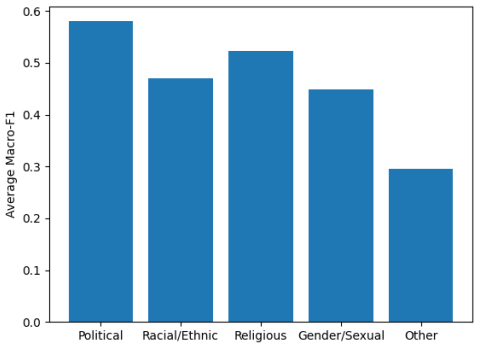


Figure 4: Per-label performance across all languages for subtask 2.

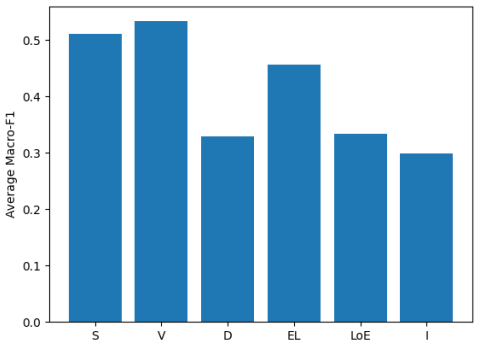


Figure 5: Per-label performance across all languages for subtask 3 – stereotype (S), vilification (V), dehumanization (D), extreme language (EL), lack of empathy (LoE) and invalidation (I).

5.1 Test Set Results

Appendix D shows our full results on the test set. Our system is not ranked highly for subtask 1, however it was expected that our unsupervised approach would not perform as well as supervised methods. Our five highest-performing languages were Burmese, Nepali, Bengali, English, and Punjabi. While English is a high-resource language, the other four were initially unexpected. However, these four languages were high-performing for the polarisation detection task across many teams, so the high performance is likely a result of the specific dataset used, rather than our method.

For subtask 2, the performance of our system varied in comparison to other submissions; we were highly ranked in the German and Italian languages but lower in others. Our system performed best when predicting "Political" polarisation, with an average macro-F1 of 0.58 across all languages, and

	Base	Polarising				Hateful			
		All	Llama	Qwen	Gemma	All	Llama	Qwen	Gemma
Subtask 1	0.707	0.707	0.894	0.893	0.894	0.657	0.879	0.877	0.912
Subtask 2	0.596	0.646	0.888	0.897	0.882	0.525	0.744	0.803	0.821

Table 2: Overall agreement scores (Krippendorff's alpha (Krippendorff, 1970)) for base models, personas annotating for polarisation, and personas annotating for hatefulness on subtasks 1 and 2. For subtask 2 (multi-label), agreement is calculated per label (political, racial/ethnic, etc.) and then averaged over labels.

worst when predicting "Other" polarisation, with an average macro-F1 of 0.30 across all languages.

For subtask 3, as with subtask 2, the performance of our system varied in comparison to other submissions, and we were again highly ranked in the German language. Our system performed best when predicting "Vilification", with an average macro-F1 of 0.53 across all languages, and worst when predicting "Invalidation", with an average macro-F1 of 0.30 across all languages.

We also considered that our results could be influenced by the characteristics of the dataset, particularly the inter-annotator agreement and the language family. However, neither of these factors significantly affected the result – see Appendix C for the results of our analysis.

When comparing our results with the label distribution (see Figure 6), however, it is clear that stark label imbalance is associated with low performance.

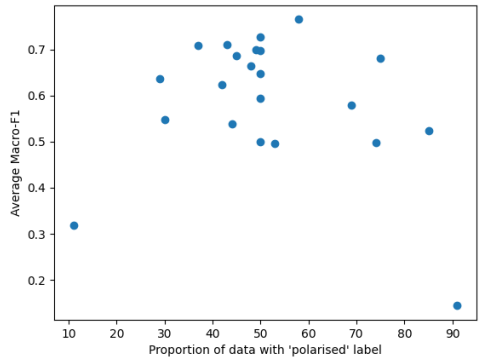


Figure 6: Comparison of model performance with dataset label balance for subtask 1.

6 Conclusion

We have applied large language models as unsupervised annotators to the polarisation detection task, focusing on sociodemographic/persona prompting. While overall our results are not highly ranked on the leaderboard, we provide evidence that while sociodemographic prompting can improve model performance, agreement among personas of the same model remains high. As the subjectivity of the task increases (such as annotating for hatefulness over polarisation), so too does the disagreement among personas; this disagreement can then be utilised as a signal.

While it is not unusual that an unsupervised approach would exhibit lower performance than supervised methods, this work provides more evidence that zero-shot LLM annotation methods should not yet to be used in place of real human annotators. Our analysis shows that LLMs do not perform well with highly imbalanced data and our results seem to be independent of language family and inter-annotator agreement.

7 Ethical Considerations

Beck et al. 2024 outline several issues with sociodemographic prompting and the use of large language models as annotators. It is important to emphasise that the outputs of sociodemographic/persona prompting should not be regarded as statements on the opinions or behaviour of any given individual or group.

Acknowledgments

This work was supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Meredith Gibbons is partially supported by Ofcom; this paper represents the views and opinions of the author and should not be taken as a statement of Ofcom opinion.

References

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI magazine*, 36(1):15–24.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 1236–1270, Singapore. Association for Computational Linguistics.

Tim Dettmers and Luke Zettlemoyer. 2023. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning*, pages 7750–7774. PMLR.

Mucong Ding, Chenghao Deng, Jocelyn Choo, Zichu Wu, Aakriti Agrawal, Avi Schwarzschild, Tianyi Zhou, Tom Goldstein, John Langford, Animashree Anandkumar, and 1 others. 2024. Easy2hard-bench: Standardized difficulty labels for profiling llm performance and generalization. *Advances in Neural Information Processing Systems*, 37:44323–44365.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Gemma-Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. Annollm: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, and 3 others. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, 30(1):61–70.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. (arXiv:2106.04511). ArXiv:2106.04511 [cs].

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jiyi Li. 2024. A comparative study on annotation quality of crowdsourcing and llm via label aggregation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529. IEEE.

Scott Lundberg, Harsha Nori, and Marco Tulio Ribeiro. 2022. Guidance: Control lm outputs. reduce latency and cost. `https://github.com/guidance-ai/guidance`.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acarürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Etienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2023. Chatgpt for text annotation? mind the hype. *SocArXiv preprint*, page 32.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the popquorn dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, page 252–265, Toronto, Canada. Association for Computational Linguistics.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Michael V Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th international conference on machine learning*, volume 202 of *Proceedings of machine learning research*, page 29971–30004. PMLR.

Max Smolaks. 2025. 'reasoning' will increase the infrastructure footprint of ai. Accessed: February 18, 2026.

Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Petter Törnberg. 2025. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 43(6):1181–1195.

Dmitry Ustalov, Nikita Pavlichenko, Vladimir Losev, Iulian Giliazev, and Evgeny Tulin. 2021. A general-purpose crowdsourcing computational quality control toolkit for python. In *The Ninth AAAI Conference on Human Computation and Crowdsourcing: Works-in-Progress and Demonstration Track*, HCOMP 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.

Mingyue Yuan, Jieshan Chen, Zhenchang Xing, Gelareh Mohammadi, and Aaron Quigley. 2025. A case study of scalable content annotation using multi-llm consensus and human review. *arXiv preprint arXiv:2503.17620*.

A LLM Prompts

A.1 Subtask 1: Binary Polarisation

A.1.1 Base Prompting

System Prompt: You are a highly accurate polarization detection system. You will be provided with a piece of text in the {LANGUAGE} language and you must respond in English with either 'yes' or 'no' as to whether the provided text is polarizing. You must respond ONLY 'yes' or 'no' and terminate your response immediately after with the end of generation token.

User Prompt: Input text: {INPUT_TEXT}
Is polarizing (yes/no):

A.1.2 Persona Prompting (polarisation)

System Prompt: You are a {POLITICAL STANCE} {RACE/ETHNICITY} {RELIGION} {SEXUAL ORIENTATION} {GENDER}. You will be provided with a piece of text in the LANGUAGE language and you must respond in English with either 'yes' or 'no' as to whether the provided text is polarizing. You must respond ONLY 'yes' or 'no' and terminate your response immediately after with the end of generation token.

User Prompt: Input text: {INPUT_TEXT}
Is polarizing (yes/no):

A.1.3 Persona Prompting (hatefulness)

System Prompt: You are a {POLITICAL STANCE} {RACE/ETHNICITY} {RELIGION} {SEXUAL ORIENTATION} {GENDER}. You will be provided with a piece of text in the LANGUAGE language and you must respond in English with either 'yes' or 'no' as to whether the provided text is hateful. You must respond ONLY 'yes' or 'no' and terminate your response immediately after with the end of generation token.

User Prompt: Input text: {INPUT_TEXT}
Is hateful (yes/no):

A.2 Subtask 2: Polarisation Type Classification

A.2.1 Base Prompting

System Prompt: You are a highly accurate polarization detection system. You will be provided with a piece of text in the {LANGUAGE} language and you must respond in English with either 'yes' or 'no' as to whether the provided text is politically/ideologically polarizing. You must respond ONLY 'yes' or 'no' and terminate your response immediately after with the end of generation token.

Political/ideological polarization guidelines: This type of extremism focuses on division, intolerance, and conflict between political parties and followers. Political polarization refers to political beliefs and affiliations becoming more extreme. People may identify more strongly with their political party, leading to deeper divides and a reduced willingness to compromise. It broadens ideological differences between political groups.

User Prompt: Input text: {INPUT_TEXT}
Is politically/ideologically polarizing (yes/no):

A.2.2 Persona Prompting (polarisation)

System Prompt: You are a {POLITICAL STANCE} {RACE/ETHNICITY} {RELIGION} {SEXUAL ORIENTATION} {GENDER}. You will be provided with a piece of text in the {LANGUAGE} language and you must respond in English with either 'yes' or 'no' as to whether the provided text is politically/ideologically polarizing. You must respond ONLY 'yes' or 'no' and terminate your response immediately after with the end of generation token.

Political/ideological polarization guidelines: This type of extremism focuses on division, intolerance, and conflict between political parties and followers. Political polarization refers to political beliefs and affiliations becoming more extreme. People may identify more strongly with their political party, leading to deeper divides and a reduced willingness to compromise. It broadens ideological differences between political groups.

User Prompt: Input text: {INPUT_TEXT} Is politically/ideologically polarizing (yes/no):

A.2.3 Persona Prompting (hatefulness)

System Prompt: You are a {POLITICAL STANCE} {RACE/ETHNICITY} {RELIGION} {SEXUAL ORIENTATION} {GENDER}. You will be provided with a piece of text in the LANGUAGE language and you must respond in English with either 'yes' or 'no' as to whether the provided text is politically/ideologically hateful, considering that you are politics. You must respond ONLY 'yes' or 'no' and terminate your response immediately after with the end of generation token.

User Prompt: Input text: {INPUT_TEXT} Is politically/ideologically hateful (yes/no):

A.3 Subtask 3: Manifestation Identification

A.3.1 Base Prompting

System Prompt: You are a highly accurate polarization detection system. You will be provided with a piece of text in the {LANGUAGE} language and you must respond in English with either 'yes' or 'no' as to whether the provided text exhibits stereotyping AND is polarizing. You must respond ONLY 'yes' or 'no' and terminate your response immediately after with the end of generation token.

Stereotype guidelines: Stereotype: This manifestation occurs when a message generalizes certain characteristics of individuals to all members of a group, ignoring individual differences. Stereotypes simplify complex personalities into one-size-fits-all representations.

User Prompt: Input text: {INPUT_TEXT} Exhibits stereotyping AND is polarizing (yes/no):

A.3.2 Persona Prompting (polarisation)

System Prompt: You are a {POLITICAL STANCE} {RACE/ETHNICITY} {RELIGION} {SEXUAL ORIENTATION} {GENDER}. You will be provided with a piece of text in the {LANGUAGE} language and you must respond in English with either 'yes' or 'no' as to whether the provided text exhibits stereotyping AND is polarizing. You must respond ONLY 'yes' or 'no' and terminate your response immediately after with the end of generation token.

Stereotype guidelines: Stereotype: This manifestation occurs when a message generalizes certain characteristics of individuals to all members of a group, ignoring individual differences. Stereotypes simplify complex personalities into one-size-fits-all representations.

User Prompt: Input text: {INPUT_TEXT} Exhibits stereotyping AND is polarizing (yes/no):

B Persona Generation

We selected the characteristics for the sociodemographic/persona prompting based on the polarisation categories within subtask 2: political affiliation, race/ethnicity, religion, gender and sexual orientation. Personas were generated for each subtask using these characteristics. To create a relatively even distribution, we generate the first 50% of the personas via random selection for each characteristic. For the other 50% of the personas, for each characteristic we select the value that is rarest in the current set of personas.

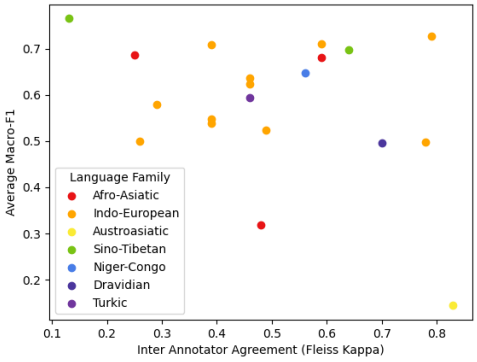
As we only use a small number of unique personas (10) per model, we limit the set of potential values for the characteristics in order to make our sets of generated personas more consistent (see Table 3). This limited set is a limitation of our method, as the generated personas are less diverse. Including a wider range of personas would likely improve the result - future work could examine this effect.

Characteristic	Potential Values
Political Affiliation	left-wing, politically moderate, right-wing
Race/Ethnicity	Black, East Asian, Middle Eastern, Mixed Race, White
Religion	Atheist, Christian, Hindu, Jewish, Muslim
Gender	man, woman
Sexual Orientation	heterosexual, homosexual

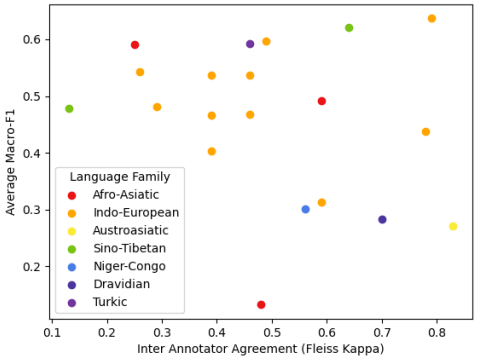
Table 3: Potential values for the persona characteristics.

C Analysis of Inter-Annotator Agreement

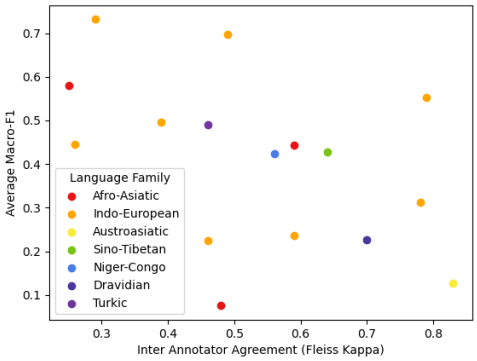
Figure 7 compares our test set results with inter-annotator agreement (reported for subtask 1 only but reused for the subtask 2 and subtask 3 analysis) and language family. Our results show high variability, and there is no clear association between the variables.



(a) Subtask 1



(b) Subtask 2



(c) Subtask 3

Figure 7: Comparison of model performance with dataset inter annotator agreement and language family.

D Full Results

Language	Subtask 1	Subtask 2	Subtask 3
Amharic (amh)	0.6811	0.4916	0.4423
Arabic (arb)	0.6863	0.5910	0.5800
Bengali (ben)	0.7099	0.3128	0.2352
German (deu)	0.6650	0.5950	0.5153
English (eng)	0.7084	0.4660	0.4959
Persian (fas)	0.4977	0.4378	0.3127
Hausa (hau)	0.3179	0.1325	0.0752
Hindi (hin)	0.5236	0.5974	0.6981
Italian (ita)	0.5394	0.5375	N/A
Khmer (khm)	0.1440	0.2719	0.1260
Burmese (mya)	0.7653	0.4780	N/A
Nepali (nep)	0.7267	0.6368	0.5524
Odia (ori)	0.6365	0.4675	0.2240
Punjabi (pan)	0.6994	0.4237	0.3842
Polish (pol)	0.6234	0.5367	N/A
Russian (rus)	0.5487	0.4037	N/A
Spanish (spa)	0.5001	0.5433	0.4456
Swahili (swa)	0.6471	0.3007	0.4243
Telugu (tel)	0.4957	0.2832	0.2260
Turkish (tur)	0.5939	0.5923	0.4904
Urdu (urd)	0.5785	0.4818	0.7316
Chinese (zho)	0.6983	0.6208	0.4283

Table 4: Macro F1 scores for each language and task.

Language	Political	Racial/Ethnic	Religious	Gender/Sexual	Other
Amharic (amh)	0.8041	0.5883	0.3883	0.2632	0.4139
Arabic (arb)	0.7026	0.6160	0.6264	0.5953	0.4149
Bengali (ben)	0.7486	0.0649	0.2690	0.2000	0.2814
German (deu)	0.6815	0.6011	0.6230	0.7333	0.3363
English (eng)	0.6290	0.5503	0.5191	0.4658	0.1659
Persian (fas)	0.7295	0.1818	0.4291	0.4472	0.4014
Hausa (hau)	0.1908	0.1895	0.1643	0.1081	0.0097
Hindi (hin)	0.8888	0.3375	0.9077	0.6463	0.2069
Italian (ita)	0.5933	0.6000	0.7065	0.4654	0.3224
Khmer (khm)	0.4969	0.3143	0.0192	0.3590	0.1699
Burmese (mya)	0.5872	0.3274	0.4103	0.3762	0.6886
Nepali (nep)	0.4972	0.7559	0.8732	0.6935	0.3640
Odia (ori)	0.6138	0.4228	0.6154	0.4337	0.2517
Punjabi (pan)	0.5937	0.3784	0.5217	0.3906	0.2340
Polish (pol)	0.7603	0.5538	0.6829	0.4751	0.2114
Russian (rus)	0.3954	0.5306	0.5000	0.5045	0.0879
Spanish (spa)	0.5783	0.5553	0.6103	0.6311	0.3416
Swahili (swa)	0.0974	0.6516	0.4689	0.1138	0.1715
Telugu (tel)	0.2505	0.2594	0.1667	0.3794	0.3601
Turkish (tur)	0.7655	0.6686	0.7601	0.5737	0.1938
Urdu (urd)	0.8277	0.3732	0.4681	0.1917	0.5486
Chinese (zho)	0.3349	0.8219	0.7838	0.8303	0.3330

Table 5: Macro F1 scores for the individual labels for subtask 2.

Language	S	V	D	EL	LoE	I
Amharic (amh)	0.6541	0.5975	0.3478	0.4671	0.3047	0.2826
Arabic (arb)	0.7364	0.7574	0.5154	0.7096	0.4448	0.3161
Bengali (ben)	0.2177	0.5918	0.3257	0.1606	0.0628	0.0524
German (deu)	0.6774	0.6039	0.4361	0.5011	0.5063	0.3671
English (eng)	0.4567	0.6553	0.4682	0.6226	0.3088	0.4640
Persian (fas)	0.3371	0.4786	0.2083	0.3947	0.2604	0.1972
Hausa (hau)	0.1390	0.0362	0.1056	0.1384	0.0275	0.0048
Hindi (hin)	0.7594	0.7939	0.4591	0.6909	0.7492	0.7358
Khmer (khm)	0.0652	0.1687	0.1524	0.1019	0.2041	0.0636
Nepali (nep)	0.7397	0.6621	0.3548	0.7672	0.2819	0.5085
Odia (ori)	0.3577	0.2939	0.1765	0.3577	0.0728	0.0851
Punjabi (pan)	0.3810	0.4508	0.4138	0.3576	0.3033	0.3988
Spanish (spa)	0.5034	0.6327	0.2726	0.5394	0.4580	0.2674
Swahili (swa)	0.6065	0.4765	0.2301	0.3555	0.4911	0.3861
Telugu (tel)	0.3026	0.2390	0.2000	0.1532	0.2713	0.1899
Turkish (tur)	0.7375	0.6681	0.3439	0.7661	0.2647	0.1619
Urdu (urd)	0.7981	0.7941	0.5647	0.7779	0.7770	0.6780
Chinese (zho)	0.7213	0.6997	0.3567	0.3604	0.2233	0.2084

Table 6: Macro F1 scores for the individual labels for subtask 3 - stereotype (S), vilification (V), dehumanization (D), extreme language (EL), lack of empathy (LoE) and invalidation (I).