ANNALES

Anali za istrske in mediteranske študije
Annali di Studi istriani e mediterranei
Annals for Istrian and Mediterranean Studies
Series Historia et Sociologia, 36, 2026, 2

UDK 009 Annales, Ser. hist. sociol., 36, 2026, 2, pp. 163-336, Koper 2026 ISSN1408-5348

UDK 009

ISSN 1408-5348
e-ISSN 2591-1775

ANNALES

Anali za istrske in mediteranske študije
Annali di Studi istriani e mediterranei
Annals for Istrian and Mediterranean Studies

Series Historia et Sociologia, 36, 2026, 2

KOPER 2026

ISSN 1408-5348
e-ISSN 2591-1775

UDK 009

Letnik 36, leto 2026, številka 2

Glavni urednik/*Redattore capo/ Editor in chief:* Darko Darovec

Odgovorni urednik/*Redattore responsabile/Responsible Editor:* Salvator Žitko

Uredniki/*Redattori/Editors:* Urška Lampe, Boštjan Udovič, Žiga Oman, Veronika Kos

Prevajalka/*Traduttrice/Translator:* Cecilia Furioso Cenci (it.)

Oblikovalec/*Progetto grafico/ Graphic design:* Dušan Podgornik, Darko Darovec

Tisk/*Stampa/Print:* Založništvo PADRE d.o.o.

Založnika/*Editori/Published by:* Zgodovinsko društvo za južno Primorsko - Koper / *Società storica del Litorale - Capodistria*© / Inštitut IRRIS za raziskave, razvoj in strategije družbe, kulture in okolja / *Institute IRRIS for Research, Development and Strategies of Society, Culture and Environment / Istituto IRRIS di ricerca, sviluppo e strategie della società, cultura e ambiente*©

Sedež uredništva/*Sede della redazione/ Address of Editorial Board:* SI-6000 Koper/*Capodistria,* Garibaldijeva/*Via Garibaldi 18*
e-mail: annaleszdjp@gmail.com, ***internet:*** https://zdjp.si

Redakcija te številke je bila zaključena 30. 06. 2026.

Sofinancirajo/*Supporto finanziario/ Financially supported by:* Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS)

Annales - Series Historia et Sociologia izhaja štirikrat letno.

Maloprodajna cena tega zvezka je 11 EUR.

Naklada/*Tiratura/Circulation:* 300 izvodov/copie/copies

Revija Annales, Series Historia et Sociologia je vključena v naslednje podatkovne baze / *La rivista Annales, Series Historia et Sociologia è inserita nei seguenti data base / Articles appearing in this journal are abstracted and indexed in*: Clarivate Analytics (USA): Arts and Humanities Citation Index (A&HCI) in/and Current Contents / Arts & Humanities; IBZ, Internationale Bibliographie der Zeitschriftenliteratur (GER); Sociological Abstracts (USA); Referativnyi Zhurnal Viniti (RUS); European Reference Index for the Humanities and Social Sciences (ERIH PLUS); Elsevier B. V.: SCOPUS (NL); Directory of Open Access Journals (DOAJ).

ANNALES · Ser. hist. sociol. · 36 · 2026 · 2

Anali za istrske in mediteranske študije - Annali di Studi istriani e mediterranei - Annals for Istrian and Mediterranean Studies

UDK 009 Volume 36, Koper 2026, issue 2 ISSN 1408-5348
e-ISSN 2591-1775

VSEBINA / *INDICE GENERALE / CONTENTS*

received: 2026-02-25 DOI 10.19233/ASHS.2026.15

TOWARDS FUTURE ARTIFICIAL INTELIGENCE AGENTS FOR IMPROVED POLITICAL DISCOURSE QUALITY WITH LARGE LANGUAGE MODELS

Tadej ŠKVORC
Institute IRRIS for Research, Development and Strategies of Society, Culture and Environment, Centur 1f, 6273 Marezige, Slovenia
e-mail: tadej.skvorc@irris.eu

Marjan HORVAT
Institute IRRIS for Research, Development and Strategies of Society, Culture and Environment, Centur 1f, 6273 Marezige, Slovenia
e-mail: marjan.horvat@irris.eu

Jure KORAŽIJA
Institute IRRIS for Research, Development and Strategies of Society, Culture and Environment, Centur 1f, 6273 Marezige, Slovenia
e-mail: jure.korazija@irris.eu

Marko ROBNIK-ŠIKONJA
University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana, Slovenia
e-mail: marko.robniksikonja@fri.uni-lj.si

ABSTRACT

Large language models have enabled large-scale analysis of many phenomena, including political discourse on social media. We analyze how finetuning models on social media posts can be used in discourse analysis. We first present a theoretical framework for analyzing political discourse and show that finetuned models are better at detecting discourse quality. We finetune models on examples that match specific discourse quality indicators and demonstrate how this process can align messages with the desired indicator.

Keywords: political discourse, AI agents, Political commemorations, Berlin Wall, Day of Europe, large language models

VERSO FUTURI AGENTI DI INTELLIGENZA ARTIFICIALE PER MIGLIORARE LA QUALITÀ DEL DISCORSO POLITICO CON I GRANDI MODELLI LINGUISTICI

SINTESI

I modelli linguistici di grandi dimensioni hanno reso possibile l'analisi su larga scala di numerosi fenomeni, tra cui il discorso politico sui social media. Analizziamo in che modo i modelli ottimizzati sulla base dei post pubblicati sui social media possano essere utilizzati nell'analisi del discorso. Presentiamo innanzitutto un quadro teorico per l'analisi del discorso politico e dimostriamo che i modelli ottimizzati sono più efficaci nel rilevare la qualità del discorso. Ottimizziamo i modelli sulla base di esempi che corrispondono a specifici indicatori di qualità del discorso e dimostriamo come questo processo possa allineare i messaggi all'indicatore desiderato.

Parole chiave: Discorso politico, agenti di intelligenza artificiale, commemorazioni politiche, Muro di Berlino, Giornata dell'Europa, modelli linguistici di grandi dimensioni

INTRODUCTION[1]

The public sphere is a key arena of political discourse in which publics and individuals contest and negotiate power, legitimacy, norms, and authority. For the public sphere to function democratically, certain criteria must be met, including inclusiveness, freedom of expression, and infrastructural conditions that enable discussion and deliberation about matters of common concern and the formation of public opinion (Habermas, 1989; 1996; Fraser, 1990). However, with the rise of social media, a growing body of research has examined how online environments reshape public discussion through platform affordances (Boyd, 2011; Gillespie, 2018), often intensifying polarization (Iyengar et al., 2019), echo chambers (Sunstein, 2018; Cinelli et al., 2021) and homophily (McPherson et al., 2001). In particular, online political discussions frequently feature high levels of incivility and hostility, which can discourage participation and undermine deliberation (Papacharissi, 2004; Coe et al., 2014). These dynamics are especially complex in debates on specific topics, such as political commemorations, as they reactivate contested collective memories and invoke embedded value systems (Meyer, 2008), turning debates about the past into heated and explicit struggles over present-day political issues (Gutman & Wüstenberg, 2023; Wüstenberg, 2017).

Assessing the democratic significance of online political discussions, including those triggered by commemorations, therefore requires attention to discourse quality, not merely its volume or sentiment. In this respect, deliberative democracy theory emphasizes interactional dimensions such as respect, reciprocity, constructiveness, and the provision of reasons (Habermas, 1996; Mansbridge et al., 2012), criteria that are more nuanced than standard computational proxies such as hate speech or toxicity. In empirical research, the Discourse Quality Index (DQI) operationalizes deliberative principles through observable coding categories and has been used to assess deliberative quality across settings (Steenbergen et al., 2003). However, applying such fine-grained frameworks at scale remains difficult, as manual annotation is costly, context-sensitive, and challenging across languages and political topics, particularly for highly contextual cases such as political commemorations.

Recent advances in large language models (LLMs) open new possibilities for bridging public sphere theory and deliberative democracy theory with scalable analysis. LLMs can support more context-aware annotation and enable novel experimental and LLM-based approaches, such as training models to detect discourse-quality indicators or to generate responses that better align with targeted qualities, potentially expanding both methodological reach and substantive insight (Ziems et al., 2024; Törnberg, 2025). This matters not only for measurement, understanding when and why discourse quality varies across political contexts, but also for the careful exploration of interventions that could support higher-quality deliberative discourse.

In recent years, methods based on natural language processing have become increasingly common for discourse analysis. Most recently, large language models dominate all kinds of text analysis. However, while LLMs have been shown to perform well in a variety of domains, their application to the analysis of political discourse has not yet been thoroughly tested, particularly for specific tasks and domains. One way to improve LLM performance on specific discourse analysis tasks and domains is through finetuning, in which general-purpose LLMs are further trained on labelled data from political discourse. However, such models are relatively rare.

A related approach to computational analysis of political discourse is the use of artificial intelligence (AI) agents. While there is no strict definition of AI agents, the term is mostly applied to entities (e.g., computer programs) that are, in some way, aware of their environment and can automatically take actions to achieve a goal. In the field of AI, such agents can be used to perform a wide range of planning and problem-solving tasks.

When dealing with text, the term AI agent commonly refers to an LLM that can generate text autonomously, without explicit human input. Commonly, such agents are trained for a specific task (e.g., mimicking human conversations, making decisions about complex systems, or answering specific questions). Due to recent advances in the field of natural language processing (NLP) (i.e., the advances in LLMs and generative AI), these types of AI agents have become increasingly ubiquitous, and the term "AI agent" now mostly refers to this type of AI agent.

1 The authors acknowledge funding from the European Union's Horizon Europe programme under grant agreement No. 101094752, *Social Media for Democracy – Understanding the Causal Mechanisms of Digital Citizenship* (SoMe4Dem). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or Horizon Europe. Neither the European Union nor the granting authority can be held responsible for them. This work was also supported by the European Union's Horizon Europe programme under grant agreement No. 101186647, *Centre of Excellence in Artificial Intelligence for Digital Humanities* (AI4DH). The article is also the result of research supported by the Slovenian Research and Innovation Agency (ARIS) within the research programme P6-0411, *Language Resources and Technologies for the Slovenian Language*; P6-0435, *Practices of Conflict Resolution Between Customary and Statutory Law in the Area of Today's Slovenia and Its Neighbouring Lands*; and research project GC-0002, *Large Language Models for Digital Humanities*.

In this work, we analyze the political discourse related to the commemorations of the Day of Europe, the fall of the Berlin Wall, and the Italian National Memorial Day of the Exiles and Foibe (*Giorno del ricordo*). Methodologically, we examine two analytical approaches.

- First, we examine whether finetuned LLMs can analyze political discourse using various indicators of discourse quality based on the discourse quality index (DQI). We manually annotate a small number of texts using eight indicators of discourse quality and use these texts to finetune general LLMs. We show that this improves the models' performance at detecting discourse quality indicators compared to larger, non-finetuned models.
- Second, we examine whether finetuning LLM models on a small amount of specific types of political discourse (e.g., antagonistic, respectful, or uncivil texts) can be used to align LLM-generated text with specific types of political discourse. We finetune the models on datasets related to the mentioned commemorations and evaluate the outputs using the classification models trained in the first part of our work. We then compare the outputs produced by our models with several huge non-finetuned models, including LLMs commonly used by the public (e.g., ChatGPT4).

We present an approach towards solving both tasks using language technologies. Specifically, we make the following scientific contributions:

1. We show that both the above tasks can be successfully tackled using finetuned LLMs, which could lead to future development of larger agentic AI systems.
2. We show that this approach can be successful even with relatively small LLMs (e.g., the openly available Gemma 3 family of models (Kamath et al., 2025) with 4b and 12b parameters that can be run on personal computers, as opposed to commercial models that require large datacenters).
3. We show this can be done by training the models either on a small number of manually-labelled examples, or on automatically-labelled data that avoids the need for time-consuming manual annotation.
4. We present a comprehensive evaluation on two case studies of social media discourse related to political commemorations (Day of Europe/the fall of the Berlin Wall and the Italian National Memorial Day of the Exiles and Foibe (*Giorno del ricordo*))

In Section "Related Work", we present past works related to our approach from a theoretical and practical standpoint. We follow with the description of our methodology in Section "Methodology", followed by Section "Results", where we present the evaluation of our approach. Section "Conclusion and Further Work" contains the conclusion and ideas for further work.

RELATED WORK

In this section, we first provide an overview of technological approaches to discourse analysis. We follow with a brief overview of works related to AI agents and present how AI agents can be used to analyze political discourse, specifically that related to the domain of commemorations in Eastern and Central Europe. We present specific methodology that can be used to train AI agents and describe how such agents can be used to generate political discourse that is either antagonistic, agonistic, or deliberative.

Technological approaches to political discourse analysis

Technological approaches to political discourse analysis have developed along several methodological waves, shaped by both platform data availability and advances in computational methods. Early work in computational political communication relied on dictionary-based content analysis and supervised models to quantify topics, frames, sentiment, and ideological positioning in large collections of political texts (Laver et al., 2003; Grimmer & Stewart, 2013). As social media became central to political communication, computational social science expanded the methodological toolkit beyond text classification by combining linguistic signals with network analysis and diffusion modelling to study polarization, selective exposure, mobilization, and information disorder at scale (Barberá, 2015; Conover et al., 2011; Vosoughi et al., 2018). In parallel, a substantial literature developed methods to detect and quantify problematic or strategic forms of online discourse, including incivility, hate speech, toxicity, and coordinated manipulation (Coe et al., 2014; Ferrara et al., 2016; Fortuna & Nunes, 2018; Tucker et al., 2018).

Speech act analysis is also common across discourse analysis on social media. Such works try to detect different kinds of speech acts (i.e., the functions and acts performed by a given text), defined by theoretic taxonomies (Searle, 1969). Multiple authors present datasets and models aimed at detecting and classifying speech acts in social media posts, including on X (Saha et al., 2019; Zhang et al., 2011) and Facebook (Ilyas & Khushi, 2012).

While these methods are powerful, much applied research still operationalizes discourse through relatively coarse proxies such as sentiment, offensiveness, or misinformation. Such proxies do not fully capture deliberative qualities such as reciprocity, respect,

constructiveness, and justification that are central to discourse quality (Habermas, 1996; Mansbridge et al., 2012). Recent progress in large language models offers new opportunities to narrow this gap through more context-sensitive modelling, including few-shot and finetuned approaches that reduce reliance on large manually labelled datasets and facilitate multilingual adaptation (Devlin et al., 2018; Brown et al., 2020; Törnberg, 2025). Beyond measurement, LLMs also enable generative approaches that can be evaluated for their capacity to shift discourse toward specific qualities, raising novel methodological possibilities for facilitating democratic communication and moderation.

AI agents and large-language models

The broader history of AI agents predates the text-based agents that are increasingly popular today. Early AI agents were simple rule-based approaches that gathered observations (precepts) from the environment and performed predefined actions for each precept following (i.e., following the condition-action rule). Russel and Norvig (2003) label this kind of AI agents as simple reflex agents and define several other categories of complexity (e.g., goal-based, modelbased, and utility agents). The most notable of those categories is learning agents, which begin from an unoptimized initial state and then gradually learn from the environment to improve their performance.

As with other agents, early text-based AI agents relied on simple logic-based methods. Commonly, this included simple NLP approaches such as pattern matching and rule-based systems. Simple machine-learning and reinforcement-learning approaches were also common, allowing agents to learn and improve from human-generated text. While these simple agents could replicate human language to a certain extent, they lacked the complexity to perform more complex tasks. A prominent early example includes ELIZA (Weizenbaum, 1976), a program designed to simulate human conversation using a variety of hard-coded rules and responses. While the program was incapable of understanding conversation or learning from human responses, it was still capable of emulating human conversation to a limited extent.

Later, methods based on machine learning and neural networks became increasingly common. Advances in NLP methods allowed such systems to become more complex and more capable of replicating human language. Text embedding methods such as Word2Vec (Mikolov et al., 2013) and BERT (Devlin et al., 2018) allowed for more powerful text representations that could take into account contextual information of words.

In recent years, text-based AI agents commonly make use of deep learning and LLMs. Most often, such systems use an LLM as a foundation and expand it with tools for reasoning and planning. This can include complex chain-of-thought reasoning for decision making (Wei et al., 2022), external memory modules (Sumers et al., 2024) and API calls that allow AI agents to use external tools and databases (Schick et al., 2023). Established methodologies for training text-based agents start by using a general-purpose pre-trained LLM as a foundational model and then perform additional finetuning based on the desired task (Parthasarathy et al., 2024). Several LLMs can be used for this task, including OpenAI's GPT models (Achiam et al., 2023), DeepSeek (Liu et al., 2024), Llama (Touvron et al., 2023), or Gemma (Kamath et al., 2025).

AI agents for political discourse

While AI agents have advanced significantly in recent years, current state-of-the-art systems are predominantly commercial and focus on specific, marketable tasks. (Casper et al., 2025) present an overview of 67 state-of-the-art AI agents and show that over 70% focus on software engineering (i.e., assisting with coding and programming) and computer use (i.e., open-ended interaction with computer programs), while others are either general-purpose reasoning engines (e.g., ChatGPT) or focus on specific areas such as robotics or assisting with scientific research (e.g., generating a list of related work). Consequently, AI agents focused on social sciences and humanities are comparatively underexplored, and there are few examples that are specifically designed for the analysis of political discourse. When AI agents are used for political discourse, it is often in negative or morally dubious contexts. For example, AI agents are often used to conduct astroturfing campaigns or to spread disinformation and fake news surrounding various political topics (Marcellino et al., 2023). While this shows that AI agents can engage in political discourse, there is a clear lack of positive examples of AI agents in this context.

Additionally, human participants often hold negative views of AI-enabled deliberation, even when AI agents are used with good intentions. (Jungherr & Rauchfleisch, 2025) show that people view AI moderation, AI summarization, AI opinion aggregation, and other AI deliberation tools less favorably than when such tasks are performed by humans.

Positive examples of AI agents assisting in deliberation are rare, and using agentic approaches to improve political deliberation is currently underexplored, despite their significant potential. Some existing examples include the work of (Tessler et al., 2024),

who propose an agentic system called the Habermas Machine that can mediate political discourse across opposing positions. The machine produces automatically generated group statements that are more commonly endorsed by the entire group than those composed by humans. This shows that AI agents can be helpful in mediating political discourse.

Another area that is currently under-explored is using AI agents as a tool for designing policy recommendations related to different political topics. For example, it is possible to design AI agents that mimic various forms of political discourse (i.e., antagonism, agonism and deliberation) surrounding specific political commemorations from different political positions (i.e., right, left, and center-wing). Such agents could be used to provide a quick overview of political positions on specific events or to synthesize group statements (as with the Habermas Machine (Tessler et al., 2024)), which could be helpful when designing policy recommendations.

While our work is still preliminary and does not present a direct practical use-case, it could lead to development of AI agents with practical applications. For example, by generating respectful, constructive replies that would guide uncivil discourse towards a more desirable outcome or represent excluded perspectives (e.g., posts backed by experience or reason) in discussions of specific commemorations (Fournier-Tombs, 2024.). With additional finetuning, similar models could also be used to perform automated discussion mediation (e.g., by generating constructive messages that lead participants towards a shared positive consensus (Tessler et al., 2024)). Other possibilities include analytical agents for monitoring discourse quality, moderator agents combined with human oversight, and agents that would suggest rewrites of offensive posts with more deliberative, civil alternatives. The end goal would be a development of carefully-monitored, human guided tools that could help improve the quality of political and public discourse.

However, each practical application comes with ethical considerations. For example, such LLMs could also be used to generate uncivil, hateful responses, exclude certain perspectives, or generate misleading messages.

METHODOLOGY

In this section, we present the multiple methodological contributions of our work. We begin by presenting the theoretical framework we use to evaluate political discourse in social media posts. The analysis is based on 8 indicators of discourse quality presented in Section "Discourse quality indicators". We used the discourse quality indicators to construct multiple datasets for finetuning our models, as described in Section "Dataset". These datasets were used in the analysis of political discourse in two ways: i) to automatically detect which indicators are present in a given social media posts, and ii) to finetune LLMs on specific types of messages so that they are better aligned with those types of messages (e.g., taking an existing AI agent or LLM that is prone to producing uncivil messages and aligning it towards civil messages). We present a detailed description of the proposed AI agents, as well as their training, in Section "Large language models for analysis of political discourse".

Discourse quality indicators

To analyse commemorative political discourse on X, we build on operational groundwork developed in our prior work on discourse quality in discussions of political commemorations. Our operationalisation follows deliberative-democracy framework associated with the Discourse Quality Index (Steenbergen et al., 2003; Bächtiger et al., 2009), and is aligned with recent computational approaches to measuring deliberative quality in large-scale online data (Beauchamp, 2020; Fournier-Tombs & MacKenzie, 2021). To capture deliberative-quality signals in X posts, we also draw on the Corpus for the Linguistic Analysis of Political Talk ONline, which provides a framework for features such as constructiveness, justification, relevance, reciprocity, empathy, and incivility and was already used to train end evaluate several machine learning models (Jaidka, 2022). We choose this set of indicators to specifically build upon recent work involving computational approaches and machine learning.

We complement this with established approaches to identifying conflict in online political communication (Canute et al., 2023). We tailor these concepts to the specificities of commemorative discourse surrounding Europe Day, the fall of the Berlin Wall, and *Giorno del ricordo*. This tailoring was carried out through an iterative operational process in which the initial coding categories derived from Jaidka (2022) and Canute et al. (2023) were refined into increasingly explicit indicator definitions, decision rules and examples. We experimented with different few-shot examples, prompt structures and models to ensure the the automatic annotations matched human-annotated samples. The final prompt reported in Appendix 1 (Škvorc et al., 2026) was selected after pilot annotations because it produced the most consistent agreement with the manually applied coding scheme. Exploratory intermediate prompt variants were not retained.

Methodologically, this paper extends the above line of work by formalising the indicator set as a multi-label target space for LLM-based modelling, enabling scalable detection and controlled generation

of discourse qualities. First, we finetune open LLMs to detect discourse-quality indicators at scale in commemorative debates. Second, we finetune generative models on indicator-specific subsets to test whether outputs can be systematically shifted toward targeted discourse qualities, probing the potential of agent-based interventions in online deliberation (Fournier-Tombs, 2024).

We use the following indicators, following the work presented by Jaidka et al. (2022) with minor modifications:

- Conflict: does the post express disagreement, critique, blame, or opposition toward any target (actors, policies, ideas, events, institutions, groups), even if it is phrased politely/respectfully or expressed implicitly (sarcasm, ironic praise, comparative shaming)?
- Positive/Respectful: Does the post show respect or empathy?
- Uncivil: Does the post include abuse, slurs, negative stereotypes, threats, or exaggeration/hyperbole used as argument?
- Reciprocity: Does the post ask genuine information-seeking questions intended to elicit the other side's views or facts?
- Constructiveness: Does the post contain fact-checking, seek common-ground, or propose solutions/next steps?
- Justification with reason: Does the post give a logical reason or explanation for its view? Specifically, does it provide a because/therefore/so rationale, causal claim, or argument?
- Justification with experience: Does the post explicitly use a first-hand or observed experience as justification for a claim in the post?
- Justification with a link: Does the post use a URL/link as evidence/source for a claim.

The original set of indicators presented by Jaidka et al. (2022) also includes a criteria called "Relevance", which checks whether a given post is even relevant to their desired use-case (i.e., analyzing political discourse). In our case, the data process (i.e., filtering by keywords) ensures the selected posts are relevant to political discourse surrounding the selected commemorations. We therefore replace this with a more specific indicator termed "Conflict", which doesn't just check for the presence of political discourse, but also checks that the discourse expresses some form of disagreement, critique, blame, or opposition.

We omit the "Justification with a link" indicator from our analysis. The indicator was rarely present in our datasets and is not interesting from the standpoint of computational analysis. The presence of links can be detected without the use of LLMs (e.g., using pattern matching tools).

Dataset

In order to finetune our models, we use multiple datasets related to the political discussion of specific commemorations on the social network X. Specifically, we focus our attention on the Day of Europe, the fall of the Berlin Wall, and the Italian *Giorno del ricordo*. We select these commemorations because they differ in how conflictual and polarising the activated discourse tends to be. Europe Day discourse is typically more ceremonial and oriented toward European integration and institutional legitimacy, whereas the Berlin Wall anniversary is a historically charged rupture frame that is readily mobilised in present-day disputes (e.g., democracy, sovereignty, geopolitics). *Giorno del ricordo* represents a particularly polarised and contested memory domain in Italy and the Italo-Slovenian borderland, structured through competing narratives of victimhood, responsibility, and national identity. These commemorations provide politically charged discourse centred on specific topics. The work builds upon the results presented in (Lampe et al., 2026) and (Horvat & Koražija, 2026), which show these commemorations are interesting from the perspective of discourse analysis.

These makes the commemorations useful for LLM-based analysis, as they should contain a wide range of discourse quality indicators. Posts related to Europe Day are more likely to contain discourse that is respectful and non-conflictful, while the opposite is true for posts related to *Giorno del ricordo*. Additionally, these commemorations allow us to examine how the detected indicators translate to political discourse that is relevant today. The selected posts contain discussions focused around migration, xenophobia, the European Union, Fascism and Democracy and other relevant topics.

We collect posts by filtering X data from April 2023 to June 2025 using event-specific keyword queries related to each commemoration. Using these commemorative contexts, we construct four datasets: two manually labelled datasets used as high-quality supervision and evaluation data, and two larger automatically labelled datasets used to scale finetuning:

1. Manually-labelled Day of Europe and Berlin Wall dataset of 193 posts on the social network X. We combine Europe Day and Berlin Wall posts into a single manual dataset because the label schema is identical, the manual sample is small, and because discussion related to both commemorations centers around similar contemporary topics (e.g., migration and politics). We manually label each post using eight criteria described in Section "Discourse quality indicators" (conflict, respectfulness,

incivility, reciprocity, constructiveness, justification reason, justification experience, and justification link). To avoid ambiguity, each example is annotated by multiple annotators. The datasets contains posts mainly in German, with a small number of English, French, Italian, and Slovene posts.

2. A manually-labelled *Giorno del ricordo* dataset. We repeat the process from the previous dataset on posts containing keywords related to the Italian *Giorno del ricordo*. This allows us to perform our analysis across multiple domains and multiple languages. The *Giorno del ricordo* dataset contains posts in Italian and Slovene.
3. Two automatically labelled datasets. Because manual annotation of political discourse on social media is time-consuming, the two datasets above contain a small number of examples. While few-shot finetuning can be used to successfully finetune LLMs, we include a larger number of examples that were automatically labelled with ChatGPT-5.2 (OpenAI, 2026). The full prompt used to perform this labelling is shown in Appendix 1 (Škvorc et al., 2026). We used an English prompt for all data available in our dataset and relied on the LLMs' native cross-lingual capabilities to process non-English text. While these annotations are not 100% reliable, our evaluation in Section "Discourse quality analysis" shows they perform reasonably well and can be used to successfully finetune LLMs.

We construct two automatically labelled datasets, each corresponding to one of the manually labelled datasets. The Day of Europe/Berlin Wall dataset contains 47.166 examples in German, Slovene, French, and Italian. The *Giorno del ricordo* dataset contains 23.389 examples in Slovene and Italian.

Human-annotated ground-truth samples and adjudication

To provide a reliable reference point for evaluation and prompt development, we created manually annotated ground-truth samples for both case studies. These samples were used as expert-coded reference data. The manually adjudicated samples therefore serve as the basis for evaluating model performance and for checking whether automatically generated annotations are meaningful.

Berlin Wall/Europe Day ground-truth sample. For the Berlin Wall/Europe Day case study, we manually annotated a sample of Slovene X posts. We drew 200 random posts from the larger corpus, 100 from the Berlin Wall subset and 100 from the Europe Day subset. The individual post was the unit of annotation. Each post was coded according to the discourse quality indicators described in Section "Discourse quality indicators", with every indicator treated as a separate binary label.

The two expert annotators coded the sampled posts independently. Complete paired annotations were available for 192 posts. For the remaining eight posts, at least one annotator did not provide a full label set, because the post could not be coded with sufficient confidence under the annotation scheme. These cases were therefore excluded from the reliability calculation and from the adjudicated reference set. The Berlin Wall/Europe Day ground-truth sample used in the analysis consequently contains 192 posts.

Inter-annotator agreement was calculated on the 192 posts for which both annotators had provided complete independent annotations. These statistics were computed before adjudication and are reported in Table 1. The annotators then reviewed and resolved disagreements, producing a final consensus label for each post and each indicator. These adjudicated labels were used as the ground-truth reference for prompt development and model evaluation.

Table 1: Expert–expert agreement before adjudication for the Berlin Wall/Europe Day ground-truth sample.

Indicator	N	Agreement	Cohen's κ	Gwet's AC1
Conflict	192	81.58%	0.6136	0.6495
Positive/Respectful	192	90.53%	0.6231	0.8735
Uncivil	192	84.21%	0.5698	0.7527
Reciprocity	192	80.32%	0.2948	0.7270
Constructiveness	192	87.30%	0.4973	0.8301
Justification–Reason	192	71.51%	0.4419	0.4360
Justification–Experience	192	96.28%	0.2148	0.9609

Giorno del ricordo ground-truth sample. For the *Giorno del ricordo* case study, we used the same expert-coding procedure. The manual reference sample consisted of 200 X posts from the Slovene and Italian corpus. As above, the individual post was the unit of annotation, and each discourse quality indicator was coded as a separate binary label.

Two expert annotators coded all posts independently using the multi-label scheme described in Section "Discourse quality indicators". Inter-annotator agreement was calculated before adjudication and is reported in Table 2. The annotators then reviewed the disagreements together with a third expert judge. This adjudication step produced a final consensus label for each post and each indicator. These adjudicated labels were used as the ground-truth reference for the *Giorno del ricordo* evaluation.

Dataset imbalance

A key issue in developing manually curated datasets related to political discourse on social media is the imbalance in the distribution of certain labels. For example, discourse related to charged political events is less likely to be civil. Table 3 shows the distribution of labels in our datasets, showing present imbalances. Indicators related to civil speech ("respectful", "reciprocity", "constructiveness") appear less often than indicators such as "conflict" and "uncivil", regardless of whether they were labelled manually or using GPT. Additionally, "justification with experience" appears infrequently across all datasets.

In order to address label imbalance, we attempt to balance the manually-labelled datasets by adding a number of the most suitable examples from the automatically-labelled datasets. We first use the sentence-transformer embedding gemma-300m model to calculate a vector representation of every example text in both the manually-labelled and the automatically-labelled datasets. We then balance the manually-labelled data by finding an automatically-labelled example with the same indicator label and the highest similarity to its text. We pick texts based on similarity to increase the likelihood of picking correctly-labelled examples. This reduces the variance of the balanced dataset and could lead to over-fitting. However, as we have not performed a thorough manual evaluation of

Table 2: Expert–expert agreement before adjudication for the Giorno del ricordo ground-truth sample.

Indicator	N	Agreement	Cohen's κ	Gwet's AC1
Conflict	200	88.0%	0.710	0.790
Positive/Respectful	200	75.5%	0.450	0.560
Uncivil	200	87.5%	0.660	0.800
Reciprocity	200	95.5%	0.390	0.950
Constructiveness	200	68.0%	0.270	0.450
Justification–Reason	200	77.0%	0.460	0.600
Justification–Experience	200	97.0%	0.000	0.970

Table 3: The characteristics of the datasets used in our analysis.

Dataset information					Annotations				
Dataset	Examples	Language	Conflict	Respectful	Uncivil	Reciprocity	Reason	Experience	Constructiveness
Berlin Wall/Day of Europe (Manual)	192	SI	125	34	40	8	91	4	34
Berlin Wall/Day of Europe (GPT)	47.166	SI, GER, FR, IT	26.760	4.441	15.243	2.645	15.398	1.606	15.456
Giorno del ricordo (Manual)	200	SI, IT	133	52	50	5	62	2	72
Giorno del ricordo (GPT)	23.389	SI, IT	16.290	7.931	7.005	1.087	10.904	637	6.562

Table 4: The characteristics of the balanced datasets used in our analysis.

Dataset	Examples	Conflict	Respectful	Uncivil	Reciprocity	Reason	Experience	Constructiveness
Berlin Wall/Day of Europe (Balanced)	2663	609	261	356	224	552	535	126
Giorno del ricordo (Balanced)	2816	640	279	384	218	614	554	127

the automatically-labelled dataset, we prioritize picking examples with the correct labels.

We repeat this process for each indicator until the indicator is balanced. We then remove duplicate posts and check that none of the added posts had previously appeared in the unbalanced version of the dataset to prevent data leakage between training and testing sets (i.e., to ensure no post in the test set is also present in the training set). Because each example can be labelled with multiple indicators (e.g., a text can be both uncivil and provide a justification with reason) this does not produce a perfectly labelled dataset but still increases the amount of messages with under-represented indicators.

We present the number of examples for each indicator present in the balanced dataset in Table 4. Conflict remains the most frequent indicator, but the indicators are now more balanced. In Section "Label co-occurences", we show that several of the indicators are highly correlated with conflict, making it difficult to find posts that contain them without also containing conflict. However, the balancing still increases the number of less frequent indicators that almost never appear in the original, unbalanced datasets (i.e. experience and reciprocity).

Label co-occurrences

Because we are dealing with multi-label classification (i.e., a post can be labelled with multiple indicators at the same time), we perform an analysis of label co-occurrences present in our datasets. We present the co-occurrences in Figure 1, which shows a significant presence between multiple indicators. All datasets show a the highest correlation between the indicators "uncivil" and "conflict". With the exception of the balanced Berlin Wall dataset, every post marked as uncivil was also marked as containing conflict (although the reverse is not true). This can be explained by the nature of our data (contentious political commemorations) and the high amount of posts containing conflict.

Correlations also exist to a smaller extent between several other pairs (e.g., conflict-constructive, constructive-reason, conflict-reason and positive-constructive pairs.) We see similar co-occurrences across all four datasets. Balancing the dataset with automatically-labelled data maintains similar relative co-occurrences while increasing the number of samples.

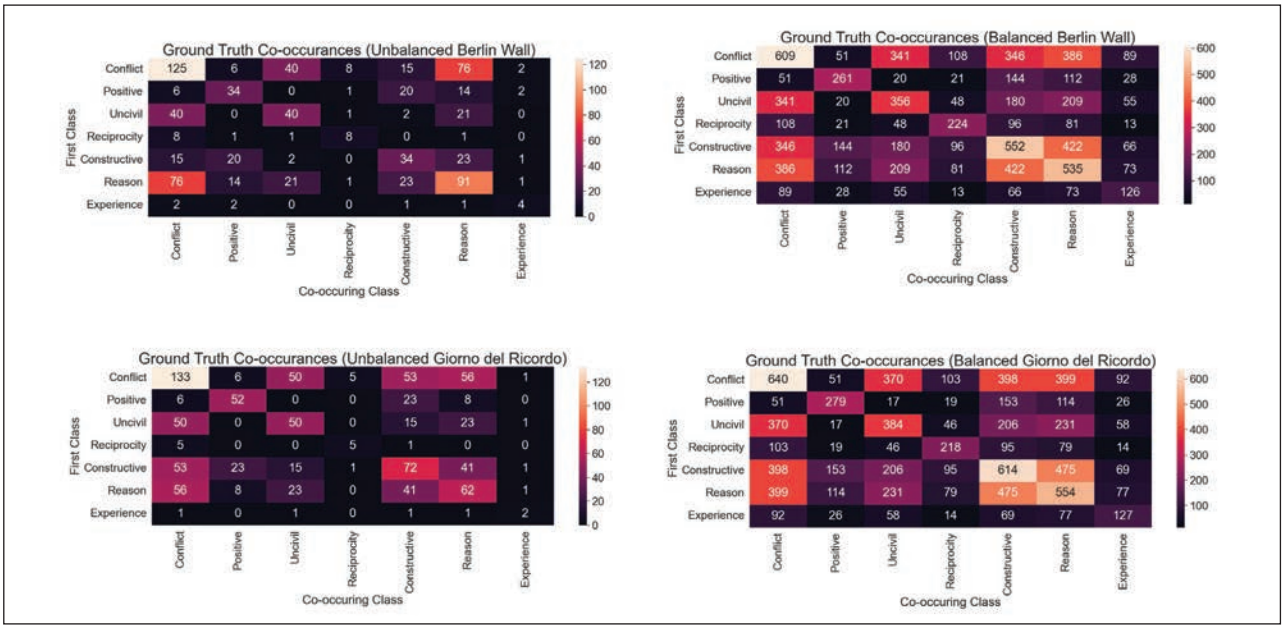

Figure 1: Label co-occurrences present in each of the four datasets.

Large language models for analysis of political discourse

After constructing the required datasets, we used two LLM approaches for the analysis of political discourse. In Section "Discourse quality analysis models", we first create finetuned models designed to analyze posts according to the discourse indicators presented in Section "Discourse quality indicators". We then use these models to evaluate whether finetuning existing LLMs on posts tied to specific discourse indicators (e.g., civility or constructiveness) can change the way these models talk about specific topics related to the chosen commemorations (e.g., producing messages that are more civil and constructive). We present this approach in Section "Aligning large language models for political discourse". To ensure reproducibility and allow for open model access, we focus our analysis on the openly available Gemma 3 models. We also attempted to finetune smaller, non-generative multilingual classification models models (e.g., bert-base-multilingual (Devlin et al., 2018) and mmBert (Marone et al., 2025)). These models underperformed even compared to the non-finetuned baselines. Due to their smaller size and smaller amounts of pre-training data (e.g., 12 trillion tokens for Gemma3-12b vs 3 trillion for mmBert), we estimate they would require a larger finetuning corpus to achieve comparable performance.

Discourse quality analysis models

To obtain LLMs capable of analyzing the quality of political discourse, we finetune several Gemma 3 models on the manually labelled Day of Europe/Berlin Wall and *Giorno del ricordo* datasets. We employ finetuning in order to:

1. Obtain models that are smaller, faster to run, and more energy efficient than the full ChatGPT-5.2 model while still achieving good performance.
2. Obtain open-source models that can be run locally, rather than being tied to corporate models running on remote data centers.
3. Imbue the models with information specific to the chosen commemorations, which may not be included in the training sets of non-finetuned LLMs.

We split each dataset into train, test, and validation sets at a 70/20/10 ratio, with the validation set used for hyper-parameter optimization. Based on this, we use a max sequence length of 2048, a learning rate of $5e^{-5}$, and perform the finetuning using low-rank adapters with $r = 8$, $\alpha = 8$, and dropout set to 0. During prediction, we set the temperature to 0 in order to ensure the model always returns the most likely output. We train the model using specially-designed prompts that include thorough descriptions of each indicator. The full prompt is available in Appendix 1 (Škvorc et al., 2026). We perform this experiment using two sets of training data:

1. The manually-labelled dataset described in Section "Dataset". This dataset contains accurate labels for each discourse indicator but only contains a small number of tweets. Additionally, such a methodology is more difficult to replicate in future research due to the time-consuming need for manual annotation and a lack of existing manually-labelled datasets.
2. The balanced version of the manually-labelled dataset described in Section "Dataset. This augments the manually-labelled dataset with examples from the automatically-labelled dataset to reduce the imbalances present in multiple indicators. We use this dataset in order to better evaluate the indicators that are rarely present in the manually-labelled dataset.

We compare our results to a non-finetuned GPT-5 mini model in a prompting mode as a baseline. A preliminary analysis showed that smaller models (e.g., the non-finetuned Gemma3-4b and Gemma3-12b model) were not powerful enough to suitably perform this task. They were unable to follow the instructions in the prompt and generated text that did not contain proper labels in a large amount of cases and were therefore omitted as a baseline for comparison.

We then perform an additional experiment using the two automatically labelled datasets. The results of this evaluation are presented in Section "Discourse quality analysis".

Aligning large language models for political discourse

After obtaining finetuned models capable of analyzing political discourse, we examine whether existing LLMs can be finetuned to produce messages more closely aligned with specific types of discourse (e.g., respectful, reciprocal, or justified). To do this, we first finetune existing Gemma 3 models on posts from different discourse types. For each post, we construct the following prompt: "Generate a message related to the following political commemoration: <commemoration> on the following topic <topic>", where <topic> is a specific topic related to the commemoration that was assigned to each message during dataset construction. The full list of topics is presented in Appendix 2 (Škvorc et al., 2026). In order to train the model, we use existing posts from our dataset as ground-truth answers to this prompt. We train one model for

each type of discourse label present in our dataset (conflict, respectful, civil, reciprocal, constrictive, justification with reason/experience) for a total of 8 different models. We then evaluate the messages produced by the aligned models using the discourse analysis models described in Section "Discourse quality analysis models". This allows us to see how the messages produced by each model shift when compared to the base model (in terms of discourse quality indicators). We also compare the aligned models with a strong, non-finetuned LLMs of a similar complexity as a baseline (i.e., GPT-5 mini).

For this experiment, we finetune all models using the automatically-labelled datasets and evaluate them using a classifier trained on the manually-labelled dataset. Due to the small number of examples, finetuning on the manually-labelled dataset produces models that do not generate a diverse set of messages and therefore have limited practical use.

We use the finetuned LLMs developed in Section "Discourse quality analysis models" to perform automatic analysis of the aligned models. While these models are not 100% accurate, we show that they perform well enough to provide useful results. The results of this analysis are presented in Section "Aligned language models".

RESULTS

We split the results section into two parts. First, we present the evaluation of models designed to detect discourse quality indicators (Section "Discourse quality analysis"). We show that these models achieve better results when compared to larger, non finetuned models. We then use these verified detection models to evaluate the LLMs oriented towards different discourse indicators. We present these results in "Aligned language models".

Discourse quality analysis

To evaluate finetuned LLMs, we perform multiple evaluations. First, we evaluate finetuned models on the manually labelled Berlin Wall/Day of Europe and Giorno del Ricordo datasets to obtain an objective measure of their classification performance. The results of this evaluation for non-balanced data are presented in Table 5 (for the Berlin Wall/Day of Europe dataset) and Table 7 (for the *Giorno del ricordo* dataset).

All results show an overall dominance of finetuned models, in particular Gemma3-12b variant. The results also show several issues with discourse quality analysis on X and dataset construction. For several indicators,

Table 5: The classification accuracy of finetuned models on the discourse quality indicators measured on the nonbalanced Berlin Wall/Day of Europe dataset. The best scores are typeset in bold.

Model	Conflict	Respectful	Uncivil	Reciprocity	Reason	Experience	Link	Constructiveness
Majority-class baseline	29/39 (0.74)	36/39 (0.92)	29/39 (0.74)	37/39 (0.95)	34/39 (0.87)	39/39 (1.00)	33/39 (0.85)	34/39 (0.87)
GPT-5 mini	33/39 (0.85)	30/39 (0.77)	31/39 (0.80)	35/39 (0.90)	28/39 (0.74)	37/39 (0.95)	34/39 (0.87)	33/39 (0.85)
Gemma3-4b	**35/39 (0.90)**	**36/39 (0.92)**	30/39 (0.77)	36/39 (0.92)	31/39 (0.80)	38/39 (0.97)	37/39 (0.95)	34/39 (0.87)
Gemma3-12b	**35/39 (0.90)**	**36/39 (0.92)**	**34/39 (0.87)**	**37/39 (0.95)**	**34/39 (0.87)**	**39/39 (1.00)**	**38/39 (0.97)**	**35/39 (0.90)**

Table 6: The classification accuracy of finetuned models on the discourse quality indicators measured on the balanced Berlin Wall/Day of Europe dataset. The best scores are typeset in bold.

Model	Conflict	Respectful	Uncivil	Reciprocity	Reason	Experience	Constructiveness
Majority-class baseline	0.590	0.787	0.681	0.766	0.527	0.872	0.543
GPT-5 mini	0.793	0.718	0.782	0.856	0.702	0.888	0.532
Gemma3-4b	0.835	0.856	0.755	0.872	0.707	0.910	0.675
Gemma3-12b	**0.936**	**0.926**	**0.867**	**0.926**	**0.814**	**0.989**	**0.771**

the majority-class baseline is close to one. Out of all annotated posts on the Berlin Wall/Day of Europe dataset, 92.3% were labelled as disrespectful and 94.9% were labelled as non-reciprocal. No post in our test set contained a justification with a personal experience. Other indicators (conflict, uncivil, justification/reason, justification/link, and justification/constructiveness) were more balanced. Similar imbalances are present in the *Giorno del ricordo* dataset (100% for reciprocity and justification/experience). Another limitation of this dataset is the small size of the test set, which contains around 40 examples (39 on the Berlin Wall/Day of Europe and 40 on the *Giorno del ricordo* dataset). This makes the evaluation difficult because a single misclassified tweet shifts the accuracy by roughly 2.5%. For clarity, we also report the absolute number of correct predictions on non-balanced datasets.

To address this, we repeat the analysis on the balanced versions of our datasets, as described in Section "Dataset". This allows us to better evaluate the under-represented indicators but means that the dataset is no longer fully manually annotated. The results for balanced data are presented in Table 6 (for the Berlin Wall/Day of Europe dataset) and Table 8 (for the *Giorno del ricordo* dataset).

As before, the finetuned models outperform the non-finetuned baseline, including on indicators that were previously imbalanced (e.g. recipricoty and experience). Due to a larger number of examples on the test set (188 in each case), we omit the absolute numbers of correct predictions.

Due to the label imbalances present in our data, particularly for non-balanced datasets, classification accuracy is not the best performance measure, as even simply predicting the majority class would lead a high classification accuracy. To address this, we also calculate F1-averaged precision, recall, and F1 score for each dataset and each indicator. We present the results in Table 9 (unbalanced Berlin Wall/Day of Europe), Table 10 (balanced Berlin Wall/Day of Europe), Table 11 (unbalanced *Giorno del ricordo*) and Table 12 (balanced *Giorno del ricordo*). This evaluation shows that the non-finetuned GPT-5 mini struggles to outperform the majority-class baseline, while the finetuned models achieve better results. As before, the Gemma3-12b model outperforms the Gemma3-4b and Gemma3-27b models in the majority of scenarios. Even on unbalanced datasets, the models are not simply predicting the majority class (except with reciprocity and experience in the non-balanced *Giorno del ricordo* dataset, where only one class appears in the test set).

Table 7: The classification accuracy of finetuned models on the discourse quality indicators measured on the nonbalanced Giorno del ricordo dataset. The best scores are typeset in bold.

Model	Conflict	Respectful	Uncivil	Reciprocity	Reason	Experience	Constructiveness
Majority-class baseline	29/40 (0.73)	32/40 (0.80)	33/40 (0.83)	**40/40 (1.00)**	35/40 (0.78)	1.00 (40/40)	25/40 (0.63)
GPT-5 mini	37/40 (0.93)	32/40 (0.80)	36/40 (0.90)	**40/40 (1.00)**	23/40 (0.58)	15/40 (0.38)	15/40 (0.38)
Gemma3-4b	37/40 (0.93)	35/40 (0.88)	**38/40 (0.95)**	**40/40 (1.00)**	30/40 (0.75)	39/40 (0.98)	**33/40 (0.83)**
Gemma3-12b	**37/40 (0.93)**	**37/40 (0.93)**	35/40 (0.88)	37/40 (0.93)	**32/40 (0.80)**	**40/40 (1.00)**	35/40 (0.78)

Table 8: The classification accuracy of finetuned models on the discourse quality indicators measured on the balanced Giorno del ricordo dataset. The best scores are typeset in bold.

Model	Conflict	Respectful	Uncivil	Reciprocity	Reason	Experience	Constructiveness
Majority-class baseline	0.705	0.664	0.589	0.774	0.541	0.938	0.561
GPT-5 mini	0.938	0.890	0.904	0.932	0.774	0.274	0.657
Gemma3-4b	0.938	0.911	0.863	0.945	0.801	0.900	0.739
Gemma3-12b	**0.945**	**0.918**	**0.911**	**0.938**	**0.863**	**0.966**	0.842
Gemma3-27b	0.932	0.904	0.884	0.925	0.781	0.815	**0.959**

Table 9: Macro averaged precision, recall, and F1 scores of finetuned models on the discourse quality indicators measured on the non-balanced Berlin Wall/Day of Europe. The best F1 scores are typeset in bold.

Model	Conflict P | R | F1	Respectful P | R | F1	Uncivil P | R | F1	Reciprocity P | R | F1	Reason P | R | F1	Experience P | R | F1	Constructiveness P | R | F1
Majority-class baseline	0.37 | 0.5 | 0.43	0.46 | 0.5 | 0.48	0.37 | 0.5 | 0.43	0.47 | 0.5 | 0.49	0.43 | 0.5 | 0.47	0.42 | 0.5 | 0.46	0.43 | 0.5 | 0.47
GPT-5 mini	0.37 | 0.5 | 0.43	0.49 | 0.48 | 0.33	0.33 | 0.35 | 0.34	0.57 | 0.65 | 0.58	0.76 | 0.65 | 0.70	0.5 | 0.15 | 0.23	0.67 | 0.60 | 0.62
Gemma3-4b	0.97 | 0.9 | **0.93**	0.46 | 0.5 | 0.48	0.77 | 0.75 | 0.76	0.47 | 0.49 | 0.48	0.79 | 0.78 | 0.78	0.5 | 0.47 | 0.49	0.95 | 0.6 | 0.64
Gemma3-12b	0.91 | 0.88 | 0.90	0.72 | 0.65 | **0.68**	0.80 | 0.83 | **0.81**	0.74 | 0.74 | **0.74**	0.79 | 0.80 | **0.79**	1 | 1 | **1**	0.96 | 0.70 | **0.76**

Table 10: Macro averaged precision, recall, and F1 scores of finetuned models on the discourse quality indicators measured on the balanced Berlin Wall/Day of Europe. The best F1 scores are typeset in bold.eset in bold.

Model	Conflict P | R | F1	Respectful P | R | F1	Uncivil P | R | F1	Reciprocity P | R | F1	Reason P | R | F1	Experience P | R | F1	Constructiveness P | R | F1
Majority-class baseline	0.28 | 0.5 | 0.36	0.40 | 0.5 |0.44	0.35 | 0.5 | 0.41	0.39 | 0.5 | 0.44	0.27 | 0.5 | 0.35	0.44 | 0.5 | 0.47	0.28 | 0.5 | 0.36
GPT-5 mini	0.28 | 0.5 | 0.36	0.49 | 0.48 | 0.48	0.33 | 0.35 | 0.34	0.57 | 0.65 | 0.60	0.76 | 0.65 | 0.70	0.5 | 0.15 | 0.23	0.68 | 0.60 | 0.64
Gemma3-4b	0.86 | 0.83 | **0.85**	0.65 | 0.67 | 0.66	0.76 | 0.77 | 0.77	0.78 | 0.77 | 0.77	0.75 | 0.75 | 0.75	0.76 | 0.81 | 0.78	0.66 | 0.67 | 0.66
Gemma3-12b	0.84 | 0.82 | 0.83	0.87 | 0.78 | **0.81**	0.84 | 0.85 | **0.84**	0.87 | 0.89 | **0.88**	0.80 | 0.78 | **0.79**	0.93 | 0.94 | **0.93**	0.78 | 0.78 | **0.78**

Table 11: Macro averaged precision, recall, and F1 scores of finetuned models on the discourse quality indicators measured on the non-balanced Giorno del ricordo dataset. The best scores are typeset in bold.

Model	Conflict P | R | F1	Respectful P | R | F1	Uncivil P | R | F1	Reciprocity P | R | F1	Reason P | R | F1	Experience P | R | F1	Constructiveness P | R | F1
Majority-class baseline	0.36 | 0.5 | 0.42	0.4 | 0.5 | 0.44	0.41 | 0.5 | 0.45	1 | 1 | **1**	0.44 | 0.5 | 0.47	1 | 1 | **1**	0.31 | 0.5 | 0.38
GPT-5 mini	0.36 | 0.5 | 0.42	0.18 | 0.18 | 0.18	0.43 | 0.43 | 0.43	0.5 | 0.43 | 0.46	0.64 | 0.69 | 0.65	0.5 | 0.14 | 0.22	0.23 | 0.5 | 0.32
Gemma3-4b	0.87 | 0.93 | 0.89	0.86 | 0.79 | **0.81**	0.96 | 0.79 | 0.84	1 | 1 | **1**	0.6 | 0.6 | 0.6	1 | 1 | **1**	0.65 | 0.62 | 0.62
Gemma3-12b	0.89 | 0.92 | **0.91**	0.83 | 0.80 | **0.81**	0.98 | 0.93 | **0.95**	1 | 1 | **1**	0.80 | 0.83 | **0.81**	1 | 1 | **1**	0.75 | 0.72 | **0.72**

Aligned language models

After showing that the finetuned models perform reasonably well at classifying discourse indicators, we conduct the second set of evaluations. We asses the generative ability of models adapted to specific discourse indicators described in Section "Aligning large language models for political discourse". To this end, we compare them to the non-finetuned version of the model to determine whether adaptation on a small set of task specific data can improve their discourse orientation.

As the evaluation in Section "Discourse quality analysis" shows, our finetuned models are reasonably accurate in detecting discourse quality indicators. Therefore, we use these models to perform an automatic evaluation. While this is not completely reliable, it gives a statistically significant insight into the differences between various models. For this evaluation, we generate 200 messages across every topic in our dataset with each model. A detailed list of topics is shown in Appendix 2 (Škvorc et al., 2026). We then use the finetuned Gemma-4b models (finetuned on the balanced Berlin Wall and *Giorno del ricordo* dataset) from Section "Discourse quality analysis models" to assess the 7 discourse quality indicators for both commemorations. In Tables 13 and 14, we show the performance of the finetuned models described in Section "Aligning large language models for political discourse", where each model is aligned to produce messages that match a specific discourse quality indicator.

The results show that finetuning a model increases the presence of the target indicator, even when the

Table 12: Macro averaged precision, recall, and F1 scores of finetuned models on the discourse quality indicators measured on the balanced Giorno del ricordo dataset. The best F1 scores are typeset in bold.

Model	Conflict P | R | F1	Respectful P | R | F1	Uncivil P | R | F1	Reciprocity P | R | F1	Reason P | R | F1	Experience P | R | F1	Constructiveness P | R | F1
Majority-class baseline	0.30 | 0.50 | 0.37	0.40 | 0.5 | 0.44	0.34 | 0.5 | 0.41	0.39 | 0.5 | 0.44	0.27 | 0.5 | 0.35	0.34 | 0.5 | 0.47	0.27 | 0.5 | 0.35
GPT-5 mini	0.30 | 0.50 | 0.37	0.14 | 0.22 | 0.15	0.23 | 0.22 | 0.23	0.45 | 0.42 | 0.43	0.72 | 0.67 | 0.64	0.51 | 0.54 | 0.26	0.47 | 0.47 | 0.46
Gemma3-4b	0.93 | 0.95 | **0.94**	0.89 | 0.90 | **0.89**	0.87 | 0.86 | **0.87**	0.92 | 0.91 | 0.91	0.79 | 0.78 | **0.78**	0.82 | 0.82 | 0.82	0.77 | 0.76 | 0.76
Gemma3-12b	0.99 | 0.83 | 0.89	0.83 | 0.85 | 0.84	0.85 | 0.85 | 0.85	0.91 | 0.93 | **0.92**	0.83 | 0.76 | 0.77	0.80 | 0.90 | 0.84	0.80 | 0.80 | **0.80**
Gemma3-27b	0.70 | 0.64 | 0.65	0.72 | 0.68 | 0.70	0.84 | 0.84 | 0.84	0.76 | 0.74 | 0.75	0.75 | 0.75 | 0.75	0.81 | 0.92 | **0.86**	0.75 | 0.74 | 0.75

Table 13: The detected presence of discourse quality indicators (in columns) after finetuning the Gemma3-4b model to align with specific discourse quality indicators (in rows) on the Berlin Wall/Day of Europe dataset. Each value represents the percentage of posts containing a given indicator. The scores of the best performing models are typeset in bold. The baseline score of Gemma3-4b is for the non-adapted LLM.

Model	Conflict	Respectful	Uncivil	Reason	Experience	Constructiveness	Reciprocity
Gemma3-4b	0.563	0.132	0.023	0.350	0.041	0.127	0.017
4b-Conflict	**0.846**	0.050	0.045	0.612	0.020	0.124	0.005
4b-Respectful	0.148	**0.529**	0.006	0.155	0.071	**0.335**	0.010
4b-Uncivil	0.880	0.019	**0.089**	0.612	0.156	0.052	0.005
4b-Reason	0.756	0.065	0.050	**0.726**	0.010	0.174	0.010
4b-Experience	0.418	0.259	0.005	0.592	**0.657**	0.144	0.005
4b-Constructiveness	0.637	0.149	0.029	0.622	0.025	0.289	0.030
4b-Reciprocity	0.575	0.035	0	0.225	0.005	0.065	**0.565**

Table 14: The detected presence of discourse quality indicators (in columns) after finetuning the Gemma3-4b model to align with specific discourse quality indicators (in rows) on the Giorno del ricordo dataset. Each value represents the percentage of posts containing a given indicator. The scores of the best performing models are typeset in bold. The baseline score of Gemma3-4b is for the non-adapted LLM.

Model	Conflict	Respectful	Uncivil	Reason	Experience	Constructiveness	Reciprocity
Gemma3-4b	0.721	0.353	0.174	0.403	0.005	0.284	0.129
4b-Conflict	**0.945**	0.095	0.234	0.443	0.010	0.284	0.119
4b-Respectful	0.284	**0.846**	0.005	0.214	0.015	0.378	0.005
4b-Uncivil	0.975	0.055	**0.532**	0.408	0.005	0.214	0.119
4b-Reason	0.886	0.224	0.219	0.408	0.010	0.373	0.010
4b-Experience	0.828	0.414	0.086	**0.539**	**0.570**	0.234	0.086
4b-Constructiveness	0.866	0.313	0.149	0.547	0.005	**0.542**	0.075
4b-Reciprocity	0.960	0.085	0.040	0.124	0.005	0.109	**0.701**

models were trained on automatically-generated labels produced by a non-finetuned ChatGPT-5.2 model. This means that LLMs can be aligned towards specific indicators through a purely automatic approach, without relying on manual annotation.

While this holds for every indicator, it is more prominent for certain indicators. For example, "conflict" reaches a value of 0.846 on the Berlin Wall/Day of Europe dataset and a value of 0.945 on the *Giorno del ricordo* dataset, while "constructiveness" reaches smaller values (0.335 and 0.542 respectively). This can be explained either through the examples present in our data (where the "conflict" indicator is more common) or through the implicit biases present in the base LLMs. Additionally, the results show a correlation between different indicators. For example, increasing the conflict indicator also increases the "uncivil" and "reason" indicators on both datasets.

From a standpoint of deliberation, these results also show that antagonistic indicators such as "conflict" and "uncivil" can be reduced by training a model on an inversely-correlated indicator (e.g., "respectful"), which can be useful for training models that avoid generating messages with such unwanted indicators.

Qualitative analysis of generated posts

In addition to the automatic evaluation of the aligned LLM models, we perform a small-scale manual, qualitative analysis of the posts generated by each model. For each commemoration and indicator, we chose 3 randomly examples for manual analysis (for a total of 42 posts). We we provide English translations of these posts in Tables 15 and 16. We remove hash symbols and at tags from the posts as part of data preprocessing, so they do not appear in the generated posts.

We provide English translations for a further list of 140 generated posts (10 for each indicator for both the the Berlin Wall/Day of Europe and the *Giorno del ricordo* datasets) in Appendix 3 (Škvorc et al., 2026). Then, we manually analyse each chosen post using the following criteria:

- Is the post grammatically sensible?
- Is the post related and contextually accurate to the desired commemoration?
- Is the post aligned with the desired indicator?

Additionally, we check the overall similarity of the chosen posts to ensure the models are not simply repeating a small number of phrases that would score well with the evaluation models.

For the Berlin Wall/Day of Europe dataset, most of the chosen posts meet the desired criteria. Posts with **conflict** focus on the fall of the Berlin Wall, communism, and its relation to modern-day politics surrounding Russia and Europe, although an unambiguous link to conflict is not always present. All the posts are grammatically sensible and contextually accurate to the commemoration. Day of Europe is rarely mentioned, likely due to the fact that it is a less contentious commemoration.

Conversely, **Respectful** posts relate to the day of Europe and clearly express respect and empathy. As with conflict, **Uncivil** posts are related to the fall of the Berlin Wall but include clear disrespect and exaggeration.

Justification with reason produces posts that include arguments based on history (in the case of the Berlin wall) or the values of Europe day. Posts that justify their arguments with **experience** contain specific phrases that signify this ("I remember", "I was there", "I have seen" ...), which could be seen as the model over-fitting to a few specific key words and may be a reflection of the small amount of training examples. Posts here also show the least amount of diversity.

Constructive posts appear to be seeking common ground around topics related to the Berlin wall and different current-day events. With **reciprocity** all the chosen posts ask some sort of question, which fits the definition of the indicator.

For *Giorno del ricordo*, the results are similar. Posts generated with the **conflict** model explicitly mention the conflict with the communist partisans (or more broadly, the left), implicitly expressing disagreement with one of the sides. The mentions of Giorgia Meloni relate the discussion to current Italian politics.

Respectful posts focus on memorializing the victims of the foibe, without mentions of contentious political events. **Uncivil** posts are similar to those produced by the conflict model, but also sometimes include explicit attacks against another person. The last post does not contain conflict, reflecting the fact that the models are not 100% accurate.

The **justification with reason** model performed poorly in the automatic analysis, which is reflected in the generated posts. One provides an explicit justification, while the other two do not. When compared with the Berlin Wall/Day of Europe dataset, justification with **experience** centres less around living during the event. The event commemorated by *Giorno del ricordo* is more distant, leading to fewer people with personal experience. Instead, an exhibition related to the foibe is mentioned by one post.

Two of the **constructive** posts seek common ground by talking about both sides of the conflict. Of the **reciprocal** posts, two ask questions related to the event, although the lack of context makes it hard to determine whether these questions are genuine.

All of the examined posts are grammatically sensible, though a further list of 10 posts per indicator shows the models are not perfect and sometimes simply generate a list of usernames. Most posts are related to the desired commemorations and include the correct historical and political contexts. As demonstrated by the automatic analysis, not all posts are correctly aligned with the desired indicator, but when they are this is not simply a result of repeating a number of repetitive key words (with the possible exception of the Berlin Wall justification with reason). We also verified that the generated posts do not match those present in the training data, meaning that the models are not simply repeating examples seen during training.

CONCLUSION AND FURTHER WORK

In this work, we presented how automatic techniques based in natural language processing and large language models can be used both for deliberative analysis of social media messages, and as a way of aligning LLM-produced messages towards specific types of discourse.

The results of our work show two cases where finetuning LLMs can be useful for discourse analysis. First, we show that finetuning LLMs on manually-labelled datasets can improve their performance when it comes to detecting discourse indicators. While non-fine tuned models already perform reasonably well on this task, finetuning further increases accuracy, allowing for a more precise discourse analysis of social media messages. Second, we show that finetuning models on specific indicators, even when the indicators were automatically labelled using non-fine tuned LLMs can align their outputs to more closely match the chosen indicator. This allows the models to produce messages that contain less conflict or uncivilly and more respect or constructiveness.

However, our evaluation contains several limitations. We limit our work to specific commemorations (Day of Europe, Fall of the Berlin Wall, and *Giorno del ricordo*) and not to general social media messages. We also limit our evaluation to the Gemma 3 family of models, comparing them with GPT-5 as a baseline. Further work is necessary to determine how these findings generalize to other LLMs. Additionally, a larger-scale qualitative analysis is necessary to evaluate the quality of the automatically generated posts. Such an analysis would require a larger number of human evaluators and was not possible in the scope of this study.

In future research, our work could lead to a development of practical applications aimed at improving political and public discourse, for example, by generating respectful replies guiding uncivil discourse to more desirable outcomes, suggesting rewrites of offensive posts with more deliberative, civil alternatives, or serving as human-assisted moderator agents and discourse quality monitoring tools. However, care must be taken to ensure such models are deployed ethically and with proper human supervision.

Table 15: Tweets generated using the aligned models trained on the Berlin Wall/Day of Europe datasets. We show 3 randomly-chosen posts per indicator.

Conflict
The former SED, which has already been banned for 30 years since the fall of the Berlin Wall, is now again on the street. You can do it like this for decades, but in the end there will be a fall of the wall, as in the fall of the Berlin Wall. If you want to know what the Russians are doing in Europe and why, read this article on the fall of the Berlin wall, the US has always had an interest in it
Respectful
On May 9th, we celebrate Europe Day. This day commemorates the signing of the Schuman declaration in 1950, the founding text of the European construction. Thank you to all those who came to the Stadtschlossgese in the Humboldtstadt! Our EuropeDay was a great success! The weather was beautiful and the visitors enjoyed the EuropeDay program and the EuropeanElections2024. On the occasion of Europe Day, the students of the 2nd grade of the 1st cycle of the AixEnProvence school have been able to discover the European Union through an exhibition organized in partnership with the French Institute and the Erasmus+ program.
Uncivil
The Moroccan flag has been banned in France, the Eiffel Tower has been covered with the Algerian flag, the Berlin Wall has been replaced by the Moroccan flag, the statue of the Eiffel Tower has been replaced by the statue of the king of Morocco, and so on. What do you mean? It is the fault of the Germans that the Berlin Wall fell? You are a ridiculous, stupid person. And you want to talk about the fall of the Berlin Wall? A wall that was built to protect us from the communist dictatorship. It was built by the people. The people of the West. And the communists want to build it again.
Justification with Reason
The USA has never respected any agreements made, the USA and NATO have repeatedly broken the agreements made with the USSR after the fall of the Berlin Wall, the NATO expansion to the East is a violation of the Minsk agreements The Russians invaded Ukraine in 2014, and then the Berlin Wall fell in 1989. Europe Day: celebrates the diversity of its cultures and languages. The diversity of our cultures and languages is a strength. It is a guarantee of our future. is a land of culture and languages is a land of diversity.
Justification with Experience
I was there 1989, I saw the Iron Curtain fall, I felt the people of the East crying and laughing. I saw the fall of the Berlin Wall and the joy of reunification. Today I see the iron curtain rise again. I have seen the fall of the Berlin Wall. I have seen the rise of the Russian Federation. I have seen the rise of the Chinese Federation. I have seen the rise of the Iranian Federation. I have seen the rise of the Korean Federation. I have seen the rise of the Turkish Federation. I remember the fall of the Berlin Wall with the great joy of a child. We were all excited, I was 20 years old. But there is also the sadness of the people who lived in the East, who did not want to leave their homes, their lives.
Constructiveness
I have lived in the GDR, I have seen the fall of the wall, I have been there when the Russians came, I have seen the suffering, the poverty, the misery, the injustice. I have never experienced such a thing. What do you know about it? You can write about Putin all day long, but you do not know what you are talking about. The former SED rulers have only been replaced, the SED is not dissolved. There are only political names that have changed. It is time to recognize the State of Palestine and we must put an iron curtain and a wall up to Jerusalem for many years to protect Israel from these massacres.
Reciprocity
Does the French government have a plan to close the border with the Maghreb? I have not seen any plan to create a wall between France and the Maghreb. Why was the Berlin Wall built? Today is Europe Day, and I am wondering if the EU is still the great project of unity that we thought it was.

Table 16: Tweets generated using the aligned models trained on the Giorno del ricordo datasets. We show 3 randomly-chosen posts per indicator.

Conflict
GiorgiaMeloni The victims of the foibe were something horrible, but the Italian communist partisans were a despicable bunch of people. It is true, the foibe were an atrocity and the massacres carried out by Tito's followers and Tito's communists must never be forgotten. But one cannot even remember the atrocities committed by the Italians in those lands. The left mobilizes for the Day of Remembrance, while Meloni does nothing. The Day of Memory is February 10, the Day of Remembrance of the foibe is February 10.You can do it like this for decades, but in the end there will be a fall of the wall, as in the fall of the Berlin Wall. If you want to know what the Russians are doing in Europe and why, read this article on the fall of the Berlin wall, the US has always had an interest in it
Respectful
So as not to forget the Day of Remembrance. In memory of the Martyrs of the foibe. My thoughts go to the innocent victims who paid the highest price of hatred and violence. Mattarella: "The foibe are one of the darkest pages in our history"
Uncivil
And what do you mean? Are you saying that your friends, the partisans, did not kill in the foibe anyone who was communist or left-wing? Do you not know what you are talking about? You are acting like the fascists who say that the Nazis killed all civilians, and more. And what about the foibe? The victims of the foibe? Do you know that they have 15,000 dead? But you're a heartless person, I believe it. The foibe are an unquestionable tragedy, but they do not justify the killing of thousands of Slavic civilians. One cannot justify the killing of civilians with the killing of civilians. Therefore, the tragedy was a tragedy, but it does not justify the tragedy.
Justification with Reason
The right has taken the liberty of commemorating the Day of Remembrance, which is not the Day of Remembrance but the Day of Shame. I remember that it was fascism that invaded Yugoslavia, massacred and deported entire populations, built concentration camps and had its headquarters in Basovizza, carried out ethnic cleansing, built extermination camps and caused the death of 10 million Slavs. The foibe were a disgrace for our nation. Today the foibe are our revenge. The foibe have taught us never to take anything for granted.
Justification with Experience
I saw an exhibition yesterday about the Foibe and the remembrance of the victims of fascism. And it made me feel a bit uneasy. My grandfather took part in one of the two. And I knew him. The memory of those times made me cry. I was in Fiume, I also saw the foibe, as a child, for the first time, I saw the funeral of a friend of mine, who had been taken away, and I understood how much fascism was a disgrace. It cannot be denied that the Foibe massacres were a tragedy, but it is another matter to deny that, in a wartime context, there were also the Foibe killings, with partisans who killed Italians.
Constructiveness
February 10 is the day in which, every year, we remember the foibe massacres and the exodus of thousands of Italians from the lands of Istria, Fiume and Dalmatia. It is the day in which we remember the victims, of all nationalities, massacred in the border areas Today is the Day of Remembrance. Forgetting the victims of the foibe massacres is a serious mistake. Whoever denies or minimizes the horror of the occupation and the forced exodus of our Slavic fellow citizens, is not antifascist, but a denier. You fascists do not remember history. The problem is that no one wants to understand that the foibe are a consequence of what the Italians did in Yugoslavia. Not the foibe, but the 20 years of Fascist occupation. Remember that without fascism there would have been no foibe.
Reciprocity
A Pasolini film came to my mind, I don't know the title. What do the foibe massacres have to do with the upcoming elections? The foibe are a historical tragedy, but they are not a political topic. What does it have to do with Foibe Remembrance Day?

ZAČETEK RAZVOJA AGENTOV UMETNE INTELIGENCE ZA IZBOLJŠANO KVALITETO POLITIČNEGA DISKURZA Z VELIKIMI JEZIKOVNIMI MODELI

Tadej ŠKVORC
Inštitut IRRIS za raziskave, razvoj in strategije družbe, kulture in okolja, Centur 1f, 6273 Marezige, Slovenija
e-mail: tadej.skvorc@irris.eu

Marjan HORVAT
Inštitut IRRIS za raziskave, razvoj in strategije družbe, kulture in okolja, Centur 1f, 6273 Marezige, Slovenija
e-mail: marjan.horvat@irris.eu

Jure KORAŽIJA
Inštitut IRRIS za raziskave, razvoj in strategije družbe, kulture in okolja, Centur 1f, 6273 Marezige, Slovenija
e-mail: jure.korazija@irris.eu

Marko ROBNIK-ŠIKONJA
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana, Slovenija
e-mail: marko.robniksikonja@fri.uni-lj.si

POVZETEK

Razvoj velikih jezikovnih modelov omogočajo poglobljeno in obsežno analizo številnih kompleksnih pojavov. Eno takšnih področij je politični diskurz v družbenih medijih, ki lahko služi kot kazalnik številnih družbenih vprašanj. V tem delu analiziramo, kako se lahko učenje velikih jezikovnih modelov na objavah v družbenih medijih, povezanih z razpravo o določenih političnih spominskih slovesnostih, uporabi tako za pomoč pri analizi diskurza kot tudi za prilagajanje sporočil, ki jih ustvarjajo ti modeli (npr. spreminjanje sporočil iz nevljudnih v vljudna). Najprej predstavimo teoretični okvir za analizo političnega diskurza, ki temelji na osmih kazalnikih kakovosti diskurza, čemur sledita dva metodološka prispevka. Najprej pokažemo, da so modeli naučeni na majhnem številu ročno označenih primerov boljši pri zaznavanju kazalnikov kakovosti diskurza kot večji, neprilagojeni modeli. Nato modele natančno prilagodimo na primerih, ki ustrezajo določenim kazalnikom kakovosti diskurza, in pokažemo, kako lahko ta proces preusmeri sporočila, ki jih ti modeli ustvarjajo, tako da se bolj uskladijo z želenim kazalnikom, npr. vljudnostjo. Ugotovitve kažejo, da bi se v prihodnosti agenti umetne inteligence, ki temeljijo na velikih jezikovnih modelih, lahko uporabljali za moderiranje javnega diskurza in izboljšanje njegove kakovosti.

Ključne besede: politični diskurz, agenti umetne inteligence, politične spominske slovesnosti, Berlinski zid, Dan Evrope, Veliki jezikovni modeli

SOURCES AND BIBLIOGRAPHY

Achiam, Josh, Adler, Steven, Agarwal, Sandhini, Ahmad, Lama, Akkaya, Ilge, Aleman, Florencia Leoni et al. (2023): GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (last access: 2026-06-23).

Barberá, Pablo (2015): Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. Political Analysis, 23, 1, 76–91.

Bächtiger, André, Shikano, Susumu, Pedrini, Seraina & Mirjam Ryser (2009): Measuring Deliberation 2.0: Standards, Discourse Types, and Sequenzialization. In: ECPR General Conference. Potsdam, 5–12.

Beauchamp, Nick (2020): Modeling and Measuring Deliberation Online. In: Foucault Welles, Brooke & Sandra González-Bailón (eds.): The Oxford Handbook of Networked Communication. Oxford, Oxford University Press, 321–349.

Boyd, Danah (2011): Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications. In: Papacharissi, Zizi (ed.): A Networked Self: Identity, Community, and Culture on Social Network Sites. New York, Routledge, 39–58.

Brown, Tom B., Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared D., Dhariwal, Prafulla et al. (2020): Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Canute, Matt, Jin, Mali, Holtzclaw, Hannah, Lusoli, Alberto, Adams, Philippa, Pandya, Mugdha, Taboada, Maite, Maynard, Diana & Wendy Hui Kyong Chun (2023): Dimensions of Online Conflict: Towards Modeling Agonism. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore, Association for Computational Linguistics, 12194–12209.

Casper, Stephen, Bailey, Luke, Hunter, Rosco, Ezell, Carson, Cabalé, Emma, Gerovitch, Michael, Slocum, Stewart, Wei, Kevin, Jurkovic, Nikola, Khan, Ariba, Christoffersen, Phillip J. K., Ozisik, A. Pinar, Trivedi, Rakshit, Hadfield-Menell, Dylan & Noam Kolt (2025): The AI Agent Index. arXiv preprint arXiv:2502.01635 (last access: 2026-06-23).

Cinelli, Matteo, De Francisci Morales, Gianmarco, Galeazzi, Alessandro, Quattrociocchi, Walter & Michele Starnini (2021): The Echo Chamber Effect on Social Media. Proceedings of the National Academy of Sciences, 118, 9, e2023301118.

Coe, Kevin, Kenski, Kate & Stephen A. Rains (2014): Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. Journal of Communication, 64, 4, 658–679.

Conover, Michael D., Ratkiewicz, Jacob, Francisco, Matthew, Gonçalves, Bruno, Menczer, Filippo & Alessandro Flammini (2011): Political Polarization on Twitter. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 5, 1, 89–96.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton & Kristina Toutanova (2018): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805 (last access: 2026-06-23).

Ferrara, Emilio, Varol, Onur, Davis, Clayton, Menczer, Filippo & Alessandro Flammini (2016): The Rise of Social Bots. Communications of the ACM, 59, 7, 96–104.

Fortuna, Paula & Sérgio Nunes (2018): A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys (CSUR), 51, 4, 1–30.

Fournier-Tombs, Eleonore (2024): An Ethical Grey Zone: AI Agents in Political Deliberations. https://carnegiecouncil.org/media/article/ethical-grey-zone-ai-agents-political-deliberation (last access: 2026-06-23).

Fournier-Tombs, Eleonore & Michael K. MacKenzie (2021): Big Data and Democratic Speech: Predicting Deliberative Quality Using Machine Learning Techniques. Methodological Innovations, 14, 2.

Fraser, Nancy (1990): Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. Social Text, 25/26, 56–80.

Gillespie, Tarleton (2018): Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. New Haven, Yale University Press.

Grimmer, Justin & Brandon M. Stewart (2013): Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, 21, 3, 267–297.

Gutman, Yifat & Jenny Wüstenberg (eds.) (2023): The Routledge Handbook of Memory Activism. London, Routledge.

Habermas, Jürgen (1989): The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society. Cambridge, Polity.

Habermas, Jürgen (1996): Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy. Cambridge, MIT Press.

Horvat, Marjan & Jure Koražija (2026): Europe Day and the Fall of the Berlin Wall on Twitter/X: Conflict, Tone, and Deliberative Quality Across France, Germany, Italy, and Slovenia. https://doi.org/10.5281/zenodo.18770520 (last access: 2026-06-23).

Ilyas, Sanaa & Qamar Khushi (2012): Facebook Status Updates: A Speech Act Analysis. Academic Research International, 3, 2, 500–507.

Iyengar, Shanto, Lelkes, Yphtach, Levendusky, Matthew, Malhotra, Neil & Sean J. Westwood (2019): The Origins and Consequences of Affective Polarization in the United States. Annual Review of Political Science, 22, 1, 129–146.

Jaidka, Kokil (2022): Talking Politics: Building and Validating Data-Driven Lexica to Measure Political Discussion Quality. Computational Communication Research, 4, 2, 486–527.

Jungherr, Andreas & Adrian Rauchfleisch (2025): Artificial Intelligence in Deliberation: The AI Penalty and the Emergence of a New Deliberative Divide. Government Information Quarterly, 42, 4, 102079.

Kamath, Aishwarya, Ferret, Johan, Pathak, Shreya, Vieillard, Nino et al. (2025): Gemma 3 Technical Report. arXiv preprint arXiv:2503.19786 (last access: 2026-06-23).

Lampe, Urška, Horvat, Marjan, Koražija, Jure, Ergaver, Angelika & Darko Darovec (2026): Agonistic Engagement in Memory Politics: Media Arenas, Normative Orientations, and Debates on Giorno del Ricordo in Italy and Slovenia. https://doi.org/10.5281/zenodo.18770931 (last access: 2026-06-23).

Laver, Michael, Benoit, Kenneth & John Garry (2003): Extracting Policy Positions from Political Texts Using Words as Data. American Political Science Review, 97, 2, 311–331.

Liu, Aixin, Feng, Bei, Xue, Bing, Wang, Bingxuan, Wu, Bochao, Lu, Chengda. et al. (2024): DeepSeek-V3 Technical Report. arXiv preprint arXiv:2412.19437 (last access: 2026-06-23).

Mansbridge, Jane, Bohman, James, Chambers, Simone, Christiano, Thomas, Fung, Archon, Parkinson, John & Mark E. Warren (2012): A Systemic Approach to Deliberative Democracy. In: Parkinson, John & Jane Mansbridge (eds.): Deliberative Systems: Deliberative Democracy at the Large Scale. Cambridge, Cambridge University Press, 1–26.

Marcellino, William, Beauchamp-Mustafaga, Nathan, Kerrigan, Amanda, Chao, Lev Navarre & Jackson Smith (2023): The Rise of Generative AI and the Coming Era of Social Media Manipulation 3.0: Next-Generation Chinese Astroturfing and Coping with Ubiquitous AI. https://www.rand.org/pubs/perspectives/PEA2679-1.html (last access: 2026-06-23).

Marone, Marc, Weller, Orion, Fleshman, William, Yang, Eugene, Lawrie, Dawn & Benjamin Van Durme (2025): mmBERT: A Modern Multilingual Encoder with Annealed Language Learning. arXiv preprint arXiv:2509.06888. https://arxiv.org/abs/2509.06888 (last access: 2026-06-23).

McPherson, Miller, Smith-Lovin, Lynn & James M. Cook (2001): Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology, 27, 1, 415–444.

Meyer, Erik (2008): Memory and Politics. In: Erll, Astrid & Ansgar Nünning (eds.): Cultural Memory Studies: An International and Interdisciplinary Handbook. Berlin, Walter de Gruyter, 173–180.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg & Jeff Dean (2013): Distributed Representations of Words and Phrases and Their Compositionality. Advances in Neural Information Processing Systems, 26. arXiv preprint arXiv:1310.4546. https://arxiv.org/abs/1310.4546 (last access: 2026-06-23).

OpenAI (2026): Predstavljamo GPT-5.2. https://openai.com/index/introducing-gpt-5-2/ (last access: 2026-06-23).

Papacharissi, Zizi (2004): Democracy Online: Civility, Politeness, and the Democratic Potential of Online Political Discussion Groups. New Media & Society, 6, 2, 259–283.

Parthasarathy, Venkatesh Balavadhani, Zafar, Ahtsham, Khan, Aafaq & Arsalan Shahid (2024): The Ultimate Guide to Finetuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. arXiv preprint arXiv:2408.13296. https://arxiv.org/abs/2408.13296 (last access: 2026-06-23).

Russell, Stuart & Peter Norvig (2003): Artificial Intelligence: A Modern Approach. Upper Saddle River, Pearson Education.

Saha, Tulika, Saha, Sriparna & Pushpak Bhattacharyya (2019): Tweet Act Classification: A Deep Learning Based Classifier for Recognizing Speech Acts in Twitter. In: 2019 International Joint Conference on Neural Networks (IJCNN), 1–8.

Schick, Timo, Dwivedi-Yu, Jane, Dessì, Roberto, Raileanu, Roberta, Lomeli, Maria, Zettlemoyer, Luke, Cancedda, Nicola & Thomas Scialom (2023): Toolformer: Language Models Can Teach Themselves to Use Tools. Advances in Neural Information Processing Systems, 36, 68539–68551.

Searle, John R. (1969): Speech Acts: An Essay in the Philosophy of Language. Cambridge, Cambridge University Press.

Steenbergen, Marco R., Bächtiger, André, Spörndli, Markus & Jürg Steiner (2003): Measuring Political Deliberation: A Discourse Quality Index. Comparative European Politics, 1, 1, 21–48.

Sumers, Theodore R., Yao, Shunyu, Narasimhan, Karthik R. & Thomas L. Griffiths (2024): Cognitive Architectures for Language Agents. Transactions on Machine Learning Research. arXiv preprint arXiv:2309.02427. https://arxiv.org/abs/2309.02427 (last access: 2026-06-23).

Sunstein, Cass R. (2018): #Republic: Divided Democracy in the Age of Social Media. Princeton, Princeton University Press.

Škvorc, Tadej, Horvat, Marjan, Koražija, Jure, & Robnik-Šikonja, Marko (2026): Appendix: LLM prompt, List of Topics and List of generated posts for „Towards future AI agents for improved political discourse quality with large language models". Zenodo. https://doi.org/10.5281/zenodo.20817044 (last access: 2026-06-23).

Tessler, Michael Henry, Bakker, Michiel A., Jarrett, Daniel, Sheahan, Hannah, Chadwick, Martin J., Koster, Raphael, Evans, Georgina, Campbell-Gillingham, Lucy, Collins, Tantum, Parkes, David C., Botvinick, Matthew & Christopher Summerfield (2024): AI Can Help Humans Find Common Ground in Democratic Deliberation. Science, 386, 6719.

Touvron, Hugo, Martin, Louis, Stone, Kevin, Albert, Peter, Almahairi, Amjad, Babaei, Yasmine et al. (2023): Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288 (last access: 2026-06-23).

Törnberg, Petter (2025): Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages. Social Science Computer Review, 43, 6, 1181–1195.

Tucker, Joshua A., Guess, Andrew, Barberá, Pablo, Vaccari, Cristian, Siegel, Alexandra, Sanovich, Sergey, Stukal, Denis & Brendan Nyhan (2018): Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3144139 (last access: 2026-06-23).

Vosoughi, Soroush, Roy, Deb & Sinan Aral (2018): The Spread of True and False News Online. Science, 359, 6380, 1146–1151.

Wei, Jason, Wang, Xuezhi, Schuurmans, Dale, Bosma, Maarten, Ichter, Brian, Xia, Fei, Chi, Ed, Le, Quoc & Denny Zhou (2022): Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems, 35, 24824–24837.

Weizenbaum, Joseph (1976): Computer Power and Human Reason: From Judgment to Calculation. San Francisco, W. H. Freeman and Company.

Wüstenberg, Jenny (2017): Civil Society and Memory in Postwar Germany. Cambridge, Cambridge University Press.

Zhang, Renxian, Dehong Gao & Wenjie Li (2011): What Are Tweeters Doing: Recognizing Speech Acts in Twitter. Analyzing Microtext: Papers from the 2011 AAAI Workshop (WS-11-05). Palo Alto, AAAI Press, 86–91.

Ziems, Caleb, Held, William, Shaikh, Omar, Chen, Jiaao, Zhang, Zhehao & Diyi Yang (2024): Can Large Language Models Transform Computational Social Science?. Computational Linguistics, 50, 1, 237–291.