Overview of Touché 2025: Argumentation Systems

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

touche@webis.de touche.webis.de

Abstract This paper is the condensed 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.

Keywords: Advertisement Detection · Argumentation · Ideology Identification · Image Generation · Image Retrieval · Retrieval-Augmented Generation · User Simulation.

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 [43]) 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 pre-trained 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 `Qwen 2.5 7B` or `Mistral 7B` 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] An extended overview of this paper is published at CEUR-WS [48].

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

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.

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 [44]. 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 [51]. Following these hypotheses, ArgueTutor [82] 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 [75] 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 [47], 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 [3,79]. Power is another factor that shapes the political discourse [16,27,28]. Automatic identification of political orientation from texts has attracted considerable interest [14,33,64,63,10], including a few recent shared tasks [32,69]. 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 [52,59,77,60]. 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 [39,61,68], 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 [21,54]. In a world where digital communication increasingly relies on visual media, visual arguments are becoming ever more significant [36]. Images can enhance the acceptability of individual premises [9], and they also have the power to evoke strong emotional responses—such as anxiety, fear, or hope—or even to prescribe specific actions [20]. 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 [50]. Therefore, some scholars argue that images cannot constitute arguments [30]—but others contend that they can [18]. An additional perspective proposes that image sequences are more effective for conveying an argument [9]. 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 [50]. Therefore, images can serve as visual reasons, either reinforcing fact-based claims or questioning established beliefs [34].

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 [17]. 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 [84] and concreteness [6] 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 [81]—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 [76]. More closely related to our task, another study found that people struggle to identify advertisements in generated responses [85]. 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 [11,42]. This also includes the specific use case of trying to hide advertisements in the output of LLMs [29,38], as well as research on detecting these types of advertisements [72]. 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 [71,83] and product placement [8,26].

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 [31] 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 [37] 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 [5]. 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 did build a docker image from the code of some git repository, ensuring that the git repository is clean (i.e., all changes are committed and no untracked files), which allows to link a docker image to the exact version of a git repository that produced an submission, whereas software submissions 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 (i.e., the images were not public while the shared task was ongoing). To improve reproducibility, TIRA executes software in a sandbox by removing the internet connection (ensuring that the software is fully installed in the Docker image which eases rerunning software later, as libraries and models must be installed in an image). For the execution, participants could select the resources that their software had available for execution, from 1 CPU core with 10 GB RAM up 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, to allocate the resources that the participants selected.

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

[2] https://tira.io

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 [35], 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 [74]. 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 [86] (checkpoint: `dunzhang/stella_en_400M_v5`). The data is available online.[3]

Additionally, participants were provided a training set of 100 claims on various topics extracted from the Change My View subreddit.[4] 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.

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

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 [35]. 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?
- **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.[5]

Team SINAI [78] 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.

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

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. A random subset of the debates will be 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 F1 individually for each of the four maxims. The respective evaluation systems are then used to assess the debate systems from sub-task 1.

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 years 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 [32,69]. 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 the other two tasks, identifying power role and populism.

5.2 Data Description

The source of the data for this task is the ParlaMint version 4.1 [24], a uniformly encoded and annotated corpus of transcripts of parliamentary speeches from

multiple national and regional parliaments.[6] 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 [61,56,62].

For all tasks, the main challenge in the creation of a dataset is to minimize the effects of covariates [12]. 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 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.[7]

[6] Although all transcripts are obtained thorough 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 [24,23].

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

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 runs 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 [80] 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 [57] 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 [53], and using support vector classifiers on the extracted embeddings.

Team TüNLP [73] submitted results for only sub-task 1 (orientation) based on fine-tuning XLM-RoBERTa [13]. 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 [7] contributes to the shared tasks with a focused participation on data from a single parliament (GB). Their approach is based on extracting salient events Mistral-7b v0.2 Instruct [45]. 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.

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 [40]. Participants using the generation approach were supported with access to a Stable Diffusion-based image generation API [25], building on the concept of the Infinite Index [15].

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 [65] embeddings to measure similarity between claims and images, and another using SBERT [67] 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 [1] 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 [66] This team evaluated several embedding approaches for retrieval between images and claims using multimodal MCIP [70] 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 [2] This team uses an image generation pipeline. First, the LLaMA 3.2-3B [19] model extracts key aspects relevant to each argument. These aspects, along with the original argument, are provided as input to Mistral-7B [45], which generates a corresponding prompt for the image generator, emphasizing the relevant aspects. Afterwards, the corresponding image is generated using 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 [55], and then computing the cosine similarity between this description and the prompt used to create the image, using SBERT.

6.4 Task Evaluation

When crafting arguments for the task, the expert dataset creator envisioned a corresponding image and pinpointed two or more key aspects crucial for effectively visualizing the argument. For the evaluation of the task, each submitted image–argument pair is assessed based on how well each of these aspects is represented. A final relevance score is then assigned to each pair based on the individual aspect scores. The nDCG@5 score for a single argument is computed by comparing all submitted images for that argument. The final score is then obtained by averaging the nDCG@5 scores across all arguments.

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 whether 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.

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.

7.2 Data Description

For development purposes, we provided participants with the Webis Generated Native Ads 2024 dataset [72]. It contains 4,868 keyword queries, suitable items to be advertised, as well as 17,344 responses generated by Microsoft Copilot and YouChat. Into a third of the responses, we inserted advertisements 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. 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`. This 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. By sending the keyword queries for each meta-topic as well as the Google Trends topics to startpage.com, we collected 11,613 unique products and services to be paired with queries. Using these query-advertisement-pairs, we asked `GPT-4o-mini` to insert advertisements into the original responses collected from the conversational search engines. This resulted in a total of 16,051 responses with advertisements.

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 the generation sub-task. 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[8] 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

[8] https://trec-rag.github.io/about/

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 the classification sub-task. 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` [72], and a naive Bayes classifier using scikit-learn.[9] 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 [46] For sub-task 1, the team uses transformer-based reranking with `all-MiniLM-L6-v2` and `ms-marco-MiniLM-L6-v2` to retrieve document segments as context. The segments are given to `Qwen 2.5 7B` 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 [72]. Specifically, they trained `MPNet-v2`, `RoBERTa-base`/`-large`, `DeBERTa-v3-base`/`-large`, as well as a `RoBERTa-base` checkpoint published on Hugging Face.[10] As input to the models, they use the full response without additional data like the query.

Team JU-NLP [22] For sub-task 1, the team fine-tuned `Mistral-7B` to generate responses. The generation model was trained with Odds Ratio Preference Optimization (ORPO) [41] on pairs of responses with preference judgments obtained

[9] https://scikit-learn.org
[10] https://huggingface.co/0x7o/roberta-base-ad-detector

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 `all-mpnet-base-v2` fine-tuned on the Webis Generated Native Ads 2024 dataset [72]. 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.

Team Pirate Passau [4] This team submitted several approaches to sub-task 2 (detection of advertisements). 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 fine-tunes the transformer models `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 is again used for classification.

TeamCMU [49] 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 finds to be the best 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 used a best-of-N sampling method, selecting responses the classifier was least likely to identify as advertisements. The classifier (`DeBERTa-base`) was first trained on the Webis Generated Native Ads 2024 dataset [72], then improved through training on the synthetic datasets and responses created from the Ad-Rewriter. The same classifier was used for sub-task 2.

7.4 Task Evaluation

The evaluation of both sub-tasks is based on precision and recall. For sub-task 1, we added a linear layer to `modernbert-embed-base`[11] 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. [72]. 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. The primary score of a submission is based on the inverse recall of our classifier. This means that the score of a submission increases with the number of ads it successfully hides from the classifier. To contextualize the recall, we also report the precision of the classifier. 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 F1-score on the classification test split.

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 `Qwen 2.5 7B` and `Mistral 7B`.

We plan to continue Touché as a collaborative platform for researchers in argumentation systems. All Touché resources are freely available, including topics,

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

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 35,000 argumentative texts, web documents, and images for 227 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).

References

1. Amaya, D.A.G., Castañeda, J.E.S., Martínez-Santos, J.C., Puertas, E.: CEDNAV–UTB at Touché: Efficient Image Retrieval for Arguments with CLIP. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
2. Anand, S., Heinrich, M.: Hanuman at Touché: Image Generation with Argument-Aspect Fusion . In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
3. Arian, A., Shamir, M.: The primarily political functions of the left-right continuum. Comparative politics **15**(2), 139–158 (1983)
4. Bouhairi, T.A., Alhamzeh, A.: Pirate Passau at Touché: Do We Need to Get Complex? A Comparative Analysis of Traditional and Advanced NLP Approaches for Advertisement Classification. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
5. Breuer, T., Keller, J., Schaer, P.: ir_metadata: An extensible metadata schema for IR experiments. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022. pp. 3078–3089. ACM (2022). https://doi.org/10.1145/3477495.3531738
6. Brysbaert, M., Warriner, A.B., Kuperman, V.: Concreteness ratings for 40 thousand generally known english word lemmas. Behavior Research Methods **46**(3), 904–911 (2014). https://doi.org/10.3758/s13428-013-0403-5
7. Callac, B., Bosser, A.G., de Saint-Cyr, F.D., Maisel, E.: DEMA^{2}IN at Touché: Salient Events Extraction for Ideology and Power Identification in Parliamentary Debates. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
8. Campbell, C., Grimm, P.E.: The challenges native advertising poses: Exploring potential federal trade commission responses and identifying research needs. Journal of Public Policy & Marketing **38**(1), 110–123 (2019)

9. Champagne, M., Pietarinen, A.V.: Why images cannot be arguments, but moving ones might. Argumentation **34**(2), 207–236 (2020). https://doi.org/10.1007/s10503-019-09484-0
10. Chen, C., Walker, D., Saligrama, V.: Ideology Prediction from Scarce and Biased Supervision: Learn to Disregard the "What" and Focus on the "How"! In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proc. of ACL (Volume 1: Long Papers). pp. 9529–9549. ACL, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.acl-long.530
11. Chen, X., Feng, W., Du, Z., Wang, W., Chen, Y., Wang, H., Liu, L., Li, Y., Zhao, J., Li, Y., Zhang, Z., Lv, J., Shen, J., Lin, Z., Shao, J., Shao, Y., You, X., Gao, C., Sang, N.: CTR-Driven Advertising Image Generation with Multimodal Large Language Models. In: Proceedings of the ACM Web Conference 2025. p. 2262–2275. WWW '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3696410.3714836
12. Çöltekin, Ç., Kopp, M., Katja, M., Morkevicius, V., Ljubešić, N., Erjavec, T.: Multilingual Power and Ideology identification in the Parliament: a reference dataset and simple baselines. In: Fiser, D., Eskevich, M., Bordon, D. (eds.) Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024. pp. 94–100. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.parlaclarin-1.14/
13. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747
14. Conover, M.D., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of Twitter users. In: Proc. of PASSAT and SocialCom. pp. 192–199. IEEE (2011). https://doi.org/10.1109/PASSAT/SocialCom.2011.34
15. Deckers, N., Fröbe, M., Kiesel, J., Pandolfo, G., Schröder, C., Stein, B., Potthast, M.: The Infinite Index: Information Retrieval on Generative Text-To-Image Models. In: Gwizdka, J., Rieh, S.Y. (eds.) ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2023). pp. 172–186. ACM (Mar 2023). https://doi.org/10.1145/3576840.3578327
16. van Dijk, T.: Discourse and Power. Bloomsbury Publishing (2008)
17. Dimitrov, D., Bin Ali, B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., Nakov, P., Da San Martino, G.: SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images. In: Proc. of SemEval. pp. 70–98. ACL (2021). https://doi.org/10.18653/v1/2021.semeval-1.7
18. Dove, I.J.: On Images as Evidence and Arguments. In: van Eemeren, F.H., Garssen, B. (eds.) Topical Themes in Argumentation Theory: Twenty Exploratory Studies, pp. 223–238. Argumentation Library, Springer Netherlands, Dordrecht (2012). https://doi.org/10.1007/978-94-007-4041-9_15
19. Dubey, A., et al.: The Llama 3 Herd of Models. CoRR **abs/2407.21783** (2024). https://doi.org/10.48550/ARXIV.2407.21783
20. Dunaway, F.: Images, Emotions, Politics. Modern American History **1**(3), 369–376 (2018). https://doi.org/10.1017/mah.2018.17
21. Dutilh Novaes, C.: Argument and Argumentation. In: Zalta, E.N., Nodelman, U. (eds.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Fall 2022 edn. (2022)

22. Dutta, A., Majumdar, A., Biswas, S., Saha, D., Pal, P.: JU-NLP at Touché: Covert Advertisement in Conversational AI-Generation and Detection Strategies. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
23. Erjavec, T., Kopp, M., Ljubešić, N., Kuzman, T., Rayson, P., Osenova, P., Ogrodniczuk, M., Çöltekin, Ç., Koržinek, D., Meden, K., et al.: ParlaMint II: Advancing Comparable Parliamentary Corpora Across Europe. Language Resources and Evaluation pp. 1–32 (2024)
24. Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., et al.: The ParlaMint Corpora of Parliamentary Üroceedings. Language resources and evaluation **57**(1), 415–448 (2023)
25. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling Rectified Flow Transformers for High-Resolution Image Synthesis (2024), https://arxiv.org/abs/2403.03206
26. Eyada, B., Milla, A.: Native Advertising: Challenges and Perspectives. Journal of Design Sciences and Applied Arts **1**(1), 67–77 (2020)
27. Fairclough, N.: Critical Discourse Analysis: The Critical Study of Language. Longman applied linguistics, Taylor & Francis (2013). https://doi.org/10.4324/9781315834368
28. Fairclough, N.: Language and Power. Language In Social Life, Taylor & Francis (2013). https://doi.org/10.4324/9781315838250
29. Feizi, S., Hajiaghayi, M., Rezaei, K., Shin, S.: Online Advertisements with LLMs: Opportunities and Challenges (2024), https://arxiv.org/abs/2311.07601
30. Fleming, D.: Can pictures be arguments? Argumentation and Advocacy **33**, 11–22 (01 1996)
31. Fröbe, M., Wiegmann, M., Kolyada, N., Grahm, B., Elstner, T., Loebe, F., Hagen, M., Stein, B., Potthast, M.: Continuous Integration for Reproducible Shared Tasks with TIRA.io. In: Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A. (eds.) Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023). pp. 236–241. Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York (Apr 2023). https://doi.org/10.1007/978-3-031-28241-6_20
32. García-Díaz, J.A., et al.: Overview of PoliticES 2022: Spanish Author Profiling for Political Ideology. Procesamiento del Lenguaje Natural **69**, 265–272 (2022). https://doi.org/10.26342/2022-69-23
33. Gerrish, S., Blei, D.M.: Predicting Legislative Roll Calls from Text. In: Getoor, L., Scheffer, T. (eds.) Proc. of ICML. pp. 489–496. Omnipress (2011)
34. Grancea, I.: Types of Visual Arguments. Argumentum. Journal of the Seminar of Discursive Logic, Argumentation Theory and Rhetoric **15**(2), 16–34 (2017)
35. Grice, H.: Studies in the Way of Words. William James lectures, Harvard University Press (1989)
36. Groarke, L.: Informal Logic. In: Zalta, E.N., Nodelman, U. (eds.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Spring 2024 edn. (2024)
37. Hagen, T., Fröbe, M., Merker, J.H., Scells, H., Hagen, M., Potthast, M.: TIREx Tracker: The Information Retrieval Experiment Tracker. In: 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2025). ACM (Jul 2025). https://doi.org/10.1145/3726302.3730297

38. Hajiaghayi, M., Lahaie, S., Rezaei, K., Shin, S.: Ad Auctions for LLMs via Retrieval Augmented Generation (2024), http://papers.nips.cc/paper_files/paper/2024/hash/20dcab0f14046a5c6b02b61da9f13229-Abstract-Conference.html
39. Hawkins, K.A., Carlin, R.E., Littvay, L., Kaltwasser, C.R. (eds.): The Ideational Approach to Populism: Concept, Theory, and Analysis. Extremism and Democracy, Routledge (2019)
40. Heinrich, M., Kiesel, J., Wolter, M., Potthast, M., Stein, B.: Touché25-Image-Retrieval-and-Generation-for-Arguments (Dec 2024). https://doi.org/10.5281/zenodo.14258397
41. Hong, J., Lee, N., Thorne, J.: ORPO: Monolithic Preference Optimization without Reference Model. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 11170–11189. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.626
42. Huang, J., Qu, M., Li, L., Wei, Y.: AdGPT: Explore Meaningful Advertising with ChatGPT. ACM Trans. Multimedia Comput. Commun. Appl. **21**(4) (Apr 2025). https://doi.org/10.1145/3720546
43. Ionescu, B., Müller, H., Stanciu, D.C., Andrei, A.G., Radzhabov, A., Prokopchuk, Y., Ştefan, Liviu-Daniel, Constantin, M.G., Dogariu, M., Kovalev, V., Damm, H., Rückert, J., Ben Abacha, A., García Seco de Herrera, A., Friedrich, C.M., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Schmidt, C.S., Pakull, T.M.G., Bracke, B., Pelka, O., Eryilmaz, B., Becker, H., Yim, W.W., Codella, N., Novoa, R.A., Malvehy, J., Dimitrov, D., Das, R.J., Xie, Z., Shan, H.M., Nakov, P., Koychev, I., Hicks, S.A., Gautam, S., Riegler, M.A., Thambawita, V., Halvorsen, P., Fabre, D., Macaire, C., Lecouteux, B., Schwab, D., Potthast, M., Heinrich, M., Kiesel, J., Wolter, M., Anand, S., Stein, B.: Overview of ImageCLEF 2025: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications. In: de Albornoz, J.C., Gonzalo, J., Plaza, L., García Seco de Herrera, A., Mothe, J., Piroi, F., Rosso, P., Spina, D., Faggioli, G., Ferro, N. (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 (Sep 2025)
44. Iordanou, K., Rapanta, C.: "Argue With Me": A Method for Developing Argument Skills. Frontiers in Psychology **12** (Mar 2021). https://doi.org/10.3389/fpsyg.2021.631203
45. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7B (2023), https://arxiv.org/abs/2310.06825
46. Kamani, S., Taqi, M., Chaudhry, M.A., Hanif, M.A.H., Alvi, F., Samad, A.: Git Gud at Touché: Unified RAG Pipeline for Native Ad Generation and Detection. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
47. Kiesel, J., Çöltekin, Ç., Heinrich, M., Fröbe, M., Alshomary, M., Longueville, B.D., Erjavec, T., Handke, N., Kopp, M., Ljubešić, N., Meden, K., Mirzakhmedova, N., Morkevičius, V., Reitis-Münstermann, T., Scharfbillig, M., Stefanovitch, N., Wachsmuth, H., Potthast, M., Stein, B.: Overview of Touché 2024: Argumentation Systems. In: Goeuriot, L., Mulhem, P., Quénot, G., Schwab, D., Nunzio, G.M.D., Soulier, L., Galuscakova, P., Herrera, A.G.S., Faggioli, G., Ferro, N. (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 (Sep 2024)

48. Kiesel, J., Çağrı Çöltekin, Gohsen, M., Heineking, S., Heinrich, M., Fröbe, M., Hagen, T., Aliannejadi, M., Anand, S., Erjavec, T., Hagen, M., Kopp, M., Ljubešić, N., Meden, K., Mirzakhmedova, N., Morkevičius, V., Scells, H., Wolter, M., Zelch, I., Potthast, M., Stein, B.: Overview of Touché 2025: Argumentation Systems. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
49. Kim, T.E., Coelho, J., Onilude, G., Singh, J.: TeamCMU at Touché: Adversarial Co-Evolution for Advertisement Integration and Detection in Conversational Search. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
50. Kjeldsen, J.E.: The Rhetoric of Thick Representation: How Pictures Render the Importance and Strength of an Argument Salient. Argumentation **29**(2), 197–215 (2015). https://doi.org/10.1007/s10503-014-9342-2
51. Kuhn, D.: Science as Argument: Implications for Teaching and Learning Scientific Thinking. Science Education **77**(3), 319–337 (1993). https://doi.org/10.1002/sce.3730770306
52. Kurtoğlu Eskişar, G.M., Çöltekin, Ç.: Emotions Running High? A Synopsis of the state of Turkish Politics through the ParlaMint Corpus. In: Fišer, D., Eskevich, M., Lenardič, J., de Jong, F. (eds.) Proc. of ParlaCLARIN. pp. 61–70. ELRA (2022), https://aclanthology.org/2022.parlaclarin-1.10
53. Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., Ping, W.: NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. arXiv preprint arXiv:2405.17428 (2024)
54. Lewiński, M., Mohammed, D.: Argumentation Theory. In: Jensen, K.B., Craig, R.T., Pooley, J., Rothenbuhler, E.W. (eds.) The International Encyclopedia of Communication Theory and Philosophy. Wiley, Hoboken, NJ (2016). https://doi.org/10.1002/9781118766804.wbiect198
55. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (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), http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html
56. Lührmann, A., Düpont, N., Higashijima, M., Kavasoglu, Y.B., Marquardt, K.L., Bernhard, M., Döring, H., Hicken, A., Laebens, M., Lindberg, S.I., Medzihorsky, J., Neundorf, A., Reuter, O.J., Ruth-Lovell, S., Weghorst, K.R., Wiesehomeier, N., Wright, J., Alizada, N., Bederke, P., Gastaldi, L., Grahn, S., Hindle, G., Ilchenko, N., von Römer, J., Wilson, S., Pemstein, D., Seim, B.: Varieties of Party Identity and Organization (V-Party) Dataset V1 (2020). https://doi.org/10.23696/vpartydsv1, date accessed: 22 February 2021
57. Marogel, M., Gheorghe, S.: Munibuc at Touché: Generalist Embeddings for Orientation and Populism Detection. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)

58. Miyaguchi, A., Johnston, C., Potdar, A.: DS@GT at Touché: Large Language Models for Retrieval-Augmented Debate. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
59. Mochtak, M., Rupnik, P., Ljubešić, N.: The ParlaSent Multilingual Training Dataset for Sentiment Identification in Parliamentary Proceedings. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proc. of LREC. pp. 16024–16036. ELRA and ICCL (2024), https://aclanthology.org/2024.lrec-main.1393
60. Navarretta, C., Haltrup Hansen, D.: Government and opposition in Danish parliamentary debates. In: Fiser, D., Eskevich, M., Bordon, D. (eds.) Proc. of ParlaCLARIN. pp. 154–162. ELRA and ICCL (2024), https://aclanthology.org/2024.parlaclarin-1.23
61. Norris, P.: Measuring populism worldwide. Party politics **26**(6), 697–717 (2020)
62. Pemstein, D., Marquardt, K.L., Tzelgov, E., Wang, Y.t., Medzihorsky, J., Krusell, J., Miri, F., von Römer, J.: The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data (2020)
63. Pla, F., Hurtado, L.F.: Political Tendency Identification in Twitter using Sentiment Analysis Techniques. In: Tsujii, J., Hajic, J. (eds.) Proc. of Coling. pp. 183–192. Dublin City University and ACL (2014), https://aclanthology.org/C14-1019
64. Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., Ungar, L.: Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In: Barzilay, R., Kan, M.Y. (eds.) Proc. of ACL. pp. 729–740. ACL (2017). https://doi.org/10.18653/v1/P17-1068
65. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), http://proceedings.mlr.press/v139/radford21a.html
66. Ramirez-delreal, T., Moctezuma, D., Ruiz, G., Graff, M., Tellez, E.: Infotec+CentroGEO at Touché: MCIP, CLIP and SBERT as retrieval score. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
67. Reimers, N., Gurevych, I.: 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 (11 2019), https://arxiv.org/abs/1908.10084
68. Rooduijn, M., Pirro, A.L.P., Halikiopoulou, D., Froio, C., Van Kessel, S., De Lange, S.L., Mudde, C., Taggart, P.: 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**(3), 969–978 (2024). https://doi.org/10.1017/S0007123423000431
69. Russo, D., et al.: PoliticIT at EVALITA 2023: Overview of the political ideology detection in Italian texts task. In: Proc. of EVALITA. CEUR Workshop Proceedings, vol. 3473. CEUR-WS.org (2023), https://ceur-ws.org/Vol-3473/paper7.pdf
70. Schall, K., Barthel, K.U., Hezel, N., Jung, K.: Optimizing CLIP Models for Image Retrieval with Maintained Joint-Embedding Alignment. In: Chávez, E., Kimia, B.B., Lokoc, J., Patella, M., Sedmidubský, J. (eds.) Similarity Search and Applications - 17th International Conference, SISAP 2024. Lecture Notes in Com-

puter Science, vol. 15268, pp. 97–110. Springer (2024). https://doi.org/10.1007/978-3-031-75823-2_9
71. Schauster, E.E., Ferrucci, P., Neill, M.S.: Native Advertising is the New Journalism: How Deception Affects Social Responsibility. American Behavioral Scientist **60**(12), 1408–1424 (2016)
72. Schmidt, S., Zelch, I., Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: Detecting Generated Native Ads in Conversational Search. In: Companion Proceedings of the ACM Web Conference 2024. p. 722–725. WWW '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3589335.3651489
73. Shamsutdinov, A., Cherta-Rodriguez, J.: TüNLP at Touché: Finetuning Multilingual Models for Ideology detection. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
74. Skitalinskaya, G., Klaff, J., Wachsmuth, H.: Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale. In: Merlo, P., Tiedemann, J., Tsarfaty, R. (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. pp. 1718–1729. Association for Computational Linguistics (2021). https://doi.org/10.18653/V1/2021.EACL-MAIN.147
75. Slonim, N., Bilu, Y., Alzate, C., Bar-Haim, R., Bogin, B., Bonin, F., Choshen, L., Cohen-Karlik, E., Dankin, L., Edelstein, L., Ein-Dor, L., Friedman-Melamed, R., Gavron, A., Gera, A., Gleize, M., Gretz, S., Gutfreund, D., Halfon, A., Hershcovich, D., Hoory, R., Hou, Y., Hummel, S., Jacovi, M., Jochim, C., Kantor, Y., Katz, Y., Konopnicki, D., Kons, Z., Kotlerman, L., Krieger, D., Lahav, D., Lavee, T., Levy, R., Liberman, N., Mass, Y., Menczel, A., Mirkin, S., Moshkowich, G., Ofek-Koifman, S., Orbach, M., Rabinovich, E., Rinott, R., Shechtman, S., Sheinwald, D., Shnarch, E., Shnayderman, I., Soffer, A., Spector, A., Sznajder, B., Toledo, A., Toledo-Ronen, O., Venezian, E., Aharonov, R.: An Autonomous Debating System. Nature **591**(7850), 379–384 (Mar 2021). https://doi.org/10.1038/s41586-021-03215-w
76. Spatharioti, S.E., Rothschild, D.M., Goldstein, D.G., Hofman, J.M.: Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. CoRR **abs/2307.03744** (2023). https://doi.org/10.48550/ARXIV.2307.03744
77. Tarkka, O., Koljonen, J., Korhonen, M., Laine, J., Martiskainen, K., Elo, K., Laippala, V.: Automated Emotion Annotation of Finnish Parliamentary Speeches Using GPT-4. In: Fiser, D., Eskevich, M., Bordon, D. (eds.) Proc. of ParlaCLARIN. pp. 70–76. ELRA and ICCL (2024), https://aclanthology.org/2024.parlaclarin-1.11
78. Vallecillo-Rodríguez, M.E., Martín-Valdivia, M.T., Montejo-Ráez, A.: SINAI at Touché: Leveraging Guided Prompt Strategies for Retrieval-Augmented Debate. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)
79. Vegetti, F., Širinić, D.: Left–right Categorization and Perceptions of Party Ideologies. Political Behavior **41**(1), 257–280 (2019)
80. Vázquez-Osorio, J., Miranda, L.A.H., Adrián Juárez-Pérez, G.S., Bel-Enguix, G.: GIL_UNAM_Iztacala at Touché: Benchmarking Classical Models for Multilingual Political Stance and Power Classification. In: Faggioli, G., Ferro, N., Rosso, P., Spina, D. (eds.) Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings (2025)

81. Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T.A., Hirst, G., Stein, B.: Computational Argumentation Quality Assessment in Natural Language. In: Proceedings of EACL 2017. pp. 176–187 (Apr 2017), https://aclanthology.org/E17-1017/
82. Wambsganss, T., Kueng, T., Soellner, M., Leimeister, J.M.: ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–13. CHI '21, Association for Computing Machinery, New York, NY, USA (May 2021). https://doi.org/10.1145/3411764.3445781
83. Wojdynski, B.W., Evans, N.J.: Going Native: Effects of Disclosure Position and Language onthe Recognition and Evaluation of Online Native Advertising. Journal of Advertising **45**(2), 157–168 (2016)
84. Wu, S., Smith, D.A.: Composition and Deformance: Measuring Imageability with a Text-to-Image Model. CoRR **abs/2306.03168** (2023). https://doi.org/10.48550/ARXIV.2306.03168
85. Zelch, I., Hagen, M., Potthast, M.: 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). https://doi.org/10.1145/3627508.3638316
86. Zhang, D., Li, J., Zeng, Z., Wang, F.: Jasper and Stella: Distillation of SOTA Embedding Models. CoRR **abs/2412.19048** (2024). https://doi.org/10.48550/ARXIV.2412.19048