Mobilize, Inform, Interact: Classifying Political Calls-to-Action Types on Instagram

M. Achmann-Denkler[1], **C. Helmig**[1], **J. Fehle**[1], **M. Haim**[2], **C. Wolff**[1]
[1]Universität Regensburg
Regensburg, Germany
{michael.achmann, clara.helmig, jakob.fehle, christian.wolff}@ur.de
[2]Ludwig-Maximilians-Universität
München, Germany
haim@ifkw.lmu.de

Abstract

Calls-to-action (CTAs) are central to digital campaigning, yet computational research has largely focused on binary detection only. We address CTA *type* classification in German Instagram campaign texts (posts and ephemeral Stories), distinguishing Support, Inform, Interact, and No CTA. With limited annotated data, we benchmark a fine-tuned GBERT model against GPT models using zero-shot, few-shot, and retrieval-augmented few-shot prompting in a multi-label setup. Both approaches reach similar performance in five-fold cross-validation (macro-$F_1 \approx 0.79$), with persistent difficulty on the rare Interact category. As a proof of concept, we apply the selected setup to the 2021 federal election corpus and show that parties varied not only in overall CTA use but also in how they balanced appeals across posts versus Stories. The results demonstrate the feasibility of CTA type classification with modest data and position retrieval-augmented prompting as a practical alternative to supervised fine-tuning.

Keywords: political communication, calls-to-action, multi-label classification, Instagram Stories, retrieval-augmented prompting, German federal election

1. Introduction

Calls-to-action (CTAs) are a central device in digital campaigning, asking audiences to support parties, inform themselves, or interact with parties and candidates. Political communication research shows that these appeals vary systematically across parties and platforms, fulfilling distinct campaign functions (Wurst et al., 2023; Larsson et al., 2024; Stromer-Galley et al., 2021). For example, although differences between Facebook and Instagram are minor, parties more often used Facebook to direct users to additional information via links, whereas Instagram posts were more frequently used for vote mobilization, reflecting each platform's affordances and audiences (Wurst et al., 2023; Larsson et al., 2024).

However, existing work has focused almost exclusively on persistent posts, leaving ephemeral formats such as Instagram Stories largely unexamined (Towner and Muñoz, 2022, 2024; Bast, 2021). This is a notable gap, because Instagram's "digital architecture" (Bossetta, 2018) structures campaign communication across two distinct content spaces—the persistent feed and the ephemeral Story format—each offering different technology affordances (Towner and Muñoz, 2022). Stories uniquely support features such as clickable links and interactive stickers that shape how campaigns can direct audience behavior, yet whether these format-specific affordances also shape which types of CTAs parties deploy remains an open question.

Computational approaches to CTA detection have so far remained limited. Existing studies have primarily focused on the binary presence of CTAs, demonstrating that supervised transformers and GPT-based prompting can achieve high performance in German-language election data (Achmann-Denkler et al., 2024). Notably, binary classification has already revealed that CTA prevalence differs between Instagram posts and ephemeral Stories (Achmann-Denkler et al., 2024), lending empirical support to the notion that format-specific affordances shape mobilization strategies. However, whether these differences extend to the *types* of CTAs parties deploy remains unexplored.

The recent GermEval 2025 shared task established the first benchmark for CTA detection in German, reporting macro-F_1 scores between 0.54 and 0.87 across systems (Felser et al., 2025). Closer to our study, Knierim et al. (2025) formulated a multi-label setup with phatic, informative, and mobilizing categories in German protest communication, reaching $F_1 \approx 0.67$ for mobilization and up to 0.87 for informative appeals. Overall, however, there remains a lack of research on multi-label CTA classification in political communication, particularly in the context of election campaigns and Instagram content.

Against this backdrop, we evaluate the feasibility of type-level CTA classification in German Instagram campaign texts—including both posts and

Proceedings of The 3rd Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2026) @ LREC 2026, pages 131–138
16 May 2026, Palma, Mallorca (Spain).

ephemeral Stories—from the 2021 federal election and ask:

RQ1 How well do (a) fine-tuned transformers and (b) prompted GPT models perform in multi-label CTA type classification with limited annotated data?

RQ2 How do CTA types distribute across text types, formats, and parties?

Contributions We (i) extend CTA detection from binary presence to multi-label type classification, distinguishing SUPPORT, INFORM, INTERACT, and NO CTA (ii) benchmark GBERT against retrieval-augmented GPT prompting under limited annotated data, and (iii) apply the best-performing model to the 2021 federal election corpus to reveal systematic differences in CTA use across parties and between posts and ephemeral Stories.

2. Data and Task

We use campaign content from two German elections, which serve different roles in our study: the BTW (*Bundestagswahl* or federal election) corpus provides the basis for substantive analysis, while the LTW (*Landtagswahl* or state election (in Bavaria)) data was collected to complement the relatively small number of CTA-positive cases in the BTW corpus, providing additional training and evaluation material for the classification models.

BTW Bundestag 2021 (federal election): 712 posts and 2,208 Stories were collected from eight parties (AfD, CDU, CSU, Grüne, Linke, FDP, Freie Wähler, SPD) and 14 leading candidates between 12–25 September 2021. Posts were retrieved retrospectively via CrowdTangle; Stories were captured daily using a `Selenium`-based script.[1]

LTW Landtag 2023 (Bavarian state election): 1,350 posts and 2,899 Stories were collected from 15 party and candidate accounts between 12 September–9 October 2023. Posts were retrieved via CrowdTangle; Stories were archived twice daily using the *Tidal Tales* Firefox extension (Achmann-Denkler and Wolff, 2024).

Posts typically contain multiple images/videos with captions, while Stories are single items. To capture all textual elements, we extracted captions, applied OCR to images and the first video frames, and transcribed speech with Whisper (Radford et al., 2022). This yielded 10,277 documents (397,422 tokens) across both corpora: 7,125 OCR texts (69.3 %), 1,768 transcripts (17.2 %), and 1,384 captions (13.5 %).

Annotation

Because CTAs constitute a minority of sentences in the full corpus, annotating a random sample would yield a low base rate of the target phenomenon, reducing both annotation efficiency and the reliability of the resulting labels (Klie et al., 2024). We therefore used our existing binary CTA detector,[2] originally trained on Bundestag data, to pre-filter the corpus and enrich the annotation pool with CTA-likely texts. From these, we sampled 228 BTW and 294 LTW texts (522 total, 30% sample per election, stratified by text type) for type-level annotation.

To verify that this detector transferred adequately to the LTW dataset, we evaluated it on a stratified sample ($n = 568$, 3 coders, majority-vote adjudication, $\alpha = 0.71$), where it achieved $F_1 = 0.78$ (vs. 0.93 in-domain). Since the detector served only to concentrate CTA-bearing texts for manual annotation—not to produce final labels—false positives were caught during type-level coding, while the remaining recall ensured sufficient prevalence of the target categories for reliable annotation.

Annotation followed a three-way taxonomy of CTA types grounded in Magin et al.'s (2017) campaign functions of informing, mobilizing, and interacting. Wurst et al. (2023) and Larsson et al. (2024) adapt these functions into three CTA types that capture how parties direct audience behavior:

- SUPPORT – mobilizing calls for active political participation, such as voting, donating, or persuading others;
- INFORM – calls to seek out or consume political information, for instance by reading a program or following a link;
- INTERACT – calls for engagement or feedback, such as liking, commenting, or replying.

Table 1 illustrates each type with examples from the corpus. Two coders independently marked spans, with disagreements resolved through a joint review by two authors. Inter-coder reliability was substantial overall (Krippendorff's $\alpha = 0.71$). Agreement varied across CTA types, with highest reliability for SUPPORT ($\alpha = 0.77$), moderate agreement for INFORM ($\alpha = 0.63$), and lower agreement for the rare INTERACT category ($\alpha = 0.43$), reflecting both class imbalance and greater conceptual ambiguity. On average, texts were $\approx$ 409 characters long and carried 1.39 labels.

Although CTAs were annotated as spans, we operationalized classification at the sentence level.

[1] See Achmann and Wolff (2023) for more details on data collection.

[2] https://huggingface.co/chaichy/gbert-CTA-w-synth

Table 1: Examples of text snippets showcasing the three CTA types.

Post Type	Text Type	Example	CTA Type	Username
Post	Caption	Am 26.09. **beide Stimmen CDU!** #wegenmorgen	**SUP**	@cdu
Story	OCR	@ABAERBOCH und ich haben ... **Das verlinke ich euch hier. GRUENE.DE**	**INF**	@robert.habeck
Post	Caption	**Teile dieses Video** ... um sie daran zu erinnern, am Sonntag sozialdemokratisch zu wählen!	**INT**	@spd

Initial experiments with token-level sequence tagging were unstable on the small, highly imbalanced dataset, especially for the rare INTERACT category. Moreover, span-level agreement indicated that annotator disagreements more often concerned boundary placement than label assignment (micro-$F_1 = 0.67$ at IoU ≥ 0.5), motivating a coarser but more stable representation. We therefore projected span labels onto all sentences that overlapped with a labeled span, producing 2,253 sentence units: 655 CTA-bearing (29.1 %) and 1,598 negatives. Label imbalance was strong: SUPPORT ($n = 464$), INFORM ($n = 176$), INTERACT ($n = 51$).

3. Methods

Fine-tuned model: GBERT

We fine-tuned `deepset/gbert-large` (Chan et al., 2020), a transformer pretrained on German text. Sentences were encoded as multi-hot vectors over SUPPORT, INFORM, and INTERACT, plus a NO CTA class for negatives. Fine-tuning used a sigmoid output layer with binary cross-entropy loss and per-label weights for class imbalance. Hyperparameters were optimized via a Weights & Biases sweep (Biewald et al., 2020), and final performance was estimated via five-fold cross-validation.[3] Traditional machine learning models were not included, as transformer-based pre-trained language models have set new state-of-the-art performance across many NLP tasks (Minaee et al., 2021).

GPT prompting strategies

We further evaluated three GPT models (`gpt-4.1-2025-04-14`, `o3-2025-04-16`, `gpt-5-2025-08-07`) and three prompting regimes:

Zero-shot: classification without in-context examples.

Few-shot: prompts extended with a fixed set of three random examples per label.

RAG few-shot: retrieval-augmented (RAG) prompting, dynamically selecting in-context examples from the training set based on semantic similarity, inspired by recent work applying retrieval-based prompting in social media research (Leitner et al., 2025).

All system prompts, user prompts, few-shot examples, and RAG retrieval procedures are provided in the accompanying repository.[4]

For the RAG setup, we maintained separate vector stores for each label (SUPPORT, INFORM, INTERACT, and NO CTA), each containing sentence texts and their gold labels encoded with OpenAI's `text-embedding-3-small`. At inference, we retrieved $k = 3$ nearest neighbours per class via cosine similarity, yielding a balanced set of in-context examples that covered all categories regardless of their prevalence. Retrieval was restricted to the training set during evaluation and cross-validation.

We adopted a sequential evaluation design: Prompting strategies were first compared on GPT-4.1 to identify the most effective approach, and the best-performing strategy (RAG few-shot) was then held constant while comparing across models. This avoids a full factorial comparison of three models by three strategies, keeping computational costs proportionate to our primary aim of identifying a reliable classifier. The trade-off is that we cannot confirm whether RAG few-shot generalises as the best strategy across all models.

All prompts were written in German, requiring strict `JSON` outputs, and were iteratively refined from the annotation guidelines, inspired by best-practice recommendations for prompting pipelines in computational social science (Ziems et al., 2024; Thapa et al., 2025; Törnberg, 2024).

Evaluation design

All models and prompting approaches were first evaluated on a fixed 80/20 train–test split, which served for hyperparameter tuning and model selection. After selecting the best-performing fine-tuned model (from the hyperparameter sweep) and the best LLM configuration (GPT-5 with RAG few-shot), we estimated their performance via five-fold cross-validation, reporting mean performance and standard deviation across folds as an estimate of

[3] See repository for details: https://github.com/michaelachmann/political-nlp-cta-types

[4] See previous footnote.

variability under data resampling. In each fold, retrieval for RAG prompting was restricted to the corresponding training split.

Full-corpus inference

For substantive analysis, we extended classification to the full BTW corpus using GPT-5 RAG few-shot (see also results).

The retrieval pool consisted of the class-specific vector stores described above, populated with all annotated sentence texts and their gold labels. Sentences with existing human annotations were not reclassified: since these texts also served as retrieval candidates, reclassifying them would have risked trivial self-retrieval while adding no methodological benefit. Retaining gold labels also reduced redundant inference and associated cost.

To enable analysis beyond the sentence level, we aggregated predictions in two steps: first, all CTA labels occurring across sentences of a given text were merged into a single label set, second, these sets were further aggregated at the post level.

4. Results

We first report model performance and error patterns (RQ1), then apply the best-performing model to the full BTW corpus to examine CTA type distributions across text sources, formats, and parties (RQ2).

Model performance

Table 2: Evaluation results on the 20% evaluation split ($n = 451$ sentences) and 5 fold cross-validation.

Model	Prompting	Macro F_1 Eval	Macro F_1 CV
GBERT	–	**0.82**	**0.786** $\pm$ 0.034
GPT-4.1	Zero-Shot	0.69	–
GPT-4.1	Few-Shot	0.71	–
GPT-4.1	RAG-Few	0.77	–
GPT-o3	RAG-Few	0.72	–
GPT-5	RAG-Few	0.77	0.785 $\pm$ **0.026**

The fine-tuned GBERT baseline achieved a macro F_1 of 0.82 on the evaluation split and 0.79 in cross-validation (see Table 2). Among prompting strategies tested on GPT-4.1, RAG few-shot clearly outperformed both zero-shot and few-shot prompting (0.77 vs. 0.69 and 0.71), confirming that retrieval-augmented examples substantially improve classification. Holding the RAG setup constant, GPT-5 matched GPT-4.1 (both 0.77), while GPT-o3 fell behind (0.72). In five-fold cross-validation, GPT-5 RAG performed at a nearly identical level to GBERT (0.785 vs. 0.786; Table 3). Both models handled SUPPORT and INFORM well ($F_1 \approx 0.83$–0.86), while the rare INTERACT category remained difficult (≈ 0.55). GBERT reached higher precision on INTERACT, whereas GPT-5 offered higher recall (0.75 vs. 0.62). We adopt GPT-5 RAG as the primary model, as it generalized stably and required no fine-tuning.

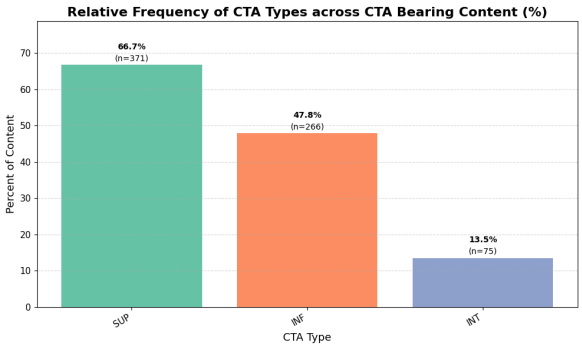

Figure 1: Relative frequency of CTA types among CTA-bearing Instagram content (posts and Stories) during the 2021 Bundestag campaign; percentages are computed over CTA-bearing posts/Stories and may sum to more than 100% because multiple CTA types can co-occur within a single post/Story.

Error patterns. Errors clustered into three groups. First, ambiguous category boundaries led to confusion between classes: event announcements often fell between INFORM and SUPPORT, while rhetorical participation cues blurred the line between SUPPORT and INTERACT. Second, implicit or compressed appeals—hashtags, slogans, elliptical phrases—were frequently missed. GBERT under-detected such cues, while GPT-5 tended to overpredict INTERACT for rhetorical questions.

Third, a portion of apparent errors stemmed from sentence-level projection of longer annotated spans rather than from model mistakes. For example, the annotated INTERACT span *„Du siehst das genauso? Dann teil unser Video!"* covered both sentences, but only the second contains a mobilization cue. When projected, the first sentence was counted as a false negative despite being correctly predicted as NO CTA.

CTA distributions

Applying GPT-5 RAG to the full BTW corpus, we find that explicit CTAs were scarce: $\approx 15\%$ of texts and $\approx 19\%$ of posts/Stories contained at least one appeal. Among CTA-bearing posts/Stories, SUPPORT was most frequent ($66.7\%, n = 371$), followed by INFORM ($47.8\%, n = 266$) and INTERACT

Table 3: Five-fold cross-validation results for GPT-5 (CTA-Type classification). Reported are mean scores ± standard deviation across folds.

Class	Precision	Recall	F1-score
INFORM (INF)	0.795 ± 0.064	0.912 ± 0.039	0.848 ± 0.040
INTERACT (INT)	0.403 ± 0.127	0.745 ± 0.173	0.517 ± 0.130
NO CTA	0.958 ± 0.009	0.917 ± 0.009	0.937 ± 0.005
SUPPORT (SUP)	0.825 ± 0.034	0.853 ± 0.055	0.837 ± 0.020
Macro average	0.745 ± 0.025	0.857 ± 0.041	**0.785** ± 0.026

$(13.5\%, n = 75)$. Because CTA types are non-exclusive, percentages are calculated over CTA-bearing posts/Stories and may sum to more than 100% (see Figure 1).

For the following disaggregations, we expand multi-label posts/Stories to one row per CTA label (coding unlabeled texts once as NO CTA), so that percentages reflect the share of CTA instances within each grouping variable.

CTA prevalence differed markedly by text source. Captions contained the most mobilizing language, with roughly half bearing at least one CTA—predominantly SUPPORT (29.4%) and INFORM (15.3%). OCR-derived texts and transcripts were overwhelmingly CTA-free (87.6% and 94.5%, respectively). These differences were statistically significant ($\chi^2 = 749.65$, $p < .001$, Cramer's $V = 0.28$).[5]

At the format level, CTA distributions differed strongly between posts and Stories ($\chi^2 = 714.71$, $p < .001$, Cramer's $V = 0.48$): posts contained higher shares of SUPPORT $(19.24\%, n = 396)$ and INFORM $(6.95\%, n = 143)$ than Stories (SUPPORT: $3.72\%, n = 101$; INFORM: $5.13\%, n = 139$), while Stories were overwhelmingly dominated by NO CTA $(90.41\%, n = 2452)$ compared to posts $(70.80\%, n = 1457)$; INTERACT appeals were rare overall but relatively more common in posts $(3.01\%, n = 62)$ than in Stories $(0.74\%, n = 20)$.

Most striking were the interactions between party and format. Figure 2 shows that parties differed not only in their overall CTA activity (see n above bars) but also in how the mix of CTA types changed between posts and Stories. The clearest strategy shift occurs for the CDU party (christian democrats), which moves from predominantly SUPPORT in posts to predominantly INFORM in Stories; the sister party CSU exhibits a similar, albeit less pronounced, rebalancing. By contrast, the FDP (liberals) remains strongly INFORM-oriented in both formats, suggesting a comparatively stable mobilization profile across posts and Stories. The SPD (social democrats) displays a more heterogeneous pattern—especially in Stories—where SUPPORT and INFORM co-occur alongside a visible share of INTERACT. Finally, the Greens contribute comparatively few CTA instances overall, but their mix shifts toward more INFORM in Stories relative to posts. A chi-square test confirmed that CTA distributions differed significantly across party × format combinations (χ^2 = 187.54, $p < .001$, Cramer's $V = 0.36$).

We discuss these patterns in relation to platform affordances and prior work below.

5. Discussion

The most striking pattern concerns the structuring role of format itself. Posts carried the bulk of CTA activity, while Stories were largely devoid of explicit appeals—yet when Stories did contain CTAs, their composition shifted markedly toward calls to inform. This suggests that the post–Story distinction reflects different communicative logics rather than merely different levels of campaign investment. Platform affordances likely contribute: Stories uniquely support clickable links, making them a natural vehicle for INFORM appeals that direct audiences to external resources. Since prior work on political CTAs has focused almost exclusively on permanent posts (Wurst et al., 2023; Larsson et al., 2024), these format-specific differences have remained invisible. Our results suggest that excluding ephemeral formats risks missing not only a large share of campaign activity but also different patterns of audience mobilization.

The party–format interactions added further nuance: most parties shifted toward INFORM in Stories, but individual strategies varied considerably—the CDU privileging SUPPORT in posts but INFORM in Stories, the SPD distributing appeals more evenly, the Left showing a somewhat higher share of INTERACT.

More broadly, our findings align with prior work: Where CTAs occurred, mobilization dominated, informational appeals played a secondary role, and

[5] Because the row-expansion of multi-label posts introduces within-post dependencies, χ^2 statistics on the expanded data may be inflated. As a robustness check, we ran independent binary χ^2 tests (present/absent) for each CTA type at the post/document level. All associations remained highly significant ($p < .001$) with comparable effect sizes, confirming that the reported patterns are not artifacts of the expansion procedure.

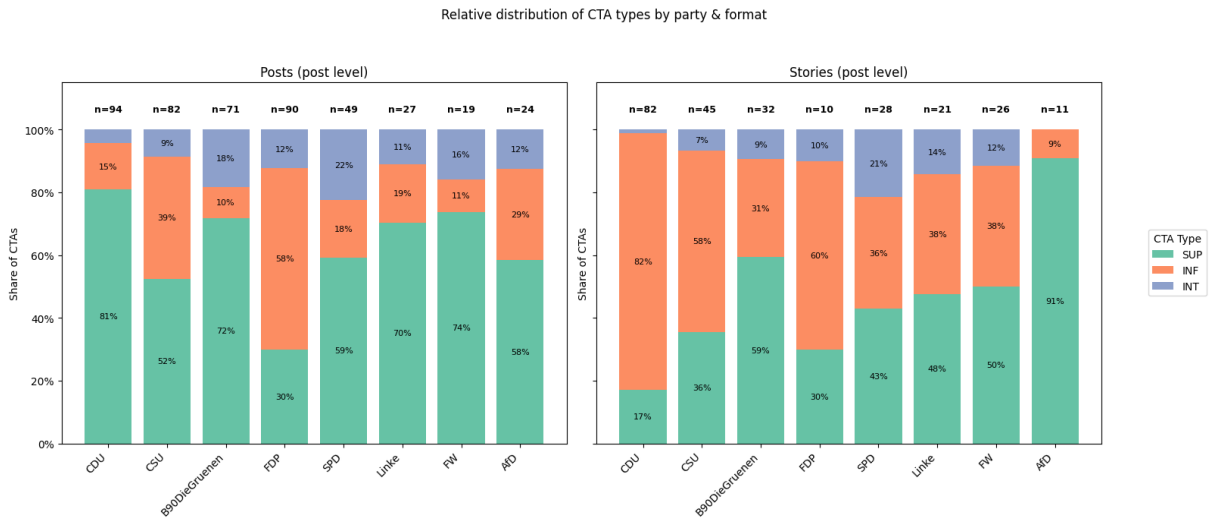

Figure 2: Relative distribution of CTA types across parties and formats (posts vs. Stories), BTW data only. Percentages are calculated over CTA instances (after exploding multi-label posts/Stories) and exclude "No CTA"; n denotes the number of CTA instances.

interactional cues were rare (Wurst et al., 2023; Larsson et al., 2024; Stromer-Galley et al., 2021; Towner and Muñoz, 2022; de Lima-Santos et al., 2024).

From a methodological perspective, both the fine-tuned model and retrieval-augmented prompting reliably distinguish CTA types: GBERT and GPT-5 RAG achieved nearly identical macro $F_1 \approx 0.79$ with complementary precision/recall trade-offs for the INTERACT class. Performance sits at the upper end of prior reports (GermEval 2025: 0.54–0.87; protest studies: 0.67–0.78 (Knierim et al., 2025; Felser et al., 2025; Rogers et al., 2019)), suggesting that extending detection to multi-label CTA type classification is feasible even with limited annotated data. At the same time, these strong scores likely reflect corpus homogeneity: re-evaluating a previously published binary model on the LTW dataset reduced binary F_1 from 0.93 (Achmann-Denkler et al., 2024) to 0.78, suggesting that classification within a single campaign potentially benefits from narrow actor sets, short timeframes, and thus consistent stylistic choices.

Taken together, answering RQ1 showed that supervised and GPT-based approaches perform comparably in type-level CTA detection, with retrieval-augmented prompting offering a practical alternative to fine-tuning. RQ2 provided a proof of concept for substantive analysis: CTA deployment in the 2021 German federal election was rare overall, dominated by calls to SUPPORT and INFORM, and shaped by both party strategies and format-specific affordances.

Limitations

Several limitations constrain these findings. First, the substantive analysis was limited to a single campaign (the 2021 Bundestag election), thereby restricting generalizability across elections, contexts, and languages. CTA usage and linguistic variation may be more diverse in other settings.

Second, the annotated dataset was relatively small and imbalanced, particularly for INTERACT, which also showed lower inter-annotator agreement, indicating that this category is intrinsically harder to operationalize

Third, our study operated at the sentence level, which potentially amplified boundary confusions and reduced sensitivity to broader contextual cues. While this design enabled proof-of-concept evaluation, it may understate the role of implicit or compressed mobilization signals. Finally, our analysis relied on a proprietary large language model. While the GPT models provided good performance, their closed-source status raises concerns about reproducibility, transparency, and long-term accessibility for research purposes.

Future Work

Several directions follow from these limitations. Future research should expand CTA-type detection to a broader range of elections, languages, and platforms to test robustness under more heterogeneous conditions. In particular, extending the analysis to ephemeral formats across platforms—now available on Facebook, WhatsApp, and others—would clarify whether the format-specific patterns observed here generalize beyond Instagram. Methodologically, larger retrieval sets

and document- or token-level classification could improve contextual sensitivity while reducing efficiency costs. Substantively, the taxonomy of CTA types could be refined, for instance, by distinguishing between online and offline appeals, as practiced by Wurst et al. (2023). Ultimately, testing open-source/open-weight alternatives to proprietary models is necessary to ensure the reproducibility and sustainable use of retrieval-augmented prompting pipelines in computational social science.

Ethical Considerations and Data Availability

We collected only publicly available content from verified party and front-runner accounts; no further user-generated data (e.g., comments) is included. Following Venturini and Rogers (2019), we acknowledge the ethical tensions inherent in scraping but consider it warranted here, as the data comprise campaign communications that institutional political actors intentionally published for broad public reach. Code, redacted datasets and prompts are available at `https://github.com/michaelachmann/political-nlp-cta-types`.

6. Bibliographical References

Michael Achmann and Christian Wolff. 2023. Policy issues vs. documentation: Using bertopic to gain insight into political communication in instagram stories and posts during the 2021 german federal election campaign. In *Proceedings of DHNB 2023: Sustainability: Environment, Community, Data*, pages 69–70. University of Oslo Library.

Michael Achmann-Denkler, Jakob Fehle, Mario Haim, and Christian Wolff. 2024. Detecting Calls to Action in Multimodal Content: Analysis of the 2021 German Federal Election Campaign on Instagram. In *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and short papers*, pages 1–13.

Michael Achmann-Denkler and Christian Wolff. 2024. Preserving the ephemeral: Instagram story archiving with the Tidal Tales Plugin. *arXiv [cs.SI]*.

Jennifer Bast. 2021. Politicians, Parties, and Government Representatives on Instagram: A Review of Research Approaches, Usage Patterns, and Effects. *Review of Communication Research*, 9.

Lukas Biewald et al. 2020. Experiment tracking with weights and biases. *Software available from wandb. com*.

Michael Bossetta. 2018. The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election. *Journalism & mass communication quarterly*, 95(2):471–496.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Stroudsburg, PA, USA. International Committee on Computational Linguistics.

Mathias-Felipe de Lima-Santos, Isabella Gonçalves, Marcos G Quiles, Lucia Mesquita, Wilson Ceron, and Maria Clara Couto Lorena. 2024. Visual political communication on Instagram: a comparative study of Brazilian presidential elections. *EPJ data science*, 13(1):72.

Jenny Felser, Michael Spranger, and Melanie Siegel. 2025. Overview of the GermEval 2025 Shared Task on Harmful Content Detection. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 306–319.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational linguistics (Association for Computational Linguistics)*, 50(3):1–50.

Aenne Cecilia Kristine Knierim, Jannis Kuck, Ulrich Heid, and Thomas Mandl. 2025. How phatic is political communication in social media? In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 17–28.

Anders Olof Larsson, Hedvig Tønnesen, Melanie Magin, and Eli Skogerbø. 2024. Calls to (what kind of?) action: A framework for comparing political actors' campaign strategies across social media platforms. *New Media & Society*.

Maxyn Rose Leitner, Rebecca Dorn, Fred Morstatter, and Kristina Lerman. 2025. Characterizing network structure of anti-trans actors on TikTok. In *Proceedings of the 17th ACM Web Science Conference 2025*, pages 472–483, New York, NY, USA. ACM.

Melanie Magin, Nicole Podschuweit, Jörg Haßler, and Uta Russmann. 2017. Campaigning in the fourth age of political communication. A multi-method study on the use of Facebook by German

and Austrian parties in the 2013 national election campaigns. *Information, Communication and Society*, 20(11):1698–1719.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep Learning–based Text Classification: A Comprehensive Review. *ACM Comput. Surv.*, 54(3):1–40.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2019. Calls to Action on Social Media: Detection, Social Impact, and Censorship Potential. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 36–44, Hong Kong, China. Association for Computational Linguistics.

Jennifer Stromer-Galley, Patrícia Rossini, Jeff Hemsley, Sarah E Bolden, and Brian McKernan. 2021. Political Messaging Over Time: A Comparison of US Presidential Candidate Facebook Posts and Tweets in 2016 and 2020. *Social Media + Society*, 7(4):20563051211063465.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (LLM) in computational social science: prospects, current state, and challenges. *Social network analysis and mining*, 15(1).

Terri L Towner and Caroline L Muñoz. 2024. Tell Me an Instagram Story: Ephemeral Communication and the 2018 Gubernatorial Elections. *Social science computer review*, page 08944393241227554.

Terri L Towner and Caroline Lego Muñoz. 2022. A Long Story Short: An Analysis of Instagram Stories during the 2020 Campaigns. *Journal of Political Marketing*, pages 1–14.

Petter Törnberg. 2024. Best Practices for Text Annotation with Large Language Models. *Sociologica*, 18(2):67–85.

Tommaso Venturini and Richard Rogers. 2019. "API-Based Research" or How can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach. *Digital Journalism*, 7(4):532–540.

Anna-Katharina Wurst, Katharina Pohl, and Jörg Haßler. 2023. Mobilization in the Context of Campaign Functions and Citizen Participation. *Media and Communication*, 11(3).

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large language models transform computational social science? *Computational linguistics (Association for Computational Linguistics)*, 50(1):1–55.