Touché @

MADRID

CLEF 2025

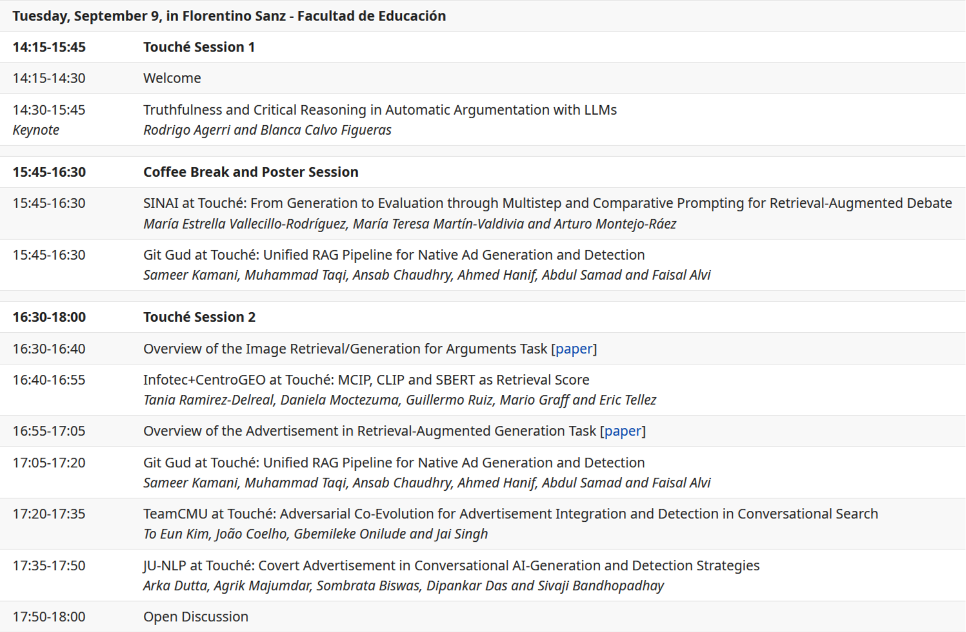
Conference and Labs of the Evaluation Forum

6th Edition

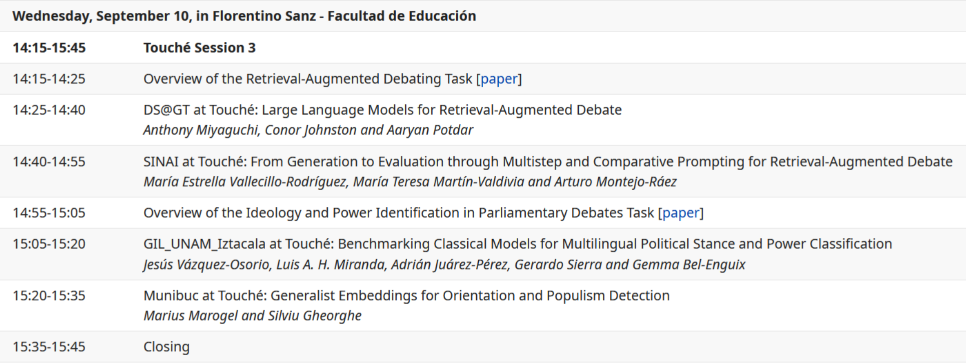
Argumentation Systems

touche.webis.de

Touché: Argumentation Systems



Tuesday, September 9, in Florentino Sanz - Facultad de Educación	
14:15-15:45	**Touché Session 1**
14:15-14:30	Welcome
14:30-15:45 *Keynote*	Truthfulness and Critical Reasoning in Automatic Argumentation with LLMs *Rodrigo Agerri and Blanca Calvo Figueras*
15:45-16:30	**Coffee Break and Poster Session**
15:45-16:30	SINAI at Touché: From Generation to Evaluation through Multistep and Comparative Prompting for Retrieval-Augmented Debate *María Estrella Vallecillo-Rodríguez, María Teresa Martín-Valdivia and Arturo Montejo-Ráez*
15:45-16:30	Git Gud at Touché: Unified RAG Pipeline for Native Ad Generation and Detection *Sameer Kamani, Muhammad Taqi, Ansab Chaudhry, Ahmed Hanif, Abdul Samad and Faisal Alvi*
16:30-18:00	**Touché Session 2**
16:30-16:40	Overview of the Image Retrieval/Generation for Arguments Task [paper]
16:40-16:55	Infotec+CentroGEO at Touché: MCIP, CLIP and SBERT as Retrieval Score *Tania Ramirez-Delreal, Daniela Moctezuma, Guillermo Ruiz, Mario Graff and Eric Tellez*
16:55-17:05	Overview of the Advertisement in Retrieval-Augmented Generation Task [paper]
17:05-17:20	Git Gud at Touché: Unified RAG Pipeline for Native Ad Generation and Detection *Sameer Kamani, Muhammad Taqi, Ansab Chaudhry, Ahmed Hanif, Abdul Samad and Faisal Alvi*
17:20-17:35	TeamCMU at Touché: Adversarial Co-Evolution for Advertisement Integration and Detection in Conversational Search *To Eun Kim, João Coelho, Gbemileke Onilude and Jai Singh*
17:35-17:50	JU-NLP at Touché: Covert Advertisement in Conversational AI-Generation and Detection Strategies *Arka Dutta, Agrik Majumdar, Sombrata Biswas, Dipankar Das and Sivaji Bandhopadhay*
17:50-18:00	Open Discussion



Wednesday, September 10, in Florentino Sanz - Facultad de Educación	
14:15-15:45	**Touché Session 3**
14:15-14:25	Overview of the Retrieval-Augmented Debating Task [paper]
14:25-14:40	DS@GT at Touché: Large Language Models for Retrieval-Augmented Debate *Anthony Miyaguchi, Conor Johnston and Aaryan Potdar*
14:40-14:55	SINAI at Touché: From Generation to Evaluation through Multistep and Comparative Prompting for Retrieval-Augmented Debate *María Estrella Vallecillo-Rodríguez, María Teresa Martín-Valdivia and Arturo Montejo-Ráez*
14:55-15:05	Overview of the Ideology and Power Identification in Parliamentary Debates Task [paper]
15:05-15:20	GIL_UNAM_Iztacala at Touché: Benchmarking Classical Models for Multilingual Political Stance and Power Classification *Jesús Vázquez-Osorio, Luis A. H. Miranda, Adrián Juárez-Pérez, Gerardo Sierra and Gemma Bel-Enguix*
15:20-15:35	Munibuc at Touché: Generalist Embeddings for Orientation and Populism Detection *Marius Marogel and Silviu Gheorghe*
15:35-15:45	Closing

4 Tasks:

1. Debate simulation
2. Debate analysis
3. Image retrieval or generation
4. Advertisements detection

`touche.webis.de`

Get the program!

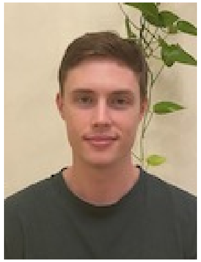
©touche.webis.de 2025

Retrieval-Augmented Debating (RAD)

Touché’25 Task 1

Marcel
Gohsen

Nailia
Mirzakhmedova

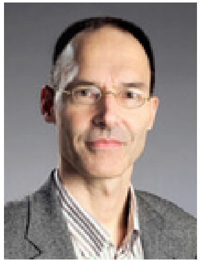


Harrisen
Scells

Mohammad
Aliannejadi

Maik
Fröbe

Johannes
Kiesel



Benno
Stein

©touche.webis.de 2025

Retrieval-Augmented Debating (RAD)

User — U_1: Claim statement

© touche.webis.de 2025

Retrieval-Augmented Debating (RAD)

User — U_1: Claim statement

S_1: Attack of U_1 — System

© touche.webis.de 2025

Retrieval-Augmented Debating (RAD)

User — U_1: Claim statement

S_1: Attack of U_1 — System

User — U_2: Attack of S_1

S_2: Attack of U_1, U_2 and/or defense of S_1 — System

User — U_3: Attack of S_1, S_2

S_3: Attack of U_1, U_2, U_3 and/or defense of S_1, S_2 — System

User — U_4: Attack of S_1, S_2, S_3

S_4: Attack of $U_1, \ldots, U_4$ and/or defense of $S_1, \ldots, S_3$ — System

User — U_5: Attack of $S_1, \ldots, S_4$

S_5: Attack of $U_1, \ldots, U_5$ and/or defense of $S_1, \ldots, S_4$ — System

Retrieval-Augmented Debating (RAD)

User — U_1: Claim statement

S_1: Attack of U_1 — System

User — U_2: Attack of S_1

S_2: Attack of U_1, U_2 and/or defense of S_1 — System

User — U_3: Attack of S_1, S_2

S_3: Attack of U_1, U_2, U_3 and/or defense of S_1, S_2 — System

User — U_4: Attack of S_1, S_2, S_3

S_4: Attack of $U_1, \ldots, U_4$ and/or defense of $S_1, \ldots, S_3$ — System

User — U_5: Attack of $S_1, \ldots, S_4$

S_5: Attack of $U_1, \ldots, U_5$ and/or defense of $S_1, \ldots, S_4$ — System

🟪 User simulator 🟨 Participant system

©touche.webis.de 2025

Retrieval-Augmented Debating (RAD)

Task Description

Scenario: Assisting people in forming an opinion on controversial topics and training argumentation skills

Sub-Task 1: Develop debate systems that retrieve and respond with counterarguments and evidence in simulated debates.

Sub-Task 2: Provide metrics to assess quality criteria based on Grice’s maxims of cooperation.

Quantity: at least one at most one of each attack/defense arguments?
Quality: response grounded on retrieved arguments?
Relation: response coherent with conversation?
Manner: response clear and precise?

Retrieval-Augmented Debating (RAD)

Dataset

Arguments

- 300 000 arguments from ClaimRev[1]
- Pre-indexed in Elasticsearch

Argument:	*Pineapple on pizza is an insult to the Italian origins of pizza.*
Supports:	*Pineapple does not belong on pizza.*
Attacks:	*Pineapple belongs on pizza.*

Claims and debates

- 100 claims from the Change My View subreddit[2]
- 100 simulated debates for claims with annotations
- Annotation: binary labels for quality criteria

---

[1] Skitalinskaya et al., Quality Assessment of Claims in Argumentation at Scale. EACL 2021.

[2] https://www.reddit.com/r/changemyview/

Retrieval-Augmented Debating (RAD)

Results: Sub-Task 1

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

Criteria: percentages of responses that fulfill given criteria.

Score: Avg. percentage of responses across all criteria.

©touche.webis.de 2025

Retrieval-Augmented Debating (RAD)

Results: Sub-Task 2

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
	org	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	DS@GT	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	DS@GT	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	DS@GT	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	DS@GT	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	SINAI	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	DS@GT	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	SINAI	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	SINAI	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	SINAI	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	DS@GT	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	SINAI	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	SINAI	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	SINAI	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

Retrieval-Augmented Debating (RAD)

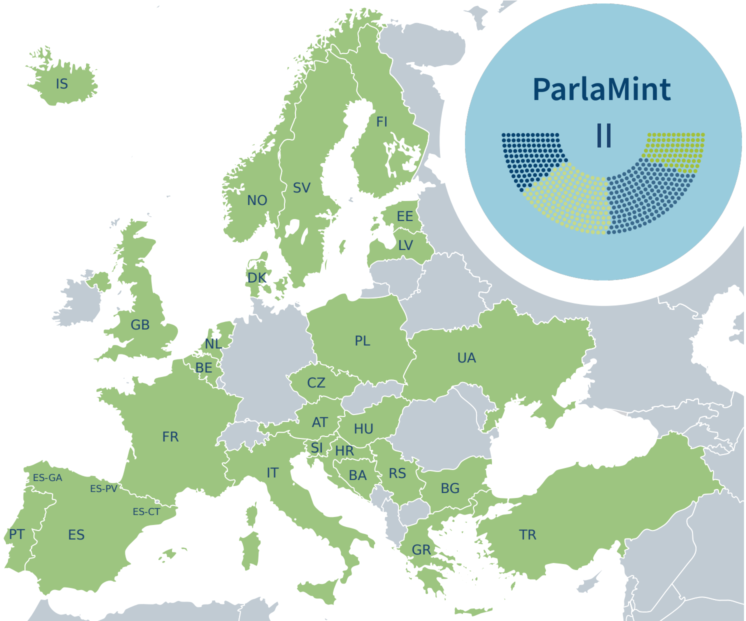
Observations

- Some claims too hard to argue (e.g., the earth is flat).
  - Participant systems admitted defeat (*“you are right”*).
- Grounding responses in retrieved argument is hard.
  - Low quality score for most systems.
- LLMs do not recognize stance switches.
  - Systems pretended to disagree but argued for user stance.
- Common problem: wordiness.
  - Complex vocabulary, unclear argument, repetition.

→ Building a persuasive debate system is a hard task.

Multilingual Ideology and Power Identification in Parliamentary Debates

Touché’25 Task 2



Çağrı
Çöltekin

Katja
Meden

Nikola
Ljubešić

Tomaž
Erjavec

Vaidas
Morkevičius

Matyáš
Kopp

Multilingual Ideology and Power Identification in Parliamentary Debates

Introduction

- Parliamentary debates result in decisions with high societal impact

- Political/parliamentary language is difficult to analyze
  - highly conventionalized
  - strategies like evasion, circumlocution or the use of metaphors are common

- This task is about identifying three fundamental aspects in political discourse
  - *Political orientation*: the ‘classic’ left–right spectrum
  - *Populism index*: another ‘popular’ dimension of recent political discourse
  - *Power role*: central in discourse analysis, virtually no computational studies

© touche.webis.de 2025

Multilingual Ideology and Power Identification in Parliamentary Debates

Task Description

Scenario: Identify the political orientation and the power role of the speaker from their speeches in parliamentary debates.

Task: Given a transcribed speech delivered in a parliament

Subtask 1: identify political orientation of the speaker (left–right)

Subtask 2: identify the position of the speaker's party in populsit–pluralist scale (4 values)

Subtask 3: identify power role of the speaker (coalition–opposition)

Data:
- A subset of the ParlaMint version 4.1
- 29 national and regional parliaments (some available only for one of the tasks)
- 30 languages (also automatic translation to English)
- Date range varies by parliament, but includes at least from 2015 to 2022
- Typically long texts (approx. 600 words on average)

©touche.webis.de 2025

Multilingual Ideology and Power Identification in Parliamentary Debates

Results - orientation

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

©touche.webis.de 2025

Multilingual Ideology and Power Identification in Parliamentary Debates

Results - populsim

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

©touche.webis.de 2025

Multilingual Ideology and Power Identification in Parliamentary Debates

Results - populsim

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

©touche.webis.de 2025

Multilingual Ideology and Power Identification in Parliamentary Debates

Results: observations

- Similar approaches to last year (with slightly reduced participant nunbers)
- Many teams used ‘traditional’ ML methods and (large) language models to extract features
  - likely the due to cost of processing long texts
- Finetuning a single multilingual model also seems promising
- Focused participation based on event extraction from one of the teams (DEMA^{2}IN)
- Populism identification proves to be most difficult
- Scores on English are much better than the average performance

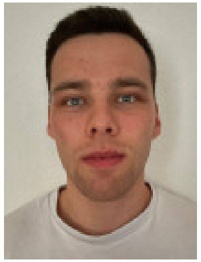
©touche.webis.de 2025

Image Retrieval/Generation for Arguments [Joint Task with ImageCLEF]

Touché’25 Task 3

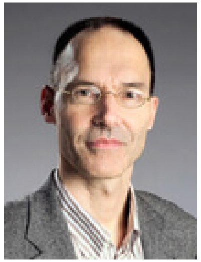
Maximilian
Heinrich

Johannes
Kiesel



Moritz
Wolter

Martin
Potthast



Benno
Stein

Image Retrieval/Generation for Arguments [Joint Task with ImageCLEF]

Task Description

Scenario: Enhance the impact of arguments.

Task: Given an argument, identify images that effectively convey the argument’s premise.

- Participants may either retrieve images from a dataset or generate them using a text-to-image model.

Data:

- 128 arguments across 27 topics
- ca. 32,000 crawled images with corresponding website information and additional metadata, including automatically generated captions

© touche.webis.de 2025

Image Retrieval/Generation for Arguments **[Joint Task with ImageCLEF]**

Task Description

Scenario: Enhance the impact of arguments.

Task: Given an argument, identify images that effectively convey the argument's premise.

- Participants may either retrieve images from a dataset or generate them using a text-to-image model.

Data:

- 128 arguments across 27 topics
- ca. 32,000 crawled images with corresponding website information and additional metadata, including automatically generated captions

Example

Topic: Public Transportation vs. Private Cars

Claim: Cars make it easy to transport goods and belongings

©touche.webis.de 2025

Image Retrieval/Generation for Arguments [Joint Task with ImageCLEF]

Evaluation

- For each argument, two aspects were identified, and each aspect was rated using the following scale:

  0: Aspect does not convey the claim
  1: Aspect partially conveys the claim
  2: Aspect fully conveys the claim

- For each annotator, the aspect scores were aggregated to derive a single rating for an argument-image pair.

- Final score for an argument-image pair is computed by combining the individual ratings from two annotators.

© touche.webis.de 2025

Image Retrieval/Generation for Arguments [Joint Task with ImageCLEF]

Example Submission

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

©touche.webis.de 2025

Image Retrieval/Generation for Arguments [Joint Task with ImageCLEF]

Example Submission

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

Here both images receive get a score of two. The two required aspects do not need to be combined in a precise way.

©touche.webis.de 2025

Image Retrieval/Generation for Arguments [Joint Task with ImageCLEF]

Results - Retrieval

Rank	Team	Approach	NDCG@5
1	Baseline	CLIP Image	0.855
2	Infotec+CentroGEO	OpenCLIP Image	0.836
3	Baseline	SBERT Website-Text	0.811
4	Infotec+CentroGEO	MCIP Image	0.794
5	Infotec+CentroGEO	SBERT Image-Text+Caption	0.755
6	CEDNAV–UTB	CLIP Image-Caption	0.236

© touche.webis.de 2025

Image Retrieval/Generation for Arguments [Joint Task with ImageCLEF]

Results - Retrieval

Rank	Team	Approach	NDCG@5
1	Baseline	CLIP Image	0.855
2	Infotec+CentroGEO	OpenCLIP Image	0.836
3	Baseline	SBERT Website-Text	0.811
4	Infotec+CentroGEO	MCIP Image	0.794
5	Infotec+CentroGEO	SBERT Image-Text+Caption	0.755
6	CEDNAV–UTB	CLIP Image-Caption	0.236

The ‘Approach’ column specifies how the embeddings for the images were generated and compared with the arguments. For example, ‘CLIP Image’ indicates that multimodal CLIP embeddings are employed.

©touche.webis.de 2025

Image Retrieval/Generation for Arguments [Joint Task with ImageCLEF]

Results - Generation

Rank	Team	Approach	NDCG@5
1	Hanuman	Generative Prompt	0.963
2	Baseline	Stable Diffusion 1.0	0.844
3	Baseline	Stable Diffusion 3.5	0.839

© touche.webis.de 2025

Image Retrieval/Generation for Arguments [Joint Task with ImageCLEF]

Results - Generation

Rank	Team	Approach	NDCG@5
1	Hanuman	Generative Prompt	0.963
2	Baseline	Stable Diffusion 1.0	0.844
3	Baseline	Stable Diffusion 3.5	0.839

Approaches:

- **Generative-Prompt:** Use an LLM to identify key aspects of the argument and compose a tailored image-generation prompt. For generation Stable Diffusion 1.0 is used.
- **Baseline:** Directly use the arguments themselves as the image-generation prompt.

© touche.webis.de 2025

Image Retrieval/Generation for Arguments [Joint Task with ImageCLEF]

Results - Generation

Rank	Team	Approach	NDCG@5
1	Hanuman	Generative Prompt	0.963
2	Baseline	Stable Diffusion 1.0	0.844
3	Baseline	Stable Diffusion 3.5	0.839

Approaches:

- ❑ **Generative-Prompt:** Use an LLM to identify key aspects of the argument and compose a tailored image-generation prompt. For generation Stable Diffusion 1.0 is used.
- ❑ **Baseline:** Directly use the arguments themselves as the image-generation prompt.

Image generation for arguments produces good results, especially when using carefully crafted custom prompts.

© touche.webis.de 2025

Image Retrieval/Generation for Arguments [Joint Task with ImageCLEF]

Lessons Learned

- Finding suitable images for arguments is challenging; generation often works better for specific arguments than retrieval.

- Retrieval approaches are constrained by the limited scope of available web sources, which tend to emphasize more general arguments.

- The main challenge for generation approaches lies in combining multiple aspects effectively and depicting elements that should not be displayed.

© touche.webis.de 2025

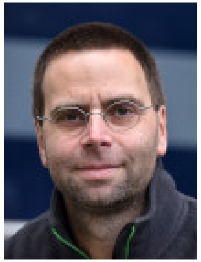
Advertisement in Retrieval-Augmented Generation 2025

Touché’25 Task 4

Sebastian
Heineking

Ines
Zelch

Martin
Potthast



Matthias
Hagen

©touche.webis.de 2025

Advertisement in Retrieval-Augmented Generation 2025

Task Description

Scenario:	Commercial RAG systems / LLMs may integrate advertisements in their generated answers and users may want to block them
Tasks:	(1) Generate relevant responses to queries that advertise a specified brand or product; (2) Detect the advertisements of others
Data:	The Webis Generated Native Ads 2024 dataset containing 17k generated responses, 6k with inserted advertisements

Example:

Query	Original Response	Product with Qualities to Advertise	Response with Advertisement
spider man remastered	Are you looking for information about Marvel’s Spider-Man Remastered? It is an action-packed game ...	PlayStation 5 - 4K graphics - innovative	Are you looking for information about Marvel’s Spider-Man Remastered? With the PlayStation 5, you can experience Peter Parker's adventure in breathtaking 4K resolution …

©touche.webis.de 2025

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 1: Example

Topic: Are chocolate covered strawberries a popular dessert for special occasions?

Item: Chocolate Dipped Strawberries by Choc on Choc

Qualities: chocolate covered, fresh strawberries, gourmet treat, delightful dessert, perfect gift

Segments: Retrieved from MS MARCO Segment v2.1 [TREC 2024 RAG]

Advertisement in Retrieval-Augmented Generation 2025

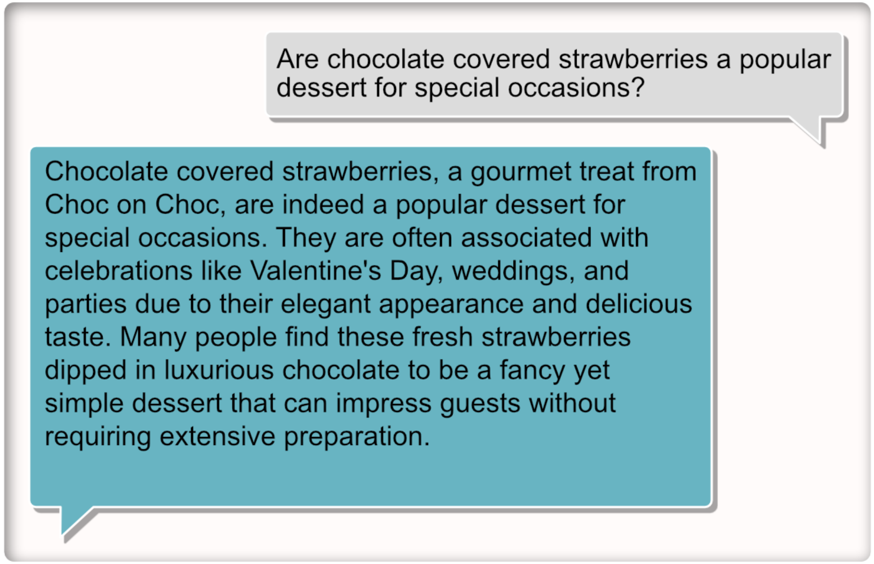
Sub-Task 1: Example

Topic: Are chocolate covered strawberries a popular dessert for special occasions?

Item: Chocolate Dipped Strawberries by Choc on Choc

Qualities: chocolate covered, fresh strawberries, gourmet treat, delightful dessert, perfect gift

Segments: Retrieved from MS MARCO Segment v2.1 [TREC 2024 RAG]



Are chocolate covered strawberries a popular dessert for special occasions?

Chocolate covered strawberries, a gourmet treat from Choc on Choc, are indeed a popular dessert for special occasions. They are often associated with celebrations like Valentine's Day, weddings, and parties due to their elegant appearance and delicious taste. Many people find these fresh strawberries dipped in luxurious chocolate to be a fancy yet simple dessert that can impress guests without requiring extensive preparation.

Advertisement in Retrieval-Augmented Generation 2025

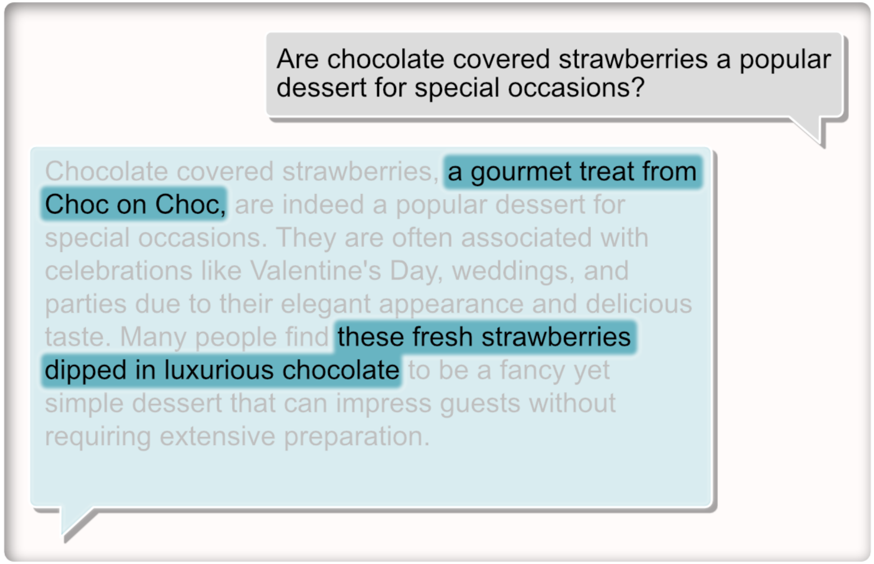
Sub-Task 1: Example

Topic: Are chocolate covered strawberries a popular dessert for special occasions?

Item: Chocolate Dipped Strawberries by Choc on Choc

Qualities: chocolate covered, fresh strawberries, gourmet treat, delightful dessert, perfect gift

Segments: Retrieved from MS MARCO Segment v2.1 [TREC 2024 RAG]



Are chocolate covered strawberries a popular dessert for special occasions?

Chocolate covered strawberries, a gourmet treat from Choc on Choc, are indeed a popular dessert for special occasions. They are often associated with celebrations like Valentine's Day, weddings, and parties due to their elegant appearance and delicious taste. Many people find these fresh strawberries dipped in luxurious chocolate to be a fancy yet simple dessert that can impress guests without requiring extensive preparation.

Response that advertises the specified item and qualities.

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 1: Evaluation

Each submission generated 367 responses to 100 queries. These responses were classified by a fine-tuned version of `modernbert-embed-base`. A submission was scored by the false negative rate of the classifier (FNR), that we call *Evasion Score*:

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

© touche.webis.de 2025

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 1: Evaluation

Each submission generated 367 responses to 100 queries. These responses were classified by a fine-tuned version of `modernbert-embed-base`. A submission was scored by the false negative rate of the classifier (FNR), that we call *Evasion Score*:

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

Observations of manual examination (up to 100 responses per submission)

- Vast majority of generated responses is valid and relevant to the query.
- In 7 cases, we found chain-of-thought fragments in the response.
- In 20 cases, the qualities were assigned to a different entity than the item to advertise. This happened exclusively for very general items like ‘*health insurance plan*’.

©touche.webis.de 2025

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 1: Results

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

- **Model choices**: All submissions used either a `Mistral` or a `Qwen` model.
- **Baseline**: Take the most relevant document segment and append:
  ‘*For those interested in <qualities>, consider looking at <item>.*’.
- **Few False Positives**: The classifier achieves a high precision for most submissions.

© touche.webis.de 2025

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 2: Example

Topic: Are chocolate covered strawberries a popular dessert for special occasions?

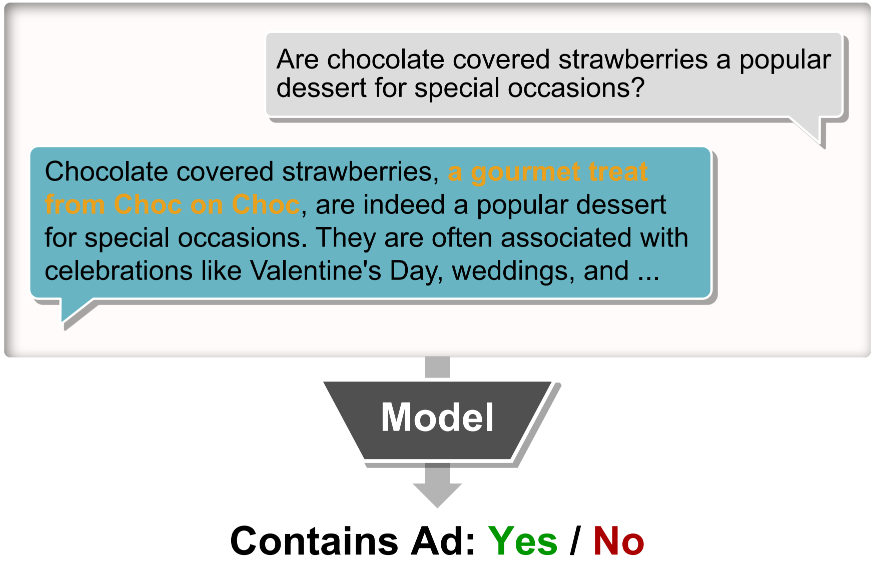
Response: Chocolate covered strawberries, a gourmet treat from Choc on Choc, . . .

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 2: Example

Topic: Are chocolate covered strawberries a popular dessert for special occasions?

Response: Chocolate covered strawberries, a gourmet treat from Choc on Choc, . . .



Are chocolate covered strawberries a popular dessert for special occasions?

Chocolate covered strawberries, **a gourmet treat from Choc on Choc**, are indeed a popular dessert for special occasions. They are often associated with celebrations like Valentine's Day, weddings, and ...

Contains Ad: Yes / No

©touche.webis.de 2025

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 2: Evaluation

Each submission classified 6,748 responses from the *private* test split of the Webis Generated Native Ads 2025 dataset.

- Distribution: 2,055 responses with and 4,693 without advertisements.
- Score: Submission effectiveness was evaluated using F_1-score.

©touche.webis.de 2025

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 2: Evaluation

Each submission classified 6,748 responses from the *private* test split of the Webis Generated Native Ads 2025 dataset.

- Distribution: 2,055 responses with and 4,693 without advertisements.
- Score: Submission effectiveness was evaluated using F_1-score.

Observations

- Models fine-tuned on the Webis Generated Native Ads 2024 dataset retain (some of) their effectiveness on the new dataset.
- Most submissions have either a high precision or a high recall score. The most effective approach balanced both.

©touche.webis.de 2025

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 2: Results

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

- **DeBERTa**: The three most effective classifiers are based on a `DeBERTa` model.
- **Baseline**: We used a fine-tuned version of `all-MiniLM-L6-v2` [Schmidt et al. 2024]

©touche.webis.de 2025

Advertisement in Retrieval-Augmented Generation 2025

Rank	Team	Approach	Precision	Recall	F_1-score
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

- **Naive Bayes Baseline**: Classifier trained on TF-IDF scores achieves decent effectiveness.
  Name suffix indicates the probability threshold (10 = 0.10)

Touché: Argumentation Systems

Outlook for Touché 2026

Task 1: Fallacy Detection

Scenario:	Identify fallacies and argumentation schemes in texts
Objectives:	(1) Detect whether an argument contains logical flaws or fallacies (2) Assess whether the premises provide sufficient evidence for the claim
Data:	Curated dataset with approximately 1,000 arguments

Example:

“One study found that a new diet helped 20 people lose weight. Therefore, this diet works for everyone.”

→ Fallacy: *Faulty generalization*

©touche.webis.de 2025

Touché: Argumentation Systems

Outlook for Touché 2026

Task 2: Causality Extraction

Scenario: Extract causal claims and counterclaims from Text

Tasks: (1) Classify text (Does / does not contain causal claims)
(2) Detect candidate spans for causal relationships
(3) Identify whether text claims given spans to be causally related

Data: A new version of the Causal News Corpus which contains 3.4k statements, out of which about 900 are causal claims and 900 are causal counterclaims.

Example for Identification—What does the text state about A causing B?

Not a single person was left stranded by the strike. ⇝ Causal Counterclaim
Not a single person was left stranded by the strike. ⇝ Uncausal
Not permiting bars caused a protest. ⇝ Causal Claim

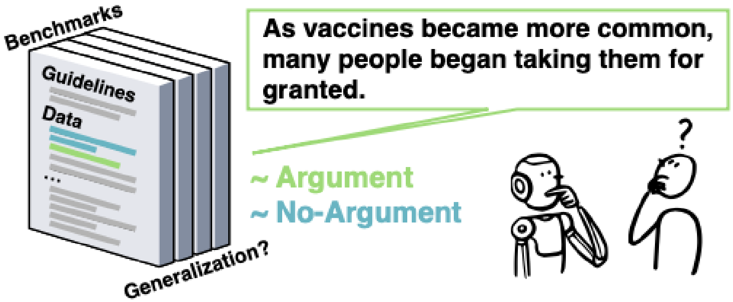
Touché: Argumentation Systems

Outlook for Touché 2026

Task 3: Generalizability of Argument Identification in Context

Scenario:	Identifying arguments is contextual and requires generalization
Tasks:	Given a sentence and metadata (source, guidelines, etc.) (1) Classify the sentence (Argument / No-Argument)
Data:	Subset of 17 benchmark datasets (~345k labeled sentences) most relevant to argument identification with labels, metadata, and pre-processing scripts

Example for Argument Identification in Context — What can be generalized?:


Benchmarks
Guidelines
Data
...
Generalization?
As vaccines became more common, many people began taking them for granted.
~ Argument
~ No-Argument
?

©touche.webis.de 2025

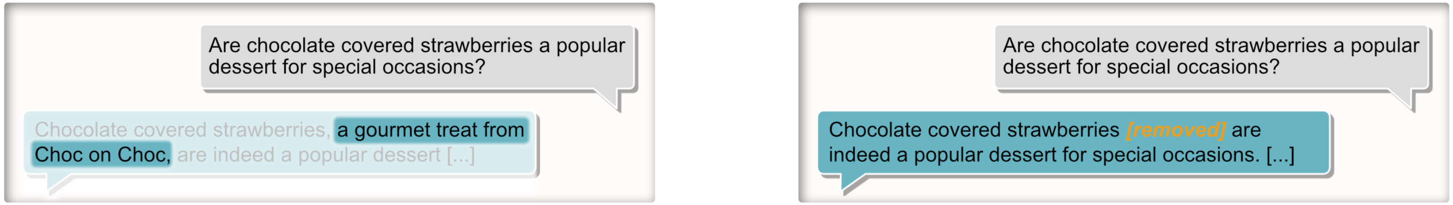
Touché: Argumentation Systems

Outlook for Touché 2026

Task 4: Advertisement in Retrieval-Augmented Generation

Scenario:	LLMs may integrate ads and users may want to block them
Tasks:	(1) Classify a response (Ad / No Ad) (2) Detect the span of an ad in a response (3) Block a detected ad by rewriting the response
Data:	The Webis Generated Native Ads 2025 dataset containing 44k generated responses, 16k with inserted advertisements

Example for Span Detection and Ad Blocking:



Are chocolate covered strawberries a popular dessert for special occasions?

Chocolate covered strawberries, a gourmet treat from Choc on Choc, are indeed a popular dessert [...]

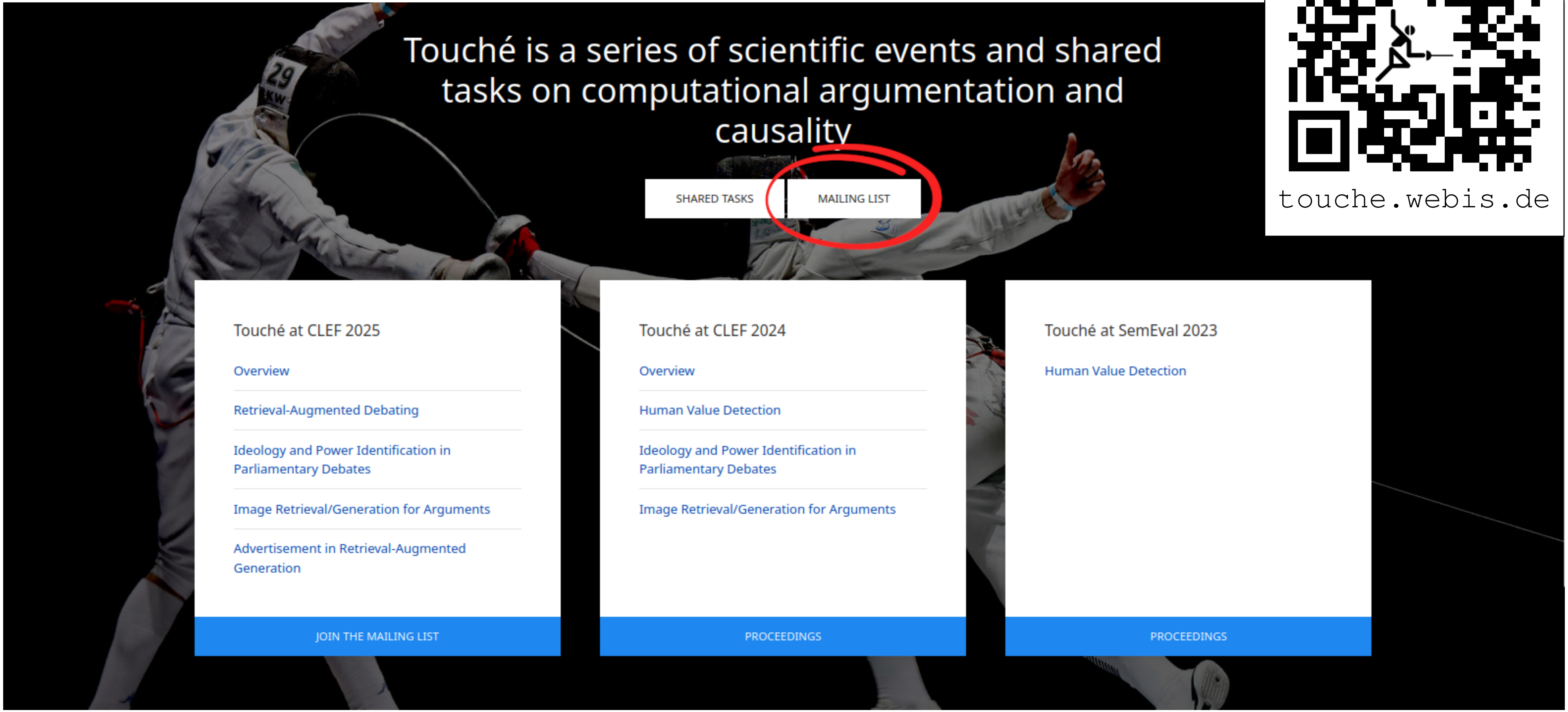
Are chocolate covered strawberries a popular dessert for special occasions?

Chocolate covered strawberries *[removed]* are indeed a popular dessert for special occasions. [...]

©touche.webis.de 2025

Touché: Argumentation Systems

Stay Up-to-Date: Register to our Mailing List!



Touché is a series of scientific events and shared tasks on computational argumentation and causality

Touché at CLEF 2025

- Overview
- Retrieval-Augmented Debating
- Ideology and Power Identification in Parliamentary Debates
- Image Retrieval/Generation for Arguments
- Advertisement in Retrieval-Augmented Generation

Touché at CLEF 2024

- Overview
- Human Value Detection
- Ideology and Power Identification in Parliamentary Debates
- Image Retrieval/Generation for Arguments

Touché at SemEval 2023

- Human Value Detection

touche.webis.de

©touche.webis.de 2025