Overview of Touché 2025: Argumentation Systems*

Extended Version

Johannes Kiesel[1,**], Çağrı Çöltekin[2], Marcel Gohsen[3], Sebastian Heineking[4], Maximilian Heinrich[3], Maik Fröbe[5], Tim Hagen[6,7], Mohammad Aliannejadi[8], Sharat Anand[3], Tomaž Erjavec[9], Matthias Hagen[5], Matyáš Kopp[10], Nikola Ljubešić[9], Katja Meden[9], Nailia Mirzakhmedova[3], Vaidas Morkevičius[11], Harrisen Scells[2], Moritz Wolter[4], Ines Zelch[4,5], Martin Potthast[6,7,12] and Benno Stein[3]

[1]*GESIS - Leibniz Institute for the Social Sciences*
[2]*University of Tübingen*
[3]*Bauhaus-Universität Weimar*
[4]*Leipzig University*
[5]*Friedrich-Schiller Universität Jena*
[6]*University of Kassel*
[7]*hessian.AI*
[8]*University of Amsterdam*
[9]*Jožef Stefan Institute*
[10]*Charles University*
[11]*Kaunas University of Technology*
[12]*ScaDS.AI*

Abstract

This paper is the extended overview of Touché: the sixth edition of the lab on argumentation systems that was held at CLEF 2025. With the goal to foster the development of support-technologies for decision-making and opinion-forming, we organized four shared tasks: (1) Retrieval-Augmented Debating (RAD), in which participants submit generative retrieval systems that argue against their users and evaluate such systems (new task); (2) Ideology and Power Identification in Parliamentary Debates, in which participants identify from a speech the political leaning of the speaker's party and whether it was governing at the time of the speech (2nd edition); (3) Image Retrieval/Generation for Arguments, in which participants find images to convey a written argument (4th edition, joint task with ImageCLEF); and (4) Advertisement in Retrieval-Augmented Generation, in which participants generate responses to queries with ads inserted and detect such inserted ads (new task). In this paper, we describe these tasks, their setup, and participating approaches in detail.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*This overview extends the one published as part of the CLEF 2025 proceedings [1]
**Corresponding author
johannes.kiesel@uni-weimar.de (J. Kiesel); ccoltekin@sfs.uni-tuebingen.de (Çağrı Çöltekin); marcel.gohsen@uni-weimar.de (M. Gohsen); sebastian.heineking@uni-leipzig.de (S. Heineking); maximilian.heinrich@uni-weimar.de (M. Heinrich); maik.froebe@uni-jena.de (M. Fröbe); tim.hagen@uni-kassel.de (T. Hagen); m.aliannejadi@uva.nl (M. Aliannejadi); sharat.annd@uni-weimar.de (S. Anand); tomaz.erjavec@ijs.si (T. Erjavec); matthias.hagen@uni-jena.de (M. Hagen); kopp@ufal.mff.cuni.cz (M. Kopp); nikola.ljubesic@ijs.si (N. Ljubešić); katja.meden@ijs.si (K. Meden); nailia.mirzakhmedova@uni-weimar.de (N. Mirzakhmedova); vaidas.morkevicius@ktu.lt (V. Morkevičius); harry.scells@uni-leipzig.de (H. Scells); moritz.wolter09@gmail.com (M. Wolter); ines.zelch@uni-jena.de (I. Zelch); martin.potthast@uni-kassel.de (M. Potthast); benno.stein@uni-weimar.de (B. Stein)
https://touche.webis.de (Touché web page)
0000-0002-1617-6508 (J. Kiesel); 0000-0003-1031-6327 (Çağrı Çöltekin); 0000-0002-1020-6745 (M. Gohsen); 0000-0002-7701-3294 (S. Heineking); 0000-0001-5450-8203 (M. Heinrich); 0000-0002-1003-981X (M. Fröbe); 0009-0000-4854-7249 (T. Hagen); 0000-0002-9447-4172 (M. Aliannejadi); 0009-0004-3903-0124 (S. Anand); 0000-0002-1560-4099 (T. Erjavec); 0000-0002-9733-2890 (M. Hagen); 0000-0001-7953-8783 (M. Kopp); 0000-0001-7169-9152 (N. Ljubešić); 0000-0002-0464-9240 (K. Meden); 0000-0002-8143-1405 (N. Mirzakhmedova); 0000-0002-2174-0396 (V. Morkevičius); 0000-0001-9578-7157 (H. Scells); 0009-0003-1550-835X [21:53] (M. Wolter); 0009-0005-2659-5326 (I. Zelch); 0000-0003-2451-0665 (M. Potthast); 0000-0001-9033-2217 (B. Stein)

1. Introduction

Decision-making and opinion-forming are everyday tasks that involve weighing pro and con arguments for or against different options. With ubiquitous access to all kinds of information on the web, everybody has the chance to acquire knowledge for these tasks on almost any topic. However, current information systems are primarily optimized for returning *relevant* results and do not address deeper analyses of arguments or multi-modality. To close this gap, the Touché lab series, running since 2020, has several tasks to advance both argumentation systems and the evaluation thereof. Previous events and tasks, data, and publications are available at https://touche.webis.de/. The 2025 edition of Touché features the following shared tasks:

1. Retrieval-Augmented Debating (RAD; new task) features two sub-tasks in argumentative agent research of (1) generating responses to argue against a simulated debate partner and (2) evaluating systems of sub-task 1.
2. Ideology and Power Identification in Parliamentary Debates (2nd edition) features three sub-tasks in debate analysis of detecting the (1) orientation on traditional left–right spectrum, (2) position of power of the speaker's party in the governance of the country or the region, and (3) position of the speaker's party on the scale of populism vs. pluralism.
3. Image Retrieval/Generation for Arguments (4th edition; joint task with ImageCLEF [2]) is about finding images to help convey an argument.
4. Advertisement in Retrieval-Augmented Generation (new task) features two sub-tasks in retrieval-augmented generation of (1) generating responses with advertisements inserted and (2) detecting whether a response contains an advertisement.

In total, 12 teams participated in Touché in 2025.

- Two teams participated in the Retrieval-Augmented Debating task (cf. Section 4) and submitted 19 runs. For debating (sub-task 1), the participants employed the provided Elasticsearch API, but used language models for query generation, answer selection, and answer generation. For evaluation (sub-task 2), the participants also focused on prompting language models.
- Four teams participated in the Ideology and Power Identification in Parliamentary Debates task (cf. Section 5) and submitted 20 runs. The approaches used traditional machine learning techniques, fine-tuning of multilingual pretrained models, and prompting large language models, among others.
- Three teams participated in the Image Retrieval/Generation for Arguments task (cf. Section 6), submitting a total of seven runs. The teams employed various approaches, including image retrieval using methods such as `CLIP`, as well as image generation using Stable Diffusion.
- Four teams participated in the Advertisement in Retrieval-Augmented Generation task (cf. Section 7) and submitted 17 runs. All teams participated in the classification sub-task and primarily submitted approaches based on fine-tuned encoder models. The generation sub-task received submissions from three teams that used models from the `Qwen` and `Mistral` families to generate responses from—in some cases re-ranked—lists of relevant document segments.

The corpora, topics, and judgments created at Touché are freely available to the research community on the lab's website.[1] A condensed version of this paper is published in the CLEF 2025 proceedings [1].

2. Background

Argumentation systems are diverse and are connected to many fields within and outside of computer science. The following sections review the related work and background for each Touché task of 2025.

[1] https://touche.webis.de/

2.1. Retrieval-Augmented Debating

Psychological literature has shown that engaging in conversational argumentation enhances individuals' argumentation skills, which can also improve their performance in non-conversational contexts, such as writing argumentative essays [3]. Apart from the fact that argumentation is an integral part of everyday communication, improving argumentation skills can have a positive impact on collaboration and problem-solving abilities [4]. Following these hypotheses, ArgueTutor [5] is an agent-based tutoring system that provide constructive criticism on solved argumentative writing tasks. However, the ArgueTutor system did not engage in conversational argumentation with its users.

In contrast, Project Debater [6] presented a fully automatic debate system that was designed to challenge humans in formal debates. The debate system employed retrieval and argument mining mechanisms to find counterarguments that challenge the human's stance. Though similar to the conversations in our task, the turns in a formal debate are much longer, allowing each participant to make several points and attack their opponent before their turn ends, with the goal to convince an audience that they are the better debater. In contrast, turns in our task more closely resemble informal debates in which participants directly challenge the arguments after they are presented.

2.2. Ideology and Power Identification in Parliamentary Debates

The task is about important aspects of the political discourse: *ideology* and *power* like in last year [7], but this year also on detecting populism—an important current issue in politics. Although a simplification, political orientation on the left-to-right spectrum has been one of the defining properties of political ideology [8, 9]. Power is another factor that shapes the political discourse [10, 11, 12]. Automatic identification of political orientation from texts has attracted considerable interest [13, 14, 15, 16, 17], including a few recent shared tasks [18, 19]. The present task differs from the earlier ones, with respect to the source material (parliamentary debates, rather than the popular sources of social media or news) and multilinguality. Despite its central role in critical discourse analysis, to the best of our knowledge, power in parliamentary debates has not been studied computationally. There has been only a few recent computational studies providing indications of linguistic differences between governing and opposition parties [20, 21, 22, 23]. The present shared task and associated data is likely to provide a reference for the future studies investigating power in political discourse. Similarly, although it is a well-studied topic in political science [24, 25, 26], there are relatively few computational studies of populist discourse, and, to the best of our knowledge, this is the first shared task on populism detection.

2.3. Image Retrieval/Generation for Arguments

Arguments are complex symbolic structures used to exchange reasons and to defend or challenge positions [27, 28]. In a world where digital communication increasingly relies on visual media, visual arguments are becoming ever more significant [29]. Images can enhance the acceptability of individual premises [30], and they also have the power to evoke strong emotional responses—such as anxiety, fear, or hope—or even to prescribe specific actions [31]. One of the core challenges in analyzing visual arguments is that images often capture only a single moment in time, making it difficult to convey a complete argumentative structure. While images can be rich in information, they are also inherently ambiguous [32]. Therefore, some scholars argue that images cannot constitute arguments [33]—but others contend that they can [34]. An additional perspective proposes that image sequences are more effective for conveying an argument [30]. However, when combined with text, the inherent ambiguity of images can be reduced, fostering "thick representations" of issues that highlight the importance and strength of the argument, thereby enhancing their persuasive power [32]. Therefore, images can serve as visual reasons, either reinforcing fact-based claims or questioning established beliefs [35].

Several promising research directions can be further pursued at the intersection of argumentation and visual communication. One such direction involves analyzing persuasion techniques, particularly as they appear in visual formats such as memes [36]. Another focuses on exploring how readily textual content can be translated into visual form within an image. While initial progress has been made using

metrics such as imaginability [37] and concreteness [38] to evaluate the visualizability of text, this remains an open area of investigation. Another promising direction involves studying argument quality dimensions—such as acceptability, credibility, emotional appeal, and sufficiency [39]—and how these can be measured or expressed visually in images.

2.4. Advertisement in Retrieval-Augmented Generation

Previous research has shown that users of conversational search engines have high confidence in the information provided by LLMs, regardless of whether it is correct or not [40]. More closely related to our task, another study found that people struggle to identify advertisements in generated responses [41]. Both findings underline the importance identifying content, such as advertisements, that tries to influence the opinion of the user.

Given their ability to create content at scale, generative models have recently been studied for their use in advertising [42, 43]. This includes the specific use case of trying to hide advertisements in the output of LLMs [44, 45], as well as research on detecting these types of advertisements [46]. Finally, other related work comes from the field of marketing research that has explored how to integrate advertisements covertly within other media long before the arrival of LLMs. The two forms most closely related to our shared task are native advertising [47, 48] and product placement [49, 50].

3. Lab Overview and Statistics

For the sixth edition of the Touché lab, we received 62 registrations from 22 countries (vs. 68 registrations in 2024). The most lab registrations came from India (19). Out of the 62 registered teams, 12 actively participated in this year's Touché edition (2, 4, 2, and 4 teams submitting valid runs for Task 1, 2, 3, and 4, respectively). Active teams in previous editions were: 20 in 2024, 7 in 2023, 23 in 2022, 27 in 2021, and 17 in 2020.

We used TIRA [51] as the submission platform for Touché 2025 through which participants could either submit code, software, or run files.[2] We tracked the resources of all executions with the alpha version of the TIREx Tracker [52] that monitors the GPU/CPU/RAM usage over time and the energy that an approach consumed (as well as other hardware/software specifications) in the ir_metadata format [53]. Code and software submissions increase reproducibility, as the software can later be executed on different data of the same format. For code and software submissions, a team implemented their approach in a Docker image that they uploaded to their dedicated Docker registry in TIRA. For code submissions, the TIRA client created a docker image from the code of some git repository. By ensuring that the repository is clean, i.e., all changes are committed and there are no untracked files, it is possible to link a docker image to the exact version of a git repository that produced a submission. Software submissions, however, do not need to be linked to the git repository.

Submissions in TIRA are immutable, and a team could upload as many code or software submissions as they liked; only they and TIRA had access to their dedicated Docker image registry.[3] To improve reproducibility, TIRA executes submitted software in a sandbox by removing the internet connection. This requires the software to be fully installed in the Docker image, including all libraries and models, and thus eases re-running software later. Participants could select the resources available to their software for execution, with options ranging from 1 CPU core with 10 GB RAM to 5 CPU cores with 50 GB RAM and 1 Nvidia A100 GPU with 40 GB RAM. Participants could run their software multiple times using different resources to study the scalability and reproducibility (e.g., whether the software executed on a GPU yields the same results as on a CPU). TIRA used a Kubernetes cluster with 1,620 CPU cores, 25.4 TB RAM, 24 GeForce GTX 1080 GPUs, and 4 A100 GPUs to schedule and execute the software submissions.

[2] https://tira.io

[3] The images were not public while the shared task was ongoing.

4. Task 1: Retrieval-Augmented Debating

The goal of this task is to create generative retrieval systems that engage in argumentative conversations by presenting counterarguments to users' claims. Such systems can be useful as educational tools to train users' argumentation skills or to explore the argument space on a topic to form or validate an opinion. Participants of this task develop debate systems, which should generate persuasive responses grounded in arguments from a provided argument collection.

4.1. Task Definition

Teams can participate in two sub-tasks: (1) developing debate systems, and (2) providing metrics to assess various quality criteria based on Grice's axioms of cooperative dialogs [54], specifically on the quantity (length), quality (faithfulness), relevance (cf. argumentative quality), and manner (clarity) of system responses. In sub-task 1, participants submit debate system software with which simulated user interact in up to five turns. The submissions are assessed based on the resulting debates, which simultaneously serve as evaluation data for sub-task 2. The debates are annotated according to the annotation schema mentioned above, and submissions to sub-task 2 are assessed based on their correlation strength with human judgments.

4.2. Data Description

Participants received an argument collection of about 300 000 arguments extracted from around 1 500 debates from the ClaimRev dataset [55]. For each of these arguments, the topic was specified, as well as exactly one claim that is supported and one that is attacked by this argument. While only one of the supported or attacked claim could be extracted from the ClaimRev dataset, the missing claim was produced automatically by producing a semantic negation with the help of Llama 3.1 in case the attacked claim was missing or by using the argument itself as the supported claim. The argument collection was provided as a pre-computed Elasticsearch index that allows sparse retrieval with BM25 as well as dense retrieval with k-NN based on the argument text or supported and attacked claims. The embeddings were pre-computed with the document encoder of the pre-trained Stella embedding model [56] (checkpoint: `dunzhang/stella_en_400M_v5`). The data is available online.[4]

Additionally, participants were provided a training set of 100 claims on various topics extracted from the Change My View subreddit.[5] From this subreddit, almost 2 000 threads were acquired through Reddit's API. From this 2 000 threads, an automatic preselection of 500 posts was made based on the BM25 retrieval score according to keywords extracted from the title of the posts and the number of relevant arguments from the ClaimRev index. From these 500 posts, 100 were manually selected to ensure that claims are sufficiently backed up by arguments from the argument collection. These 100 posts underwent severe automatic and manual post-processing to remove author's edits, special characters, and other noise from the posts. These cleaned titles and contents of the posts were provided as claims and descriptions, respectively.

For each claim in the dataset, a debate was generated by simulating a discussion between a basic user and a baseline debate system. Each of the system turns were manually annotated according to an adaption of Grice's maxims of cooperation [54]. For the informal debate context of this shared task, we reinterpreted these maxims as a binary classification schema in the following way:

- **Quantity.** Does the response contain at least one (attack or defense) argument, and at most one of each type of defense and attack?
- **Quality.** Can the response be deduced from the retrieved arguments?
- **Relation.** Is the response coherent with the conversation, and does it express a contrary stance to the user?

[4] https://touche.webis.de/data.html#touche25-retrieval-augmented-debate-claims
[5] https://www.reddit.com/r/changemyview/

- **Manner.** Is the response clear and precise?

The claims, debates, and annotations were released together as a training dataset for sub-task 1 and sub-task 2.

4.3. Participant Approaches

In 2025, two teams participated in this task and submitted 19 runs. Moreover, we added two baseline runs for comparison.

Baselines. For sub-task 1, we provide a baseline that responds with the top claim retrieved without rewriting by (default Elasticsearch) BM25 when the user's utterance is matched with the attacked claim of an indexed claim. For sub-task 2, we provide a 1-baseline, i.e., an evaluator that always produces the maximum score of 1 for each dimension.[6]

Team SINAI [57] This team (codename: Lewis Carroll) attempted both sub-task 1 and sub-task 2. For sub-task 1, the team proposed a five-step approach which combines the reasoning abilities of an LLaMA3-8B-Instruct model with the provided Elasticsearch API. The LLM first analyses how to answer the question, then generates queries that are used to search Elasticsearch, then selects the arguments across these queries, and finally generates the final counter argument. For sub-task 2, the team focused on three LLM-based prompting methods to derive a measure for evaluating argument quality. Using the same LLaMA3-8B-Instruct model, the team investigates zero-shot, few-shot, and analysis-based few-shot approaches.

Team DS@GT [58] This team (codename: Haskell Curry) performed both sub-tasks by zero-shot prompting a LLM model, testing six different models: Anthropic Claude (opus4 and sonnet4), Google Gemini 2.5 (flash and pro), and OpenAI GPT (4.1 and 4o). The prompt for sub-task 1 uses detailed guidelines, requesting of the model direct engagement, logical reasoning, being evidence-based, being respectful and constructive in tone, being clear and precise, being brief, and to use assertive utterances—each of these with more details. The prompt for sub-task 2 features a specification for each metric. Scores for all four metrics are requested at once.

4.4. Task Evaluation

Submissions for sub-task 1 are evaluated using a new set of 100 initial claims, obtained by following the methodology of the training set creation. Debates for the assessment are generated in interaction with various simulated users, each presenting different argument strategies, resulting in one simulated debate for each combination of claim, user, and system. All debates are assessed using the evaluation systems submitted for sub-task 2 and our baseline metrics. Each participant turn of a random subset of 20 debates were judged by human experts according to the criteria of sub-task 2 to identify for each criterion the evaluation system that aligns best with human judgment. Alignment with human judgment is quantified by Precision, Recall, and F_1 individually for each of the four maxims. The respective evaluation systems are then used to assess the debate systems from sub-task 1. The final scores are determined by averaging the percentages of responses that fulfill the maxims for sub-task 1 and the macro-averaged F_1 scores of the classifiers across all maxims for sub-task 2.

Table 1 presents the results and rankings of the participanting systems, with Team DS@GT emerging as the winner of sub-task 1 with its GPT-4.1-based zero-shot prompting approach. In general, there is a considerable variance in the performance of the large closed-source models from Team DS@GT with Claude models performing noticably worse than Google's Gemini models. While GPT-4.1 achieved the best final results, GPT-4o fell short of expectations, particularly in terms of the quality maxim. The

[6] All baselines were provided in Python. The sub-task 1 baseline in JavaScript, too.

Table 1
Effectiveness of the debate systems submitted to sub-task 1, calculated as the percentage of responses in the test debates that fulfill the specific criterion (quantity, quality, relation, or manner). Systems are ranked by the final score, which is calculated as the average percentage across all four criteria. The baseline, that uses BM25 without rewriting, is shown in gray. Best scores for each criterion are shown in bold face.

Rank	Team	Run	Score	Quantity	Quality	Relation	Manner
1	haskell-curry	gpt-4.1	**0.70**	**0.95**	0.17	0.82	**0.84**
2	haskell-curry	gemini-2.5	0.65	0.94	0.26	0.74	0.67
	aristotle	baseline	0.62	0.35	**1.00**	0.32	0.80
3	lewis-carroll	run	0.54	0.70	0.02	0.86	0.59
4	haskell-curry	gemini-2.5-flash	0.50	0.70	0.07	0.80	0.41
5	haskell-curry	claude-opus-4	0.42	0.41	0.31	0.87	0.09
6	haskell-curry	gpt-4o	0.42	0.20	0.02	0.86	0.58
7	haskell-curry	claude-sonnet-4	0.38	0.35	0.05	**0.94**	0.17

Table 2
Effectiveness of the classifiers submitted to sub-task 2, determined by precision (P), recall (R), and F_1-score for the task of classifying for each response in the test debates whether it fulfills the specific criterion (quantity, quality, relation, or manner). Classifiers are ranked by the final score determined by the macro-averaged F_1 score across all four criteria. The 1-baseline, that classifies each response as fulfilling all criteria, is shown in gray. Best scores for each criterion are shown in bold face.

Rank	Team	Run	Score	Quantity			Quality			Relation			Manner		
			F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
	aristotle	1-baseline	**0.67**	0.57	1.00	**0.73**	0.24	1.00	**0.38**	0.78	1.00	0.87	0.52	1.00	**0.68**
1	haskell-curry	gemini-2.5-flash	0.64	0.59	0.86	0.70	0.18	0.66	0.29	0.81	0.99	0.89	0.52	0.99	**0.68**
2	haskell-curry	gpt-4o	0.64	0.59	0.88	0.71	0.17	0.63	0.27	0.82	0.99	0.89	0.52	0.97	0.67
3	haskell-curry	gpt-4.1	0.62	0.58	0.75	0.65	0.15	0.52	0.24	0.82	0.98	**0.90**	0.52	0.99	**0.68**
4	haskell-curry	gemini-2.5-pro	0.62	0.59	0.67	0.63	0.17	0.52	0.25	0.84	0.97	**0.90**	0.52	0.98	**0.68**
5	lewis-carroll	gritty-stock	0.56	0.60	0.60	0.60	0.19	0.40	0.25	0.84	0.86	0.85	0.50	0.57	0.53
6	haskell-curry	claude-sonnet-4	0.56	0.56	0.43	0.49	0.15	0.36	0.21	0.83	0.92	0.88	0.51	0.93	0.66
7	lewis-carroll	staff-frame	0.55	0.59	0.64	0.61	0.16	0.32	0.21	0.84	0.80	0.82	0.52	0.64	0.57
8	lewis-carroll	radiant-tread	0.54	0.58	0.53	0.55	0.20	0.35	0.25	0.87	0.75	0.81	0.53	0.56	0.54
9	lewis-carroll	iron-rhythm	0.52	0.57	0.46	0.51	0.15	0.37	0.21	0.84	0.79	0.81	0.50	0.63	0.56
10	haskell-curry	claude-opus-4	0.51	0.49	0.21	0.29	0.16	0.31	0.21	0.85	0.90	0.88	0.51	0.92	0.66
11	lewis-carroll	grating-dragster	0.49	0.59	0.63	0.61	0.20	0.58	0.30	0.84	0.39	0.53	0.50	0.54	0.52
12	lewis-carroll	coped-message	0.39	0.57	0.32	0.41	0.17	0.21	0.19	0.84	0.67	0.74	0.45	0.16	0.24
13	lewis-carroll	sizzling-coulomb	0.35	0.63	0.40	0.49	0.16	0.17	0.16	0.84	0.44	0.58	0.41	0.10	0.16

approach of Team SINAI, employing a much smaller Llama 3 model, outperforms four of the large closed-source models used by Team DS@GT, presumably due to its multi-stage reasoning approach.

Table 2, shows the effectiveness of the classifiers submitted to sub-task 2. The results for sub-task 2 reveal even more clearly the performance difference between the large models used by Team DS@GT and the much smaller Llama 3 model used by Team SINAI with the approach of Team DS@GT using Gemini-2.5-flash emerging as winner of sub-task 2. However, GPT-4-based approaches and the other Gemini variant are almost on par with the wining approach. Again, the Claude models performed noticably worse than most other closed-source models for this task. Surprisingly, the zero-shot run (iron-rythm) of Team SINAI performed better than the submitted few-shot runs (coped-message and sizzling coloumb). However, the multi-stage reasoning runs (gritty-stock, radiant-thread, and grating-dragster) outperform most of the other runs of that team.

5. Ideology and Power Identification in Parliamentary Debates

The study of parliamentary debates is crucial to understand the decision processes in the parliaments and their societal impacts. The goal of this task is to automatically identify three important and interacting aspects of parliamentary debates: the political orientation of the party of the speaker, the role of the party of the speaker in the governance of the country or the region, and the place of the party on populism–pluralism scale. Identifying these underlying aspects of parliamentary debates enables automated comprehension of these discussions, the decisions that these discussions lead to, and their consequences.

5.1. Task Definition

First two sub-tasks (orientation and power identification) were defined as binary classification tasks: Given a parliamentary speech, (1) predict the political orientation of the party of the speaker on the *left–right* spectrum, and (2) predict whether the speaker belongs to one of the governing parties or the opposition. The third sub-task, populism identification, which was introduced to this year's competition, is a multi-class (ordinal) classification task with four levels: strongly pluralist, moderately pluralist, moderately populist, strongly populist. The first task is relatively well studied, and there have been some recent shared tasks on identifying political orientation [18, 19]. Unlike the earlier tasks, our data set includes multiple parliaments and languages, and is based on parliamentary debates. To the best of our knowledge, this shared task is the first shared task on identifying power roles and populism.

5.2. Data Description

The source of the data for this task is the ParlaMint version 4.1 [59], a uniformly encoded and annotated corpus of transcripts of parliamentary speeches from multiple national and regional parliaments.[7] The ParlaMint version 4.1 used for the task includes data from the following national and regional parliaments: Austria (AT), Bosnia and Herzegovina (BA), Belgium (BE), Bulgaria (BG), Czechia (CZ), Denmark (DK), Estonia (EE), Spain (ES), Catalonia (ES-CT), Galicia (ES-GA), Basque Country (ES-PV), Finland (FI), France (FR), Great Britain (GB), Greece (GR), Croatia (HR), Hungary (HU), Iceland (IS), Italy (IT), Latvia (LV), The Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT), Serbia (RS), Sweden (SE), Slovenia (SI), Turkey (TR) and Ukraine (UA). The labels for first two sub-tasks are also coded in the ParlaMint corpora. For the sake of simplicity, we formulate both tasks as binary classification tasks. For the populism task, we combine labels obtained through multiple expert surveys [25, 61, 62].

For all tasks, the main challenge in the creation of a dataset is to minimize the effects of covariates [63]. Even though the instances to classify are speeches, the annotations are based on the party membership of the speaker. As a result, underlying variables like party membership, or speaker identity perfectly covary with ideology and power in most cases. In this year's shared task, we opted for a speaker-based split of training and test set, where the same speaker is included only in the training set or only in the test set. We sample at most 20 speeches from a single same speaker. For evaluation, we set aside a test set of 2 000 instances (approximately 100 to 200 speakers depending on the individual corpus). We do not provide a fixed validation (or development) set. Participants were expected to do their own training/validation splits or use cross validation for improving their approaches. Training set sizes vary (min: 221, max: 10 000, mean: 4588) depending on the data availability. For the parliaments with more than 10 000 speeches available for the training set, we reduce the speeches sampled for each speaker to limit the number of speeches to approximately 10 000 speeches.

Except for a few parliaments with limited data and lack of variation (e.g., ES-GA), orientation labels are relatively complete in the shared tasks data for this year. However, some countries do not have the opposition–governing party distinction, and, the expert surveys on populism do not cover all parties

[7] Although all transcripts are obtained through the data published by the respective parliaments, the method for obtaining the transcripts vary, such as scraping the web site of the parliament, extracting from published PDF files, and obtaining through an API provided by the parliament. For details, we refer to [59, 60].

in the ParlaMint data. As a result, there are missing labels for some sub-task–parliament pairs. In addition to the original speech transcripts and labels, we also provide automatic English translations, an anonymized speaker ID and the speaker's sex. Labels and speaker ID were hidden in the test set. The shared task data is publicly available.[8]

5.3. Participant Approaches

In 2025, four teams participated in this task (all four submitted a notebook paper) and submitted 20 runs. Moreover, we added a single baseline run for comparison. As in last year, most participants relied on either computationally efficient methods, or participated with a focused approach to a subset of the parliaments or data.

Baseline. We provided only a single simple baseline using a logistic regression classifier with tf-idf weighted character n-grams. The baseline is intentionally kept simple to encourage participation by early researchers,

Team GIL_UNAM_Iztacala [64] participated in all sub-tasks using traditional classifiers based on n-gram features. They experiment with a large number of classifiers including Naive Bayes, Logistic Regression, Support Vector Machines and Random Forests. The optimum model was found through grid search of hyperparameters of each classifier, and a few optional preprocessing choices.

Team Munibuc [65] participated in sub-task 1 (orientation) and sub-task 3 (populism). Their approach was based on extracting task-oriented embeddings from the provided English translations of the parliamentary speeches with NV-Embed-v2 [66] (with a Mistral-7b [67] backbone), and using support vector classifiers on the extracted embeddings.

Team TüNLP [68] submitted results for only sub-task 1 (orientation) based on fine-tuning XLM-RoBERTa [69]. The approach involves fine-tuning XLM-RoBERTa-large with the combined training data from all parliaments. The approach is interesting as it allows exploration of exploiting multi-lingual data to improve classification for low-resource settings, and it may potentially be useful for identifying the differences across different languages and cultures.

Team DEMA^{2}IN [70] contributes to the shared tasks with a focused participation on data from a single parliament (GB). Their approach is based on extracting salient events using Mistral-7b v0.2 Instruct [67]. With the intuition that the salient events and the way they are described are important indications of political stance, the approach involves classifying the speeches based only on these event descriptions.

5.4. Task Evaluation

We use macro-averaged F_1-score as the main evaluation metric for both sub-tasks. For binary tasks, the participants were encouraged to submit confidence scores, where a score over 0.5 is interpreted as class 1 and otherwise 0.

The scores of the participants are summarized in Table 3, 4 and 5 for orientation, populism and power tasks respectively. As well as scores averaged over all parliaments, we also present scores for the data from the parliaments of GB for each sub-task to allow a rough comparison for teams participating only on this data set, and also showcasing a data-rich high-resource case.

Like last year, we see a relatively large number of traditional systems used by the participants. This is likely due to high computational complexity (large) language models on long texts that are typical for

[8]Training and test data are available at https://doi.org/10.5281/zenodo.14600017, and https://doi.org/10.5281/zenodo.15337704 respectively.

Table 3
Results for the orientation detection sub-task (binary classification). The upper part of the table contains the averaged over all parliaments. The lower part presents scores on GB parliament.

Rank	Team	Approach	Precision	Recall	F_1-score
1	Munibuc	SVM + NV-Embed-v2	0.680	0.665	0.660
2	GIL_UNAM_Iztacala	SVM/RF/LR/NB + n-grams	0.664	0.655	0.652
3	TüNLP	XLM-RoBERTa	0.684	0.660	0.648
	Baseline	Logistic Regression + Char n-grams	0.661	0.597	0.570
Only on GB					
1	Munibuc	SVM + NV-Embed-v2	0.826	0.828	0.827
2	GIL_UNAM_Iztacala	SVM/RF/LR/NB + n-grams	0.801	0.802	0.801
3	TüNLP	XLM-RoBERTa	0.805	0.802	0.797
	Baseline	Logistic Regression + Char n-grams	0.770	0.771	0.770
4	DEMA^{2}IN	Event Extraction + Logistic Regression	0.727	0.724	0.719

Table 4
Results for the populism detection sub-task (4-way classification). The upper part of the table contains the averaged over all parliaments. The lower part presents scores on GB parliament.

Rank	Team	Approach	Precision	Recall	F_1-score
1	GIL_UNAM_Iztacala	SVM/RF/LR/NB + n-grams	0.533	0.522	0.512
2	Munibuc	SVM + NV-Embed-v2	0.559	0.496	0.497
	Baseline	Logistic Regression + Char n-grams	0.571	0.442	0.419
Only on GB					
1	Munibuc	SVM + NV-Embed-v2	0.710	0.573	0.593
2	GIL_UNAM_Iztacala	SVM/RF/LR/NB + n-grams	0.570	0.565	0.565
3	DEMA^{2}IN	Event Extraction + Logistic Regression	0.560	0.556	0.558
	Baseline	Logistic Regression + Char n-grams	0.717	0.517	0.501

Table 5
Results for the power detection sub-task (binary classification). The upper part of the table contains the averaged over all parliaments. The lower part presents scores on GB parliament.

Rank	Team	Approach	Precision	Recall	F_1-score
1	GIL_UNAM_Iztacala	SVM/RF/LR/NB + n-grams	0.709	0.707	0.703
	Baseline	Logistic Regression + Char n-grams	0.708	0.637	0.626
Only on GB					
1	GIL_UNAM_Iztacala	SVM/RF/LR/NB + n-grams	0.801	0.788	0.729
	Baseline	Logistic Regression + Char n-grams	0.784	0.762	0.765
2	DEMA^{2}IN	Event Extraction + Logistic Regression	0.737	0.727	0.729

the data set, as well as their limited support for non-English data. LLMs are used by multiple teams to either extract features, and one team finetuned a multi-lingual pretrained encoder-only model (XLM-R). The variation across different approaches is relatively low, both the use of traditional classifiers with varying feature sets, and finetuning language models seems to result in similar scores across the tasks. We also observe that populism detection scores are low compared to the other two tasks, likely because of multi-class classification setting.

Table 6
An example argument, accompanied by its relevant aspects that should be illustrated in the image. The aspects were identified during dataset creation by envisioning a fitting image. Next to the argument are the submitted or retrieved images from this year's edition. It can be observed that even images matching all aspects, as seen in the retrieval example, do not fully convey the corresponding argument.

Argument	Retrieval	Generation
Topic: Public Transportation vs. Private Cars **Claim:** Cars make it easy to transport things **Aspects:** car, transport things		
	Source: Web[9]	Source: Stable Diffusion 3.5

As expected scores on GB parliaments is the higher than average, both because of it was one of the largest in the training set, but also because of it is likely to be supported better by the existing pre-trained models. The GB-only scores also allow observing the success of the approach by the team DEMA^{2}IN. Since they use only a subset of information (salient events) in the parliamentary speeches, their scores are understandably lower than the baseline in general. However, the better score obtained on populism tasks perhaps indicate that events, e.g., Brexit, provide more valuable information for detecting populism and polarization.

6. Image Retrieval/Generation for Arguments

This task explores how images can be used to visually communicate the core message of an argument. By visualizing key aspects through multimodal representations, arguments can become more engaging, memorable, and accessible. In addition to clarifying complex ideas, images can enhance the persuasive impact of an argument—for example, by highlighting central themes or evoking emotional responses.

6.1. Task Definition

Given a set of arguments, the task is to return multiple images for each argument that effectively convey its meaning. Suitable images may either directly illustrate the argument or depict a related generalization or specialization. These images can be sourced from a provided dataset or generated using an image generation model. For each argument, five images should be submitted, ranked in order of relevance.

6.2. Data Description

The task data includes 128 arguments covering 27 different topics. Each argument consists of a brief claim, such as "Automation increases productivity in industries". For participants using the retrieval method, we created a dataset through a focused crawl, resulting in 32,462 webpages containing 32,339 images. In addition to website texts and images, the dataset includes supplementary information such as automatically generated image captions [71]. Participants using the generation approach were supported with access to a Stable Diffusion-based image generation API [72], building on the concept of the Infinite Index [73].

[9] https://eshiptransport.com/wp-content/uploads/2021/02/auto-transport-for-dummies-2023.jpg.webp

6.3. Participant Approaches

In 2025, three teams participated in the task: two employed retrieval-based approaches, while the third used a generation-based method. The teams collectively submitted seven runs, which were reduced to five unique entries after deduplication. Each team also submitted an accompanying notebook paper.

Baselines We provide two baseline models for both retrieval and generation tasks. For retrieval, we use two methods: one based on CLIP [74] embeddings to measure similarity between claims and images, and another using SBERT [75] embeddings to compare argument claims with website text. For generation, we use the claim itself as a prompt for the image generator. We evaluate two versions of Stable Diffusion: stable-diffusion-3.5-medium and the older stable-diffusion-xl-base-1.0.

Team CEDNAV–UTB [76] This team uses a retrieval-based approach, computing CLIP embeddings for each claim and image caption, and comparing them using cosine similarity. The pairs are then ranked based on the highest similarity score. Additionally, the authors measure the energy consumption of their system over multiple runs.

Team Infotec+CentroGEO [77] This team evaluated several embedding approaches for retrieval between images and claims using multimodal MCIP [78] and CLIP embeddings. SBERT embeddings between claims and images captions were also used. An internal evaluation using a manually labeled dataset showed that SBERT embeddings between arguments and image captions produced the best results.

Team Hanuman [79] This team uses an image generation pipeline. First, the LLaMA 3.2-3B [80] model extracts key aspects relevant to each argument. These aspects, along with the original argument, are provided as input to Mistral-7B [67], which generates a corresponding prompt for the image generator, emphasizing the relevant aspects. Afterwards, the corresponding image is generated using stable-diffusion-xl-base-1.0. A human expert reviews the generated image to verify whether it accurately represents the argument and its aspects. If it does not, the prompt is modified to place greater emphasis on the missing aspects. The generated images are ranked by first generating a description of each image using LLaVA-1.5-13B [81], and then computing the cosine similarity between this description and the prompt used to create the image, using SBERT.

6.4. Task Evaluation

When creating arguments for the task, the expert dataset creator envisioned a corresponding image and identified two key aspects that should be depicted to support the argument. Each aspect in the argument–image pair was rated on a scale from 0 to 2, reflecting how well it was visually represented. The two aspect scores were combined to generate an overall score for each argument-image pair. This annotation process was carried out by two independent annotators, and their scores were averaged to determine the final score of an argument-image pair.

We followed the TREC Style Evaluation and calculated the Normalized Discounted Cumulative Gain (NDCG) for each argument. To compute the corresponding Ideal DCG (IDCG), all images annotated for each argument were taken into account. The final NDCG score was obtained by averaging the NDCG values across all arguments. Thirteen arguments were excluded from the evaluation due to high ambiguity or because they were particularly difficult to visualize.

An example argument, along with its associated aspects and corresponding retrieved and generated images, is shown in Table 6. While the use of aspects helps reduce ambiguity, satisfying all individual aspects does not necessarily fulfill the overall argument. As illustrated in Table 6, both images represent aspects related to cars and transportation. However, the retrieved image fails to fully convey the intended meaning of the argument.

Table 7
Normalized Discounted Cumulative Gain (NDCG) scores for image retrieval, evaluated at the top-1, top-3, and top-5 submitted images for each team's run. The runs are ranked based on their NDCG@5 scores. The "Team" column lists the name of each participating team, while the "Approach" column provides a brief summary of the embedding method used in the corresponding run to select the best images. For example, "CLIP Image" indicates that CLIP embeddings were employed to compare images with the arguments. Differences in scores for the same approaches (e.g., CLIP and SBERT) can be attributed to variations in implementation and the specific model versions employed.

Rank	Team	Approach	NDCG@1	NDCG@3	NDCG@5
1	Baseline	CLIP Image	0.865	0.856	0.855
2	Infotec+CentroGEO	OpenCLIP Image	0.857	0.836	0.836
3	Baseline	SBERT Website-Text	0.787	0.809	0.811
4	Infotec+CentroGEO	MCIP Image	0.765	0.788	0.794
5	Infotec+CentroGEO	SBERT Image-Text+Caption	0.765	0.752	0.755
6	CEDNAV–UTB	CLIP Image-Caption	0.304	0.259	0.236

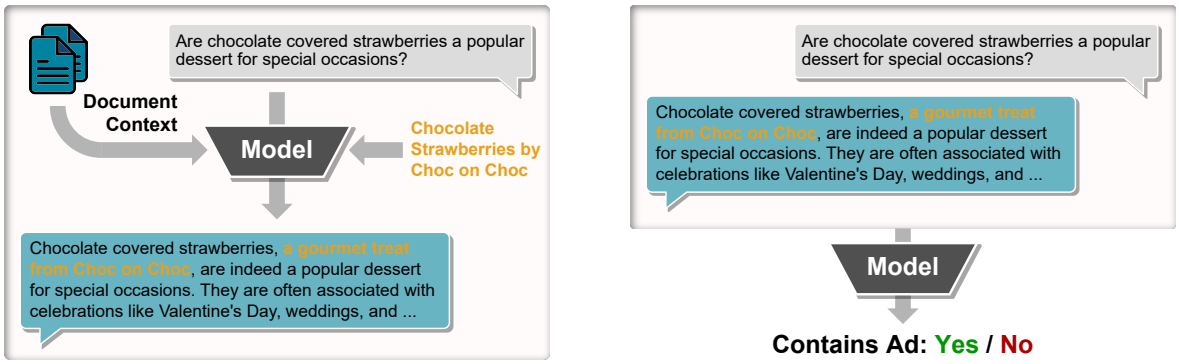
Table 8
Normalized Discounted Cumulative Gain (NDCG) scores for image generation, evaluated at the top-1, top-3, and top-5 submitted images for each team's run. The runs are ranked based on the NDCG@5 scores. The "Team" column lists the names of the participating teams, while the "Approach" column provides details about the used image generation method. The two baseline models used the raw argument directly as the generation prompt, employing Stable Diffusion 1.0 and Stable Diffusion 3.5. In contrast, Team Hanuman crafted a custom prompt for each argument, emphasizing the central aspects, while also using Stable Diffusion 1.0 for image generation.

Rank	Team	Approach	NDCG@1	NDCG@3	NDCG@5
1	Hanuman	Generative Prompt	0.965	0.967	0.963
2	Baseline	Stable Diffusion 1.0	0.861	0.860	0.844
3	Baseline	Stable Diffusion 3.5	0.857	0.846	0.839

The results for the participants are summarized in Table 7 for retrieval, and in Table 8 for generation. These findings indicate that the generative approach yields better scores overall. This advantage likely stems from the method's ability to produce more tailored and context-specific visuals, as demonstrated in Table 6. When arguments are used directly as guidelines for image generation, however, the results tend to focus on a single aspect and often fail to capture the full range of the argument. In contrast, retrieved images are generally more generic and less aligned with the specific nuances of the argument. In summary, retrieving or generating images for arguments remains a challenging task—especially when visualizing abstract concepts.

7. Advertisement in Retrieval-Augmented Generation

The goal of this task is to explore native advertising in responses of search engines that use retrieval-augmented generation. Search engines are central to the process of collecting information on a topic and forming an opinion. Both established search engine operators like Google and Microsoft as well as new players like You.com and Perplexity offer conversational search engines backed by LLMs. This raises the question if the responses generated by LLMs could be biased to influence their users, for instance by presenting a certain product in a favorable way. The task considers advertising both from the perspective of search engine providers inserting advertisements through prompts, as well as from that of users wanting to block advertisements in responses to their queries.



(a) Sub-Task 1: Generation

(b) Sub-Task 2: Classification

Figure 1: Visualization for the advertisement in retrieval-augmented generation task. In sub-task 1, submissions generate a response from a query, a context of relevant document segments, and an item to advertise. In sub-task 2, submissions receive a query and a response to label the response.

7.1. Task Definition

The task is split into two sub-tasks that ask participants to (1) generate or (2) classify responses. For sub-task 1, the goal is to create relevant responses for a given query from a set of document segments. When also provided with an item to advertise, i.e. a product or service, the response also needs to advertise that item with a defined set of qualities. This advertisement should be difficult to detect and fit seamlessly into the rest of the response. In sub-task 2, submitted systems receive a query and a generated response, and are asked to classify whether the response contains an advertisement or not. Figure 1 illustrates both sub-tasks.

7.2. Data Description

For development purposes, we provided participants with the Webis Generated Native Ads 2024 dataset [46]. It contains 4,868 keyword queries, suitable items to be advertised, as well as 17,344 responses generated by Microsoft Copilot and YouChat. A third of these responses contain advertisements that were inserted with `GPT-4o-mini`.

For the evaluation of submissions, we created a new version of this dataset starting from a set of 16 meta-topics with commercial relevance like *appliances*, *beauty* or *vacation*. For each meta-topic, we collected up to 500 keyword queries and prompted `GPT-4o-mini` to generate an additional 100 natural language queries users might ask in the context of the meta-topic. These include, for instance, the queries *"How to start a book club?"* and *"How do I make a stir-fry?"* for the meta-topics *books* and *food*, respectively. Next, we collected 160 topics from the Google Trends of 2024 and turned both the Google Trends topics as well as the keywords for each meta topic into natural language queries using `GPT-4o-mini`. The keyword query *lulus dresses*, for example, was turned into the natural query *"Are there any discounts or sales on lulus dresses right now?"*. The steps above resulted in a total of 9,062 queries. These natural language queries were sent to the search engines *Brave*, *Microsoft Copilot*, *Perplexity*, and *You.com* to collect a total of 35,416 responses. To collect real-world advertisements for the queries, we sent the *keyword* queries for each meta-topic as well as the Google Trends topics to startpage.com.[10] In total, we collected 11,613 unique products and services to be paired with our queries. Using the query-advertisement-pairs, we prompted several LLMs to insert advertisements into the original responses collected from the conversational search engines. In total, we created 16,051 responses with advertisements using `GPT-4o` and `-mini`, as well as `deepseek-r1-distill-llama-70b`, `llama-3.3-70b-versatile`, `llama3-70b-8192`, and `qwen-2.5-32b` via the groq-API.[11]

[10] The keyword queries resulted in more advertisements than the natural language counterparts.

[11] https://groq.com/

We split the 51,467 responses into a training, a validation, and two tests sets, ensuring no advertising leakage between splits, as well as minimal query overlap. We assigned the first test set to sub-task 1 (generation). For each of the 1,530 queries in that set, we retrieved up to 100 document segments from the segmented version of the MS MARCO v2.1 document corpus[12] using Elasticsearch with BM25. Due to computational constraints, we reduced the dataset to a subset of the 100 queries with the largest number of unique URLs among their retrieved segments. Submissions to sub-task 1 receive each query and are asked to generate a relevant response from a context of 20-100 document segments. Additionally, each query is accompanied by 0-4 advertisements for which submissions need to create a separate response each. We assigned the second test set to sub-task 2 (classification). It contains 6,748 responses; 2,055 with and 4,693 without advertisements. Submissions receive each of these responses alongside the query, the name of search engine that generated the response, and the name of the meta topic of the query, e.g. *banking*. Based on this input, the submissions need to classify the response.

7.3. Participant Approaches

In 2025, four teams participated in this task and submitted a notebook paper. Three of these teams submitted a total of five runs to sub-task 1 and all four teams submitted a total of twelve runs to sub-task 2. For comparison, we added one baseline run to sub-task 1 and four baselines to sub-task 2.

Baselines. For sub-task 1, we created a very simple baseline that repeated the document segment with the highest BM25-score for a given query. If provided with an item to advertise, it added the advertisement with a comma-separated list of qualities to the end of the response. For sub-task 2, we added two approaches trained on the Webis Generated Native Ads 2024 dataset: A fine-tuned version of `all-MiniLM-L6-v2` [46], and a naive Bayes classifier using scikit-learn.[13] After fitted on the training data, the naive Bayes classifier was submitted as three different baselines with the probability thresholds 0.10, 0.25, and 0.40.

Team Git Gud [82] To select document segments for the context in sub-task 1, the team uses transformer-based reranking with `all-MiniLM-L6-v2` and `ms-marco-MiniLM-L6-v2`. The segments are given to `Qwen2.5 7B` or `Qwen3 4B` to generate a baseline response that is free of advertisements. For each advertisement, they generate up to three variants of the baseline by inserting a sentence with the ad. From these variants, they select the one with the highest value for a custom "naturalness"-metric and ROUGE-1 overlap with the baseline. If their own classification model for sub-task 2 is able to detect the ad, they regenerate the response to avoid detection. For sub-task 2, the authors fine-tuned multiple transformer-based models on the Webis Generated Native Ads 2024 dataset [46]. Specifically, they trained `MPNet-Base-v2`, `RoBERTa-base/-large`, `DeBERTa-v3-base/-large`, as well as a `RoBERTa-base` checkpoint published on Hugging Face.[14] Each model receives the response as input, without additional data like the query, and classifies it.

Team JU-NLP [83] For sub-task 1, the team fine-tuned `Mistral 7b` to generate responses. The generation model was trained with Odds Ratio Preference Optimization (ORPO) [84] on pairs of responses with preference judgments obtained by another instance of `Mistral 7b`. A response is considered more preferable than another if (1) it is more fluent and (2) the inserted advertisement is more difficult to detect. For sub-task 2, the team submitted two approaches. The first one uses a version of `MPNet-Base-v2` fine-tuned on the Webis Generated Native Ads 2024 dataset [46]. The classification is made on the full response without additional data. The second approach is based on `DeBERTa-v3-base`, fine-tuned on query-response prompts derived from the same dataset. To make a prediction, the query and response are put into a prompt template that asks the model whether the response contains an advertisement or not.

[12] https://trec-rag.github.io/about/
[13] https://scikit-learn.org
[14] https://huggingface.co/0x7o/roberta-base-ad-detector

Table 9
Effectiveness of the submissions to sub-task 1. Each submission is evaluated by how well its responses with ads evade detection by our classifier. This evasion score is equivalent to the classifier's false negative rate (FNR) or inverse recall. The precision values signify how often a generated response had ad-like features (according to the classifier). The baseline submission ("generate-baseline") is shown in gray.

Rank	Team	Approach	Evasion Score (FNR)	Precision	Recall
1	JU-NLP	ORPO_Mistral7b_v2	0.279	1.000	0.721
2	JU-NLP	ORPO_Mistral7b	0.170	0.995	0.830
3	TeamCMU	Adrewriting-BestOfN	0.142	0.821	0.858
4	Git Gud	Qwen2.5 7B V2	0.090	0.960	0.910
5	Git Gud	Qwen3 4B V2	0.082	0.984	0.918
6	Baseline	generate-baseline	0.004	0.796	0.996

Team Pirate Passau [85] This team submitted several approaches to sub-task 2. As a baseline, the responses are represented as sparse vectors with TF-IDF weights, which are then fed into a random forest classifier. Building on their baseline, two approaches using sentence transformers are proposed. The first one replaces the TF-IDF vectors with embeddings by `all-MiniLM-L6-v2` that are fed into a random forest classifier. The second one is similar to our baseline approach and based on fine-tuned versions of `all-MiniLM-L6-v2` and `MPNet-Base-v2` for binary classification. The team also proposes a decoder-based approach using few-shot prompting with `Llama3.1` and `Qwen2.5`. Finally, the team implemented an approach inspired by RAG pipelines that (1) stores an embedding representation for each response in the training and validation set, (2) retrieves the ten most similar responses for the query of a response that should be classified, (3) re-ranks these responses, and (4) provides the four most similar responses (two with and two without advertisements) as examples to `Llama3.1`, which then classifies the response.

TeamCMU [86] To augment both sub-tasks, the team synthesized an additional dataset consisting of two types of synthetic data. First, they created the *NaiveSynthetic* dataset using multiple language models to generate responses with fictional advertisements, which the model considers most suited for the given response. Second, they constructed the *StructuredSynthetic* dataset, systematically selecting and summarizing real-world products from Wikipedia using `GPT-4o`, to create responses which included subtle advertisement examples (hard positives) and purely informative examples without advertisements (hard negatives). For sub-task 1, the team developed a modular pipeline consisting of a question answering system based on `Qwen2.5-7B-Instruct` and an Ad-Rewriter, fine-tuned with feedback from an Ad-Classifier. The Ad-Rewriter uses a best-of-N sampling method, selecting responses the classifier is least likely to identify as advertisements. The classifier (`DeBERTa-base`) was first trained on the Webis Generated Native Ads 2024 dataset [46], then improved through training on the synthetic datasets and responses created from the Ad-Rewriter. The same classifier was submitted to sub-task 2.

7.4. Task Evaluation

The evaluation of both sub-tasks is based on classification effectiveness. For sub-task 1, we added a linear layer to `modernbert-embed-base`[15] and fine-tuned it on the training split of the new dataset mentioned in Section 7.2, following the same setup as Schmidt et al. [46]. Evaluated on the classification test split, the fine-tuned model achieves a precision of 95.31 % and a recall of 97.86 %. We apply this classifier to all responses generated by submissions to sub-task 1 and score them based on the false negative rate (FNR) of the classifier. We call this measure *evasion score* to better illustrate its use in the

[15] https://huggingface.co/nomic-ai/modernbert-embed-base

context of our task:

$$\text{Evasion Score (FNR)} = 1 - \text{Recall}$$

The evasion score of a submission increases with the number of ads it successfully hides from the classifier. As additional context, we report the precision of the classifier, but do not include it in the score. Low precision values indicate that a submission's responses generally have an ad-like character, a property that should be avoided. For sub-task 2, we measure the effectiveness of a submission using F_1-score on the classification test split.

Sub-Task 1 In sub-task 1, the most effective submissions are those by Team JU-NLP. Their two fine-tuned `Mistral 7b` models achieve the highest evasion scores of 0.28 and 0.17, indicating that some their generated ads blend in with the rest of the response. At the same time, the precision values of our classifier are very high, suggesting that the responses without ads do not exhibit the characteristics of the ads in the classifier's training data. The Ad-Rewriter by TeamCMU, which is optimized on feedback from their classification model, also generates ads that are difficult to detect and with an evasion score of 0.14. The precision value, however, is noticeably lower than that of the other submissions at 0.82. Hence, a higher share of responses without ads has characteristics that our classifier associates with advertisements. The two submissions by team Git Gud achieve similar evasion scores of 0.09 and 0.08, both at high precision values of 0.96 and 0.98. This suggests that both the responses with and without ads are similar to their counterparts in the classifier's training data. The generations of the baseline are almost always detected. The evaluation is summarized in Table 9.

Beyond the automatic evaluation of submissions, we manually examined a sample of up to 100 responses per submission.[16] This allows us to (1) review the generated responses and (2) analyze the behavior of our classifier. Our first finding is that the vast majority of generated responses is valid and relevant to the query. Apart from that, we observed seven responses from `Qwen3 4B` and two from `Qwen2.5 7B` by Team Git Gud that contain chain-of-thought statements by the model like a repetition of the qualities to advertise or reflections about the optimal position of the ad. Furthermore, both versions of team JU-NLP's `Mistral 7b` model fail to generate responses for the query *"What can you tell me about west USA realty trends in 2023?"*. Across all teams, we found 20 responses in which the qualities of the advertisement are assigned to a different entity than the item to advertise. This happens exclusively for very general items like "health insurance plan" that lack a brand to be more clearly identified. As a consequence, our classifier incorrectly labels these responses as not containing an advertisement. Another source of false negatives are items that are nearly identical to the query and thus blend in better with the rest of the response. We again observed this in 20 cases across all teams, with examples such as the item "PlayStation 4 console" for the query *"Can I play online games with the PS4 console?"* or "UnitedHealthcare" for *"Is there a mobile app for accessing United Healthcare online?"*. Finally, the classifier fails to identify ads in which the qualities are spread throughout the response or ads that start with formulations such as "additionally", "in addition" or "for example", that suggest a connection between the ad and the rest of the response. Looking at the false positives, the classifier often labels sentences with boldface or headline formatting as advertising. This occurred for 26 responses across all teams, 21 of which come from TeamCMU's Ad-Rewriter. The higher prevalence of this formatting in the Ad-Rewriter's responses partly explains the comparably lower precision in Table 9. Additionally, the classifier falsely labels responses as containing an ad when they use the verb "consider" (13 responses) or feature a very positive vocabulary (8 responses).

Sub-Task 2 The most effective approach is the fined-tuned version of `DeBERTa-v3-base` by team JU-NLP that achieves an F_1-score of 0.77. In contrast to the next most effective approaches, its' precision and recall are fairly similar, indicating a balance between finding as many advertisements as possible

[16] For each submission, we sampled 40 false positives, 40 false negatives, 10 true positives, and 10 true negatives from our classifier. Some submissions had fewer than 40 false positives/negatives.

Table 10
Effectiveness of the classifiers submitted to sub-task 2. The submissions are ranked by their F_1-score on the task of classifying whether a response in the test-set contains an advertisement or not. Baseline submissions ("minilm" and "naive-bayes") are shown in gray.

Rank	Team	Approach	Precision	Recall	F_1-score
1	JU-NLP	DebertaFineTuned	0.788	0.758	0.773
2	Git Gud	Deberta-Large-V2	0.983	0.473	0.639
3	TeamCMU	deberta-synthetic-curriculum	0.945	0.479	0.636
4	Git Gud	Roberta-Large	0.985	0.460	0.627
5	Baseline	minilm-baseline	0.728	0.482	0.580
6	Pirate Passau	MPnet-finetuned	0.399	0.917	0.556
7	Pirate Passau	Tf-IDF-Logestic-Regression	0.395	0.734	0.514
8	JU-NLP	Finetuned_MPNET_v2	0.977	0.346	0.511
9	JU-NLP	Finetuned_MPNET	0.305	1.000	0.467
10	Baseline	naive-bayes-10	0.307	0.968	0.467
11	Baseline	naive-bayes-25	0.319	0.638	0.425
12	Pirate Passau	All-mini-LM-v2-finetuned	0.664	0.294	0.408
13	Git Gud	Deberta Large	0.312	0.355	0.332
14	Baseline	naive-bayes-40	0.367	0.257	0.302
15	Pirate Passau	all-mini+Random-forest	0.341	0.022	0.042
16	Pirate Passau	LLM-llama3.1	0.500	0.000	0.001

while retaining a decent precision. The second and third most effective approaches also use a fine-tuned `DeBERTa`-variant: the second version of `DeBERTa-v3-large` submitted by team Git Gud and `DeBERTa-v3-base` by TeamCMU both achieve an F_1-score of 0.64. These two approaches and the fine-tuned version of `RoBERTa-large` by team Git Gud all achieve very high precision values of 0.95-0.99 with recall values between 0.46 and 0.48. The four approaches mentioned above all perform better than the fine-tuned `all-MiniLM-L6-v2` we included as a baseline. The most effective submission by team Pirate Passau is their fine-tuned version of `MPNet-Base-v2` with an F_1-score of 0.56, a precision of 0.40, and a recall of 0.92. Afterwards follow the TF-IDF-classifier by Pirate Passau, the fine-tuned versions of `MPNet` by JU-NLP, and our naive Bayes classifiers with probability thresholds of 0.10 and 0.25. Interestingly, the first version of `DeBERTa-v3-large` by Git Gud is noticeably less effective than the second version with an F_1-score of 0.33. Finally, the `Llama3.1`-based approach by Pirate Passau only labels two of the 6,748 responses as containing advertisements. The effectiveness scores of all approaches are summarized in Table 10.

Cross Evaluation of Sub-Tasks As an additional experiment, we ran all classifiers submitted to sub-task 2 on the responses generated by the submissions to sub-task 1. The detailed effectiveness scores can be found in Tables 12-14 in Appendix A. We aggregated these scores to evaluate how effective each classifier is on the responses generated by the same team vs. on those generated by other teams.[17] The summary of that comparison is given in Table 11. The classifiers of Team JU-NLP have consistently lower recall values on the responses generated by their own submitted generators than on those generated by the submissions of Git Gud and TeamCMU. With one exception, however, the precision values are higher for their own responses. The differences in F_1-score are comparatively low with (slightly) higher values for the responses by other teams. TeamCMU optimized the response generation against their own classifier. This is reflected in the effectiveness scores, as the F_1-score of the classifier is more than

[17]The approach "Finetuned_MPNET" by JU-NLP fails on the responses generated by `Qwen3` and is omitted from the analyses for that dataset.

Table 11
Average effectiveness of the classifiers submitted to sub-task 2 on the responses generated for sub-task 1. The table reports the average value for the responses generated by the *Same* team vs. those generated by *Other* teams. The submissions are sorted alphabetically by team and approach. Baseline submissions are shown in gray and the "modernbert-embed-base"-baseline is the classifier used for evaluation of sub-task 1.

		Precision		Recall		F1	
Team	Approach	Own	Other	Own	Other	Own	Other
Git Gud	Deberta-Large-V2	0.988	0.961	0.498	0.220	0.656	0.352
Git Gud	Roberta-Large	0.997	0.975	0.626	0.186	0.763	0.304
JU-NLP	DebertaFineTuned	0.941	0.928	0.669	0.739	0.781	0.822
JU-NLP	Finetuned_MPNET	0.773	0.727	0.812	1.000	0.775	0.842
JU-NLP	Finetuned_MPNET_v2	0.959	0.961	0.234	0.276	0.376	0.395
Pirate Passau	All-mini-LM-v2-finetuned	—	0.917	—	0.342	—	0.495
Pirate Passau	all-mini+Random-forest	—	0.917	—	0.015	—	0.030
Pirate Passau	MPnet-finetuned	—	0.833	—	0.899	—	0.861
Pirate Passau	Tf-IDF-Logestic-Regression	—	0.834	—	0.680	—	0.740
TeamCMU	deberta-synthetic-curriculum	0.952	0.966	0.225	0.660	0.364	0.780
Baseline	minilm-baseline	—	0.829	—	0.229	—	0.352
Baseline	modernbert-embed-base	—	0.952	—	0.847	—	0.893
Baseline	naive-bayes-10	—	0.742	—	0.865	—	0.798
Baseline	naive-bayes-25	—	0.783	—	0.343	—	0.474
Baseline	naive-bayes-40	—	0.791	—	0.087	—	0.154

twice as high on responses by other teams (0.78 vs. 0.36). This difference stems almost exclusively from a lower recall of 0.23 on their own responses vs. 0.66 on those generated by Git Gud and JU-NLP. Team Git Gud also use their classifier in response generation by regenerating a response if it is detected by the classifier. This, however, does not translate into the same effect as for TeamCMU. Instead, their classifier is consistently more effective on their own responses than on those by JU-NLP and TeamCMU.

8. Conclusion

The sixth edition of the Touché lab on argumentation systems featured four tasks: (1) Retrieval-Augmented Debating, (2) Ideology and Power Identification in Parliamentary Debates, and (3) Image Retrieval/Generation for Arguments, and (4) Advertisement in Retrieval-Augmented Generation. We added two new tasks, one featuring interactive evaluation of argumentation systems and the other one focusing on the generation and detection of advertisement in generative retrieval systems. In comparison to last year the Ideology and Power Identification in Parliamentary Debates task included an additional sub-task on populism classification. Moreover, for the Image Retrieval/Generation for Arguments task, we changed the task from providing pro and con images to a topic to the less ambiguous providing images that convey a claim.

Of the 62 registered teams, 12 participated in the tasks and submitted a total of 60 runs. Unsurprisingly, large language models and generative approaches were used across tasks. For the Retrieval-Augmented Debating task, teams prompted language models in various ways to retrieve, select, phrase, and evaluate. For the Ideology and Power Identification in Parliamentary Debates task, teams used varying approaches, including traditional classifiers, fine-tuning encoder-only language models and prompting-based approaches using large language models. For the Image Retrieval/Generation for Arguments task, teams used CLIP to retrieve relevant images to Stable Diffusion to generate new ones. For the Advertisement in Retrieval-Augmented Generation task, teams primarily used encoder models like

MiniLM, MPNet, RoBERTa and DeBERTa-v3 to perform advertisement detection. The generation of responses was done with different versions of the Qwen and Mistral models.

We plan to continue Touché as a collaborative platform for researchers in argumentation systems. All Touché resources are freely available, including topics, manual relevance, argument quality, and stance judgments, and submitted runs from participating teams. In all Touché labs combined, we received 384 runs from 106 teams. We manually labeled the relevance and quality of more than 42,000 argumentative texts, debates, web documents, and images for 327 topics (topics and judgments are publicly available at the lab's web page, https://touche.webis.de). These resources and other events such as workshops will help to further foster the community working on argumentation systems.

Acknowledgments

This work was partially supported by the European Commission under grant agreement GA 101070014 (https://openwebsearch.eu) and by the German Federal Ministry of Education and Research (BMBF) through the project "DIALOKIA: Überprüfung von LLM-generierter Argumentation mittels dialektischem Sprachmodell" (01IS24084A-B).

Declaration on Generative AI

During the preparation of this work, the authors used DeepL, Grammarly, and Language Tool in order to: Grammar and spelling check, paraphrase and reword. Further, the authors used Stable Diffusion 3.5 for Table 6 in order to: Generate images (in line with the section's core topic). Further, the authors used ChatGPT in order to: Paraphrase and reword, improve writing style. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Kiesel, Çağrı Çöltekin, M. Gohsen, S. Heineking, M. Heinrich, M. Fröbe, T. Hagen, M. Aliannejadi, S. Anand, T. Erjavec, M. Hagen, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, H. Scells, M. Wolter, I. Zelch, M. Potthast, B. Stein, Overview of Touché 2025: Argumentation Systems, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 16th International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, 2025.

[2] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, S. Anand, B. Stein, Overview of ImageCLEF 2025: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 16th International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[3] K. Iordanou, C. Rapanta, "Argue With Me": A Method for Developing Argument Skills, Frontiers in Psychology 12 (2021). doi:10.3389/fpsyg.2021.631203.

[4] D. Kuhn, Science as Argument: Implications for Teaching and Learning Scientific Thinking, Science Education 77 (1993) 319–337. doi:10.1002/sce.3730770306.

[5] T. Wambsganss, T. Kueng, M. Soellner, J. M. Leimeister, ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–13. doi:10.1145/3411764.3445781.

[6] N. Slonim, Y. Bilu, C. Alzate, R. Bar-Haim, B. Bogin, F. Bonin, L. Choshen, E. Cohen-Karlik, L. Dankin, L. Edelstein, L. Ein-Dor, R. Friedman-Melamed, A. Gavron, A. Gera, M. Gleize, S. Gretz, D. Gutfreund, A. Halfon, D. Hershcovich, R. Hoory, Y. Hou, S. Hummel, M. Jacovi, C. Jochim, Y. Kantor, Y. Katz, D. Konopnicki, Z. Kons, L. Kotlerman, D. Krieger, D. Lahav, T. Lavee, R. Levy, N. Liberman, Y. Mass, A. Menczel, S. Mirkin, G. Moshkowich, S. Ofek-Koifman, M. Orbach, E. Rabinovich, R. Rinott, S. Shechtman, D. Sheinwald, E. Shnarch, I. Shnayderman, A. Soffer, A. Spector, B. Sznajder, A. Toledo, O. Toledo-Ronen, E. Venezian, R. Aharonov, An Autonomous Debating System, Nature 591 (2021) 379–384. doi:10.1038/s41586-021-03215-w.

[7] J. Kiesel, Ç. Çöltekin, M. Heinrich, M. Fröbe, M. Alshomary, B. D. Longueville, T. Erjavec, N. Handke, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, T. Reitis-Münstermann, M. Scharfbillig, N. Stefanovitch, H. Wachsmuth, M. Potthast, B. Stein, Overview of Touché 2024: Argumentation Systems, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. D. Nunzio, L. Soulier, P. Galuscakova, A. G. S. Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[8] A. Arian, M. Shamir, The primarily political functions of the left-right continuum, Comparative politics 15 (1983) 139–158.

[9] F. Vegetti, D. Širinić, Left–right Categorization and Perceptions of Party Ideologies, Political Behavior 41 (2019) 257–280.

[10] T. van Dijk, Discourse and Power, Bloomsbury Publishing, 2008.

[11] N. Fairclough, Critical Discourse Analysis: The Critical Study of Language, Longman applied linguistics, Taylor & Francis, 2013. doi:10.4324/9781315834368.

[12] N. Fairclough, Language and Power, Language In Social Life, Taylor & Francis, 2013. doi:10.4324/9781315838250.

[13] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, F. Menczer, Predicting the political alignment of Twitter users, in: Proc. of PASSAT and SocialCom, IEEE, 2011, pp. 192–199. doi:10.1109/PASSAT/SocialCom.2011.34.

[14] S. Gerrish, D. M. Blei, Predicting Legislative Roll Calls from Text, in: L. Getoor, T. Scheffer (Eds.), Proc. of ICML, Omnipress, 2011, pp. 489–496.

[15] D. Preoţiuc-Pietro, Y. Liu, D. Hopkins, L. Ungar, Beyond Binary Labels: Political Ideology Prediction of Twitter Users, in: R. Barzilay, M.-Y. Kan (Eds.), Proc. of ACL, ACL, 2017, pp. 729–740. doi:10.18653/v1/P17-1068.

[16] F. Pla, L.-F. Hurtado, Political Tendency Identification in Twitter using Sentiment Analysis Techniques, in: J. Tsujii, J. Hajic (Eds.), Proc. of Coling, Dublin City University and ACL, 2014, pp. 183–192. URL: https://aclanthology.org/C14-1019.

[17] C. Chen, D. Walker, V. Saligrama, Ideology Prediction from Scarce and Biased Supervision: Learn to Disregard the "What" and Focus on the "How"!, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proc. of ACL (Volume 1: Long Papers), ACL, Toronto, Canada, 2023, pp. 9529–9549. doi:10.18653/v1/2023.acl-long.530.

[18] J. A. García-Díaz, et al., Overview of PoliticES 2022: Spanish Author Profiling for Political Ideology, Procesamiento del Lenguaje Natural 69 (2022) 265–272. doi:10.26342/2022-69-23.

[19] D. Russo, et al., PoliticIT at EVALITA 2023: Overview of the political ideology detection in Italian texts task, in: Proc. of EVALITA, volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3473/paper7.pdf.

[20] G. M. Kurtoğlu Eskişar, Ç. Çöltekin, Emotions Running High? A Synopsis of the state of Turkish Politics through the ParlaMint Corpus, in: D. Fišer, M. Eskevich, J. Lenardič, F. de Jong (Eds.), Proc. of ParlaCLARIN, ELRA, 2022, pp. 61–70. URL: https://aclanthology.org/2022.parlaclarin-1.10.

[21] M. Mochtak, P. Rupnik, N. Ljubešić, The ParlaSent Multilingual Training Dataset for Sentiment Identification in Parliamentary Proceedings, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proc. of LREC, ELRA and ICCL, 2024, pp. 16024–16036. URL: https://aclanthology.org/2024.lrec-main.1393.

[22] O. Tarkka, J. Koljonen, M. Korhonen, J. Laine, K. Martiskainen, K. Elo, V. Laippala, Automated Emotion Annotation of Finnish Parliamentary Speeches Using GPT-4, in: D. Fiser, M. Eskevich, D. Bordon (Eds.), Proc. of ParlaCLARIN, ELRA and ICCL, 2024, pp. 70–76. URL: https://aclanthology.org/2024.parlaclarin-1.11.

[23] C. Navarretta, D. Haltrup Hansen, Government and opposition in Danish parliamentary debates, in: D. Fiser, M. Eskevich, D. Bordon (Eds.), Proc. of ParlaCLARIN, ELRA and ICCL, 2024, pp. 154–162. URL: https://aclanthology.org/2024.parlaclarin-1.23.

[24] K. A. Hawkins, R. E. Carlin, L. Littvay, C. R. Kaltwasser (Eds.), The Ideational Approach to Populism: Concept, Theory, and Analysis, Extremism and Democracy, Routledge, 2019.

[25] P. Norris, Measuring populism worldwide, Party politics 26 (2020) 697–717.

[26] M. Rooduijn, A. L. P. Pirro, D. Halikiopoulou, C. Froio, S. Van Kessel, S. L. De Lange, C. Mudde, P. Taggart, The PopuList: A Database of Populist, Far-Left, and Far-Right Parties Using Expert-Informed Qualitative Comparative Classification (EiQCC), British Journal of Political Science 54 (2024) 969–978. doi:10.1017/S0007123423000431.

[27] C. Dutilh Novaes, Argument and Argumentation, in: E. N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, Fall 2022 ed., Metaphysics Research Lab, Stanford University, 2022.

[28] M. Lewiński, D. Mohammed, Argumentation Theory, in: K. B. Jensen, R. T. Craig, J. Pooley, E. W. Rothenbuhler (Eds.), The International Encyclopedia of Communication Theory and Philosophy, Wiley, Hoboken, NJ, 2016. doi:10.1002/9781118766804.wbiect198.

[29] L. Groarke, Informal Logic, in: E. N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, Spring 2024 ed., Metaphysics Research Lab, Stanford University, 2024.

[30] M. Champagne, A.-V. Pietarinen, Why images cannot be arguments, but moving ones might, Argumentation 34 (2020) 207–236. doi:10.1007/s10503-019-09484-0.

[31] F. Dunaway, Images, Emotions, Politics, Modern American History 1 (2018) 369–376. doi:10.1017/mah.2018.17.

[32] J. E. Kjeldsen, The Rhetoric of Thick Representation: How Pictures Render the Importance and Strength of an Argument Salient, Argumentation 29 (2015) 197–215. doi:10.1007/s10503-014-9342-2.

[33] D. Fleming, Can pictures be arguments?, Argumentation and Advocacy 33 (1996) 11–22.

[34] I. J. Dove, On Images as Evidence and Arguments, in: F. H. van Eemeren, B. Garssen (Eds.), Topical Themes in Argumentation Theory: Twenty Exploratory Studies, Argumentation Library, Springer Netherlands, Dordrecht, 2012, pp. 223–238. doi:10.1007/978-94-007-4041-9_15.

[35] I. Grancea, Types of Visual Arguments, Argumentum. Journal of the Seminar of Discursive Logic, Argumentation Theory and Rhetoric 15 (2017) 16–34.

[36] D. Dimitrov, B. Bin Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino, SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images, in: Proc. of SemEval, ACL, 2021, pp. 70–98. doi:10.18653/v1/2021.semeval-1.7.

[37] S. Wu, D. A. Smith, Composition and Deformance: Measuring Imageability with a Text-to-Image Model, CoRR abs/2306.03168 (2023). doi:10.48550/ARXIV.2306.03168. arXiv:2306.03168.

[38] M. Brysbaert, A. B. Warriner, V. Kuperman, Concreteness ratings for 40 thousand generally known english word lemmas, Behavior Research Methods 46 (2014) 904–911. doi:10.3758/s13428-013-0403-5.

[39] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, B. Stein, Computational Argumentation Quality Assessment in Natural Language, in: Proceedings of EACL 2017, 2017, pp. 176–187. URL: https://aclanthology.org/E17-1017/.

[40] S. E. Spatharioti, D. M. Rothschild, D. G. Goldstein, J. M. Hofman, Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment, CoRR abs/2307.03744 (2023). doi:10.48550/ARXIV.2307.03744.

[41] I. Zelch, M. Hagen, M. Potthast, A User Study on the Acceptance of Native Advertising in Generative IR, in: ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2024), ACM, 2024. doi:10.1145/3627508.3638316.
[42] X. Chen, W. Feng, Z. Du, W. Wang, Y. Chen, H. Wang, L. Liu, Y. Li, J. Zhao, Y. Li, Z. Zhang, J. Lv, J. Shen, Z. Lin, J. Shao, Y. Shao, X. You, C. Gao, N. Sang, CTR-Driven Advertising Image Generation with Multimodal Large Language Models, in: Proceedings of the ACM Web Conference 2025, WWW '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 2262–2275. doi:10.1145/3696410.3714836.
[43] J. Huang, M. Qu, L. Li, Y. Wei, AdGPT: Explore Meaningful Advertising with ChatGPT, ACM Trans. Multimedia Comput. Commun. Appl. 21 (2025). doi:10.1145/3720546.
[44] S. Feizi, M. Hajiaghayi, K. Rezaei, S. Shin, Online Advertisements with LLMs: Opportunities and Challenges, 2024. URL: https://arxiv.org/abs/2311.07601. arXiv:2311.07601.
[45] M. Hajiaghayi, S. Lahaie, K. Rezaei, S. Shin, Ad Auctions for LLMs via Retrieval Augmented Generation, 2024. URL: http://papers.nips.cc/paper_files/paper/2024/hash/20dcab0f14046a5c6b02b61da9f13229-Abstract-Conference.html.
[46] S. Schmidt, I. Zelch, J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Detecting Generated Native Ads in Conversational Search, in: Companion Proceedings of the ACM Web Conference 2024, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 722–725. doi:10.1145/3589335.3651489.
[47] E. E. Schauster, P. Ferrucci, M. S. Neill, Native Advertising is the New Journalism: How Deception Affects Social Responsibility, American Behavioral Scientist 60 (2016) 1408–1424.
[48] B. W. Wojdynski, N. J. Evans, Going Native: Effects of Disclosure Position and Language onthe Recognition and Evaluation of Online Native Advertising, Journal of Advertising 45 (2016) 157–168.
[49] C. Campbell, P. E. Grimm, The challenges native advertising poses: Exploring potential federal trade commission responses and identifying research needs, Journal of Public Policy & Marketing 38 (2019) 110–123.
[50] B. Eyada, A. Milla, Native Advertising: Challenges and Perspectives, Journal of Design Sciences and Applied Arts 1 (2020) 67–77.
[51] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.
[52] T. Hagen, M. Fröbe, J. H. Merker, H. Scells, M. Hagen, M. Potthast, TIREx Tracker: The Information Retrieval Experiment Tracker, in: 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2025), ACM, 2025. doi:10.1145/3726302.3730297.
[53] T. Breuer, J. Keller, P. Schaer, ir_metadata: An extensible metadata schema for IR experiments, in: E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, G. Kazai (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 3078–3089. doi:10.1145/3477495.3531738.
[54] H. Grice, Studies in the Way of Words, William James lectures, Harvard University Press, 1989.
[55] G. Skitalinskaya, J. Klaff, H. Wachsmuth, Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021, Association for Computational Linguistics, 2021, pp. 1718–1729. doi:10.18653/V1/2021.EACL-MAIN.147.
[56] D. Zhang, J. Li, Z. Zeng, F. Wang, Jasper and Stella: Distillation of SOTA Embedding Models, CoRR abs/2412.19048 (2024). doi:10.48550/ARXIV.2412.19048. arXiv:2412.19048.
[57] M. E. Vallecillo-Rodríguez, M. T. Martín-Valdivia, A. Montejo-Ráez, SINAI at Touché: Leveraging Guided Prompt Strategies for Retrieval-Augmented Debate, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum,

CEUR Workshop Proceedings, 2025.
[58] A. Miyaguchi, C. Johnston, A. Potdar, DS@GT at Touché: Large Language Models for Retrieval-Augmented Debate, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[59] T. Erjavec, M. Ogrodniczuk, P. Osenova, N. Ljubešić, K. Simov, A. Pančur, M. Rudolf, M. Kopp, S. Barkarson, S. Steingrímsson, et al., The ParlaMint Corpora of Parliamentary Üroceedings, Language resources and evaluation 57 (2023) 415–448.
[60] T. Erjavec, M. Kopp, N. Ljubešić, T. Kuzman, P. Rayson, P. Osenova, M. Ogrodniczuk, Ç. Çöltekin, D. Koržinek, K. Meden, et al., ParlaMint II: Advancing Comparable Parliamentary Corpora Across Europe, Language Resources and Evaluation (2024) 1–32.
[61] A. Lührmann, N. Düpont, M. Higashijima, Y. B. Kavasoglu, K. L. Marquardt, M. Bernhard, H. Döring, A. Hicken, M. Laebens, S. I. Lindberg, J. Medzihorsky, A. Neundorf, O. J. Reuter, S. Ruth-Lovell, K. R. Weghorst, N. Wiesehomeier, J. Wright, N. Alizada, P. Bederke, L. Gastaldi, S. Grahn, G. Hindle, N. Ilchenko, J. von Römer, S. Wilson, D. Pemstein, B. Seim, Varieties of Party Identity and Organization (V-Party) Dataset V1, 2020. doi:10.23696/vpartydsv1, date accessed: 22 February 2021.
[62] D. Pemstein, K. L. Marquardt, E. Tzelgov, Y.-t. Wang, J. Medzihorsky, J. Krusell, F. Miri, J. von Römer, The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data, 2020.
[63] Ç. Çöltekin, M. Kopp, M. Katja, V. Morkevicius, N. Ljubešić, T. Erjavec, Multilingual Power and Ideology identification in the Parliament: a reference dataset and simple baselines, in: D. Fiser, M. Eskevich, D. Bordon (Eds.), Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 94–100. URL: https://aclanthology.org/2024.parlaclarin-1.14/.
[64] J. Vázquez-Osorio, L. A. H. Miranda, G. S. Adrián Juárez-Pérez, G. Bel-Enguix, GIL_UNAM_Iztacala at Touché: Benchmarking Classical Models for Multilingual Political Stance and Power Classification, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[65] M. Marogel, S. Gheorghe, Munibuc at Touché: Generalist Embeddings for Orientation and Populism Detection, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[66] C. Lee, R. Roy, M. Xu, J. Raiman, M. Shoeybi, B. Catanzaro, W. Ping, NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models, arXiv preprint arXiv:2405.17428 (2024).
[67] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7B, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.
[68] A. Shamsutdinov, J. Cherta-Rodriguez, TüNLP at Touché: Finetuning Multilingual Models for Ideology detection, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[69] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.
[70] B. Callac, A.-G. Bosser, F. D. de Saint-Cyr, E. Maisel, DEMA²IN at Touché: Salient Events Extraction for Ideology and Power Identification in Parliamentary Debates, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.
[71] M. Heinrich, J. Kiesel, M. Wolter, M. Potthast, B. Stein, Touché25-Image-Retrieval-and-Generation-for-Arguments, 2024. doi:10.5281/zenodo.14258397.

[72] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, R. Rombach, Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, 2024. URL: https://arxiv.org/abs/2403.03206. arXiv:2403.03206.

[73] N. Deckers, M. Fröbe, J. Kiesel, G. Pandolfo, C. Schröder, B. Stein, M. Potthast, The Infinite Index: Information Retrieval on Generative Text-To-Image Models, in: J. Gwizdka, S. Y. Rieh (Eds.), ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2023), ACM, 2023, pp. 172–186. doi:10.1145/3576840.3578327.

[74] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[75] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[76] D. A. G. Amaya, J. E. S. Castañeda, J. C. Martínez-Santos, E. Puertas, CEDNAV–UTB at Touché: Efficient Image Retrieval for Arguments with CLIP, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[77] T. Ramirez-delreal, D. Moctezuma, G. Ruiz, M. Graff, E. Tellez, Infotec+CentroGEO at Touché: MCIP, CLIP and SBERT as retrieval score, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[78] K. Schall, K. U. Barthel, N. Hezel, K. Jung, Optimizing CLIP Models for Image Retrieval with Maintained Joint-Embedding Alignment, in: E. Chávez, B. B. Kimia, J. Lokoc, M. Patella, J. Sedmidubský (Eds.), Similarity Search and Applications - 17th International Conference, SISAP 2024, volume 15268 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 97–110. doi:10.1007/978-3-031-75823-2_9.

[79] S. Anand, M. Heinrich, Hanuman at Touché: Image Generation with Argument-Aspect Fusion , in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[80] A. Dubey, et al., The Llama 3 Herd of Models, CoRR abs/2407.21783 (2024). doi:10.48550/ARXIV.2407.21783. arXiv:2407.21783.

[81] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual Instruction Tuning, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

[82] S. Kamani, M. Taqi, M. A. Chaudhry, M. A. H. Hanif, F. Alvi, A. Samad, Git Gud at Touché: Unified RAG Pipeline for Native Ad Generation and Detection, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[83] A. Dutta, A. Majumdar, S. Biswas, D. Saha, P. Pal, JU-NLP at Touché: Covert Advertisement in Conversational AI-Generation and Detection Strategies, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[84] J. Hong, N. Lee, J. Thorne, ORPO: Monolithic Preference Optimization without Reference Model, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 11170–11189. doi:10.18653/v1/2024.emnlp-main.626.

[85] T. A. Bouhairi, A. Alhamzeh, Pirate Passau at Touché: Do We Need to Get Complex? A Comparative

Analysis of Traditional and Advanced NLP Approaches for Advertisement Classification, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[86] T. E. Kim, J. Coelho, G. Onilude, J. Singh, TeamCMU at Touché: Adversarial Co-Evolution for Advertisement Integration and Detection in Conversational Search, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

A. Cross-Submission Results of Touché 2025: Advertisement in Retrieval-Augmented Generation

Table 12
Achieved precision of the classifiers submitted to sub-task 2 on the responses generated for sub-task 1. The column names (submissions to sub-task 1) are shortened (see Table 9 for the full names) and the submissions to sub-task 2 are sorted alphabetically by team and approach. Baseline submissions are shown in gray and the "modernbert-embed-base"-baseline is the classifier used for evaluation of sub-task 1.

		Git Gud		JU-NLP		TeamCMU
Team	Approach	Qwen3	Qwen2.5	Mistral7b_v2	Mistral7b	Adrewriting
Git Gud	Deberta-Large-V2	0.994	0.981	0.988	0.953	0.941
Git Gud	Roberta-Large	1.000	0.993	1.000	0.926	1.000
JU-NLP	DebertaFineTuned	0.973	0.959	0.965	0.916	0.852
JU-NLP	Finetuned_MPNET	—	0.726	0.728	0.817	0.728
JU-NLP	Finetuned_MPNET_v2	1.000	0.983	1.000	0.918	0.900
Pirate Passau	All-mini-LM-v2-finetuned	0.942	0.975	0.937	0.922	0.807
Pirate Passau	all-mini+Random-forest	1.000	1.000	1.000	0.750	0.833
Pirate Passau	MPnet-finetuned	0.865	0.859	0.856	0.853	0.733
Pirate Passau	Tf-IDF-Logestic-Regression	0.798	0.859	0.871	0.882	0.760
TeamCMU	deberta-synthetic-curriculum	0.939	0.995	0.982	0.947	0.952
Baseline	minilm-baseline	0.988	0.652	0.812	0.854	0.838
Baseline	modernbert-embed-base	0.984	0.960	1.000	0.995	0.821
Baseline	naive-bayes-10	0.734	0.747	0.745	0.742	0.742
Baseline	naive-bayes-25	0.758	0.753	0.800	0.850	0.752
Baseline	naive-bayes-40	0.810	0.714	0.852	0.848	0.730

Table 13
Achieved recall of the classifiers submitted to sub-task 2 on the responses generated for sub-task 1. The column names (submissions to sub-task 1) are shortened (see Table 9 for the full names) and the submissions to sub-task 2 are sorted alphabetically by team and approach. Baseline submissions are shown in gray and the "modernbert-embed-base"-baseline is the classifier used for evaluation of sub-task 1.

		Git Gud		JU-NLP		TeamCMU
Team	Approach	Qwen3	Qwen2.5	Mistral7b_v2	Mistral7b	Adrewriting
Git Gud	Deberta-Large-V2	0.603	0.393	0.311	0.228	0.120
Git Gud	Roberta-Large	0.738	0.513	0.288	0.187	0.082
JU-NLP	DebertaFineTuned	0.809	0.697	0.723	0.614	0.712
JU-NLP	Finetuned_MPNET	—	1.000	1.000	0.623	1.000
JU-NLP	Finetuned_MPNET_v2	0.543	0.217	0.258	0.210	0.067
Pirate Passau	All-mini-LM-v2-finetuned	0.483	0.292	0.333	0.352	0.251
Pirate Passau	all-mini+Random-forest	0.015	0.015	0.015	0.011	0.019
Pirate Passau	MPnet-finetuned	0.936	0.753	0.891	0.936	0.978
Pirate Passau	Tf-IDF-Logestic-Regression	0.607	0.479	0.682	0.757	0.876
TeamCMU	deberta-synthetic-curriculum	0.805	0.678	0.618	0.539	0.225
Baseline	minilm-baseline	0.296	0.057	0.196	0.287	0.311
Baseline	modernbert-embed-base	0.918	0.910	0.721	0.830	0.858
Baseline	naive-bayes-10	0.835	0.828	0.910	0.914	0.839
Baseline	naive-bayes-25	0.270	0.262	0.375	0.446	0.363
Baseline	naive-bayes-40	0.064	0.037	0.086	0.146	0.101

Table 14
Achieved F_1-score of the classifiers submitted to sub-task 2 on the responses generated for sub-task 1. The column names (submissions to sub-task 1) are shortened (see Table 9 for the full names) and the submissions to sub-task 2 are sorted alphabetically by team and approach. Baseline submissions are shown in gray and the "modernbert-embed-base"-baseline is the classifier used for evaluation of sub-task 1.

		Git Gud		JU-NLP		TeamCMU
Team	Approach	Qwen3	Qwen2.5	Mistral7b_v2	Mistral7b	Adrewriting
Git Gud	Deberta-Large-V2	0.751	0.561	0.473	0.369	0.213
Git Gud	Roberta-Large	0.849	0.677	0.448	0.312	0.152
JU-NLP	DebertaFineTuned	0.883	0.807	0.827	0.735	0.776
JU-NLP	Finetuned_MPNET	—	0.841	0.843	0.707	0.842
JU-NLP	Finetuned_MPNET_v2	0.704	0.356	0.411	0.341	0.125
Pirate Passau	All-mini-LM-v2-finetuned	0.639	0.450	0.492	0.509	0.383
Pirate Passau	all-mini+Random-forest	0.030	0.030	0.030	0.022	0.037
Pirate Passau	MPnet-finetuned	0.899	0.802	0.873	0.893	0.838
Pirate Passau	Tf-IDF-Logestic-Regression	0.689	0.615	0.765	0.815	0.814
TeamCMU	deberta-synthetic-curriculum	0.867	0.806	0.759	0.687	0.364
Baseline	minilm-baseline	0.455	0.104	0.316	0.429	0.454
Baseline	modernbert-embed-base	0.950	0.935	0.838	0.905	0.839
Baseline	naive-bayes-10	0.781	0.785	0.820	0.819	0.787
Baseline	naive-bayes-25	0.398	0.389	0.510	0.585	0.490
Baseline	naive-bayes-40	0.118	0.071	0.156	0.249	0.178