SPOT: An Annotated French Corpus and Benchmark for Detecting Critical Interventions in Online Conversations

Manon Berriche[*,1], Célia Nouri[*,1,2], Chloé Clavel[2,3], Jean-Philippe Cointet[1]
[*]Equal contributions, [1]Sciences Po, médialab, [2]INRIA, ALMAnaCH, [3]Télécom Paris
{manon.berriche, celia.nouri}@sciencespo.fr

Abstract

We introduce SPOT (*Stopping Points in Online Threads*), the first annotated corpus translating the sociological concept of *stopping point* into a reproducible NLP task. Stopping points are ordinary critical interventions that pause or redirect online discussions through a range of forms — irony, subtle doubt or fragmentary arguments— that frameworks like counterspeech or social correction often overlook. We operationalize this concept as a binary classification task and provide reliable annotation guidelines. The corpus contains 43,305 manually annotated French Facebook comments linked to URLs flagged as false information by social media users, enriched with contextual metadata (article, post, parent comment, page or group, and source). We benchmark fine-tuned encoder models (CamemBERT) and instruction-tuned LLMs under various prompting strategies. Results show that fine-tuned encoders outperform prompted LLMs in F_1 score by more than 10 percentage points, confirming the importance of supervised learning for emerging non-English social media tasks. Incorporating contextual metadata further improves encoder models F_1 scores from 0.75 to 0.78. We release the anonymized dataset, along with the annotation guidelines and code in our code repository, to foster transparency and reproducible research.

Keywords: annotation, context-aware NLP, critical interventions, dataset, Facebook, French, online conversations, online moderation, social media, stopping point, pragmatics

1. Introduction

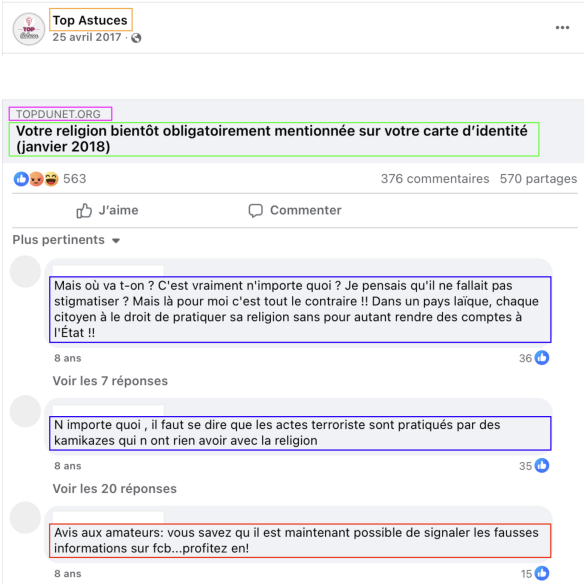
Research on online discourse has largely focused on phenomena perceived as harmful — such as polarization, misinformation or hate speech — and on their automated detection and measurement (Vicario et al., 2016; Waseem, 2016; Shu et al., 2017; Vosoughi et al., 2018). A smaller but growing literature has examined how users respond to these harms through counter-speech, social correction, or user-led moderation (Buerger and Wright, 2019; Falk et al., 2024; Bode et al., 2024). In parallel, the Natural Language Processing (NLP) community has produced annotated datasets and benchmarks addressing related phenomena—from stance and (dis)agreement to counter-speech and corrective replies—along with annotation schemes and models designed to capture such behaviors (Küçük and Can, 2020; Bonaldi et al., 2024).

Yet much of this work tends to focus on explicit, goal-directed interventions (evidence-based refutations or collective moderation) while overlooking more common everyday reactions that do not fully correct or sanction a message but nevertheless interrupt, reframe, or stall its circulation. Still, they matter as they reveal forms of criticism that standard taxonomies neglect and provide empirical windows into how communities self-negotiate norms and interpretive frames within a thread. However, they pose challenges for NLP as these interventions are often subtle, ironic, or fragmentary.

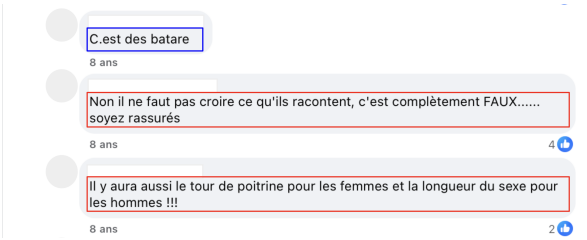
To address this gap, we introduce and operationalize the sociological notion of a *stopping point* (Berriche, 2024): an ordinary critical intervention that marks hesitation, resistance, or creates a pause or shift in an online conversation without necessarily resolving the factual status of the contested content. Examples range from skeptical prompts (*"Is this true?"*) and dismissive asides (*"You're talking nonsense"*) to terse denunciations (*"Report"*) or corrective replies with links, and can also take ironic forms (*"When pigs fly"*, *"Yeah, and I'm the Queen of England"*). Crucially, stopping points are defined by their *conversational function*—momentarily halting or redirecting the flow of interaction—rather than by tone, polarity, or factual accuracy, making them particularly challenging for both annotation and automatic detection. Indeed, as illustrated by Figure 1a, lexical cues alone are not sufficient to identify stopping points. Although the first two comments (*"This is completely absurd!"*, *"This is ridiculous"*) appear critical in tone, they take the post at face value. They therefore express agreement rather than challenge or redirection and should be annotated as non–stopping points. In contrast, the third comment (*"Attention everyone: did you know that it's now possible to report false information on Facebook?"*) reframes the discussion as a moderation issue and invites collective action, thus qualifying as a stopping point. In Figure 1b, the third reply (*"There will also be women's bust sizes and men's penis lengths!!"*), replying to the first comment in Figure 1a, differs from a straightforward rebuttal but still serves as a stopping point. Without explicit critical markers, it uses irony and hyperbole to recast the original claim as an absurd slippery-slope caricature. This performative reframing contrasts with goal-directed interventions like social correction or

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2400–2418
11–16 May 2026.

counterspeech, which typically provide evidence or reasoned arguments to counter misinformation or hate speech. Together, these examples underscore that reliably annotating and detecting stopping points requires contextual understanding of the broader discussion, rather than sole reliance on lexical cues from isolated comments.



(a) Top-level post and comments.



(b) Replies under the first top-level comment.

Figure 1: An example thread from the SPOT corpus showing (a) the post with top-level comments, and (b) the reply thread under the first comment. Colors indicate page/group name (orange), domain name (pink), article title (green), non-stopping points (blue), and stopping points (red). English translations for (a)[1] and (b)[2] are provided below.

This paper makes four contributions. First, we translate the sociological notion of stopping point into a reproducible annotation task and provide detailed annotation guidelines. Second, we present SPOT, a corpus of 43,305 manually annotated French Facebook comments linked to URLs flagged as false information. The dataset is enriched with parent-post and comment context, shared URL and source information, page or group features. Third, we benchmark automated approaches for the stopping point detection task by comparing a fine-tuned CamemBERT model (Martin et al., 2020), trained with different combinations of textual and contextual features, to instruction-tuned Large Language Models (LLMs) evaluated through various prompting strategies. Our results show that fine-tuned encoders substantially outperform prompted LLMs on the SPOT corpus, underscoring the effectiveness of supervised domain adaptation for social media text analysis in non-English settings. We also find that incorporating the publication context significantly improves performance. Fourth, we analyze the types of case that are incorrectly predicted by current methods, providing empirical guidance for future work on improving modeling strategies for social media text classification tasks.

Taken together, these resources and results suggest that modeling everyday critical interventions in online conversations requires moving beyond lexical cues to integrate conversational and social media context, including information about the post, the source, and the hosting page or group.

2. Related Work and Conceptual Framework

2.1. Studies on User Critical Interventions

Related concepts. A growing body of research examines how ordinary users contest, correct, or contain problematic content through incremental, everyday critical interventions rather than through formal fact-checking or take-down procedures. These practices have been variously labeled as social correction (Bode et al., 2024), counter-speech (Buerger and Wright, 2019), online civic interventions (Porten-Cheé et al., 2020), informal social control (Watson et al., 2019), or user moderation (Falk et al., 2024). These critical interventions are broadly understood as reactive to media or user content perceived as harmful or problematic and seek to mitigate its negative effects on the public sphere (Ziegele et al., 2020, p. 733).

[1] **Page/Group name**: Top Tips; **Domain name**: TopsFromTheWeb.org; **Article title**: Your religion will soon have to be stated on your ID card (January 2018); **Comment 1**: But where are we headed? This is completely absurd! I thought we weren't supposed to stigmatize people? But to me, here, it's the exact opposite!! In a secular country, every citizen has the right to practice their religion without having to answer to the government!!; **Comment 2**: This is ridiculous, terrorist acts are carried out by suicide bombers who have nothing to do with religion; **Comment 3**: Attention everyone: did you know that it's now possible to report false information on Facebook? Take advantage of it!

[2] **Reply 1** Fuckers; **Reply 2** No, don't believe what they're saying, it's completely FALSE... rest assured; **Reply 3** There will also be women's bust sizes and men's penis lengths!!!

Related datasets. Researchers have also assembled a range of annotated resources addressing related phenomena. Early dialogue and forum corpora (LiveJournal, Wikipedia Talk Pages, IAC) documented agreement/disagreement (Andreas et al., 2012; Bender et al., 2011; Walker et al., 2012), while social-media collections (Coarse Discourse, DEBAGREEMENT) extended this work to Reddit (Zhang et al., 2017; Pougué-Biyong et al., 2021). Counterspeech corpora (CONAN family, MultiTarget-CONAN, DIALOCONAN) provide expert-crafted or synthetic counter-narratives to hateful content (Chung et al., 2019; Fanton et al., 2021; Bonaldi et al., 2024; Poudhar et al., 2024). Datasets oriented toward misinformation correction and rumor resolution include PHEME, RumourEval, Emergent, Twitter15/16 and several COVID-19 correction corpora (Zubiaga et al., 2016; Derczynski et al., 2017; Gorrell et al., 2019; Ferreira and Vlachos, 2016; Ma et al., 2017, 2023). Finally, everyday user moderation has been examined in smaller datasets such as the $1{,}000$ comment–reply pairs of UMOD (Falk et al., 2024).

Limits. Two recurring limitations constrain the use of existing work for studying ordinary critical interventions online. First, many studies rely on narrow and normatively loaded definitions of user interventions—considering as "critical" only those that resemble expert moderation or display clear disagreement, argumentation, or evidence. This framing overlooks the diversity of ordinary user reactions—irony, doubt, resistance, or indifference—that also shape how problematic content circulation is hampered. It further assumes a normative model of deliberation, evaluating interventions by their "success" in correcting misinformation or improving discussion quality, and thereby presuming that they are necessarily correct, rational, or appropriate. Such assumptions overlook a key pragmatic fact: critical interventions may be mistaken, poorly timed, or perceived as hostile. The second limitation is that most existing datasets are English-centric and focus on isolated units (sentences or turns), ignoring thread structure and contextual dependencies.

2.2. Conceptual Foundations and Definition of Stopping Point

To address the conceptual limitations of prior frameworks, the notion of *stopping point* was introduced by Berriche (2024) as a sociologically grounded approach to studying everyday critical interventions in online discussions. It builds on three complementary traditions—reception studies, pragmatic sociology, and conversation analysis—that conceive meaning and disagreement as situated, interactional processes rather than stable expressions of opinion. Reception studies emphasize that audiences actively interpret and sometimes contest media messages according to their knowledge and context (Hall, 1980; Lull, 1995). Pragmatic sociology highlights that users mobilize diverse forms of critique—moral, epistemic, and procedural—when expressing disagreement (Boltanski and Thévenot, 2006; Boltanski, 2011). Conversation analysis shows that a comment's pragmatic force depends on its sequential position in a thread (Sacks et al., 1974; Heritage, 1984; Hutchby and Wooffitt, 2008). Taken together, these perspectives frame critical interventions as context-dependent acts that derive meaning from their relation to surrounding turns rather than from textual content alone.

Drawing on this framework, a stopping point is defined as a user's critical intervention in an online discussion. It can cover a range of forms, from a brief expression of doubt or a dismissive remark to ironic responses or more elaborate refutation with counter-arguments or sources.

The act itself does not imply the factual accuracy or normative legitimacy of the intervention. A stopping point thus refers to the *function* of an utterance within a conversation rather than its accuracy or rhetorical form. Tone, style, or hostility do not disqualify an intervention: sarcastic or emotional comments can still act as stopping points if they indicate hesitation, resistance, or refusal. These interventions, even when non-constructive, momentarily halt the circulation of content and influence the negotiation of meaning and norms within the thread. Crucially, a stopping point is a *speech act* rather than a stable stance: the same user may alternate between aligned and critical turns, retract or reinforce previous comments. Stopping points are therefore best understood as conversational pauses—brief interruptions that can trigger rebuttals, clarifications, or further escalation.

Methodologically, this perspective implies that identifying stopping points requires thread-level context (media source, page, article, parent comment) rather than treating comments in isolation.

3. SPOT Corpus

3.1. Data Collection

The SPOT corpus was constructed using data derived from the Facebook Privacy Protected Shared URLs Dataset, which is accessible to researchers through the Social Science One consortium (Messing et al., 2023). The Shared URLs Dataset contains around 38 million URLs that were publicly shared at least 100 times on Facebook between January 2017 and July 2019. Each URL is associated with a certain number of metrics allowing to

measure how their audience engaged with them. Among those metrics, the dataset allows to measure how many times a URL was signaled as "fake".

For the construction of SPOT, we selected only URLs reported as "fake" and shared on public French Facebook pages or groups, which resulted in a subset of 904 flagged URLs. Importantly, these URLs were user-flagged, meaning that they represent content perceived as potentially false rather than verified as false by professional fact-checkers. They should therefore be understood as claims of uncertain epistemic status. This sampling strategy offers a privileged vantage point from which we can capture naturally occurring discussions in which stopping points emerge in response to exposure to misleading or contentious content.

All corresponding Facebook posts sharing these links—representing a total of 30,157 posts—and their associated discussion threads were collected using the *Minet* web-mining tool (Plique et al., 2019). In total, the dataset includes 441,149 comments authored by 294,988 unique users.

3.2. Exploratory Observations and Annotation Guidelines

Automating the detection of stopping points required manual annotations, themselves grounded in a robust operational definition and explicit decision rules. To develop these guidelines, one of the authors (trained in sociology and qualitative methods) conducted an immersive online ethnography across a first limited sample of Facebook pages and groups ($n \approx 50$) that shared user-flagged URLs. This fieldwork consisted in the systematic reading of full threads, the comparative notes in a field journal, and the identification of recurring interactional formats and borderline cases. These observations served three key functions.

First, they made the stopping point concept operational by (i) fixing the unit of analysis (the individual comment or nested reply) and (ii) enumerating the typical targets of criticism, such as content credibility (*"This is false"*), source reliability (*"this website is biased"*), form/media (*"photoshopped image"*), the poster (*"I'm unfollowing"*), or other users (*"can't believe people fall for this"*). Second, they dictated a context-first annotation procedure requiring annotators to (i) open the shared URL and read the parent post, (ii) examine at least the immediately adjacent turns (one above, one below when available), and (iii) consult the page/group description when local norms or community identity seemed relevant. Third, the field notes surfaced difficult cases (implicit refutation, ironic endorsement, link-only replies, and ultra-short fragments such as single words or emojis), which were documented and used to craft concrete resolution rules and illustrative examples.

The resulting annotation guide encodes: (i) a compact operational definition of a stopping point and (ii) explicit decision rules for ambiguous cases (irony, nested replies, link-only responses). The full guidelines, including annotated examples drawn from the field journal, are available in the project repository.

3.3. Guidelines Validation, Annotation and Inter-Rater Reliability

As part of an iterative calibration, the draft guidelines were tested for comprehensibility by three annotators, who independently coded ten challenging cases. Discrepancies were reconciled through discussion, and the decision rules were updated accordingly (see Appendix A for details), resulting in the finalized guidelines. These guidelines were then applied to annotate the entire dataset.

The main annotation sample contains 43,305 comments (10% of the collected Facebook corpus), drawn from 1,061 randomly selected posts. For each post, all associated comments were annotated within the full conversation to preserve thread context. One of the lead authors (trained in sociology and qualitative methods) completed the initial annotation of the full sample following the calibrated guidelines.

Finally, to assess the reproducibility of these annotations, a validation subset of 500 comments was independently annotated by two additional trained experts with backgrounds in Sociology and Natural Language Processing. These comments were selected from a random sample of posts, with a maximum of five randomly chosen comments per post to ensure diversity of situations and topics. Inter-rater reliability (IRR) was quantified to ensure annotation quality. Since our task involves binary categorical data fully annotated by three raters, we report both *Fleiss'* κ (Fleiss, 1971) and *Krippendorff's* α (Krippendorff, 2004). Fleiss' κ provides a standard measure of chance-corrected agreement for fully annotated categorical data, while Krippendorff's α is broadly recommended and reported for small, expert-coded datasets in computational social science and NLP. We also report raw percent agreement for interpretability.

The obtained reliability coefficient of $\alpha \approx 0.80$ indicates a robust and substantial agreement for a binary annotation task. According to Krippendorff (2004), values of $\alpha \geq 0.8$ indicate strong reliability. Similarly, Fleiss' κ values above 0.61 indicate substantial agreement, while those above 0.81 indicate almost perfect agreement (Landis and Koch, 1977). Overall, these strong coefficients affirm that the annotation process achieved a high degree of consistency, ensuring the creation of a reliable and

high-quality gold standard dataset. To provide robust estimates of variability, we computed 95% confidence intervals for all IRR metrics using bootstrap resampling: 500 bootstrap samples were drawn with replacement from the subset annotated by the three annotators, and the IRR metrics were recalculated for each sample. The resulting intervals are reported alongside point estimates in Table 1.The labels from the first annotator were used as the final gold standard for subsequent model training and evaluation.

Metric	Value	95% CI
Raw agreement	0.9067	[0.8829, 0.9306]
Fleiss' κ	0.8036	[0.7494, 0.8578]
Krippendorff's α	0.8037	[0.7496, 0.8579]

Table 1: Inter-annotator agreement on the 500-sample dataset, with bootstrap 95% confidence intervals.

3.4. Corpus Description and Availability

The SPOT Corpus contains 43,305 manually annotated comments drawn from 1,061 posts and 253 shared URLs, published across 275 public French Facebook pages and groups. Each comment is linked to its post, the shared article (URL, title, description), and the hosting community (page or group name).

Each post includes several elements forming its *publication context* (Figure 1a): the *account name*, which provides cues about the page or group's thematic or ideological orientation), the *post message* (which may reproduce, criticize, or comment on the article title), and metadata about the shared article (domain name, title, description). Comments can be direct comments to the post (28,457; 65.7%) or replies to another comment (14,848; 34.3%), in which case the parent comment is part of the publication context (Figure 1b)

Overall, 4,306 comments (9.9%) were labeled as stopping points. SPOT provides a large-scale, conversation-level resource for studying how users collectively problematize, contest, or nuance potentially misleading content in authentic social media contexts.

The SPOT corpus is made available to the research community upon request through a secure institutional data repository. Access is granted only for academic purposes after evaluation of the research project, ensuring both reproducibility and the protection of users' privacy.

4. Classification Task

4.1. Models

Encoder-based Models. Automatic comment classification in *Computational Social Science* (CSS) research—covering tasks such as counter-speech (Bonaldi et al., 2024), disagreement detection (De Kock and Vlachos, 2021), and hate speech analysis (Fortuna and Nunes, 2018)—has predominantly relied on *encoder-based transformers*. Pre-trained models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are typically fine-tuned on domain-specific datasets via a classification head, enabling effective transfer to specialized applications.

Following this line of work, we fine-tune the French pre-trained model CamemBERT (Martin et al., 2020) to detect whether a comment constitutes a *stopping point* ($y{=}1$) or not ($y{=}0$). The model serves as a strong French baseline due to its robust pretraining and established performance in social media tasks. We experimented with both CamemBERT and CamemBERT-v2 (Antoun et al., 2024), and ultimately retained the original CamemBERT model, as it achieved the best performance on our dataset.

The interpretation of a social media comment often depends on its broader publication context, including the post, article, hosting page or group, or parent comment. Encoder-based models—originally designed for sentence-level tasks—tend to overlook this context due to their limited input window and their only later adaptation to social media and conversational data (Castelle, 2018; Flek, 2020). Recent studies have shown that incorporating contextual elements, such as parent comments, post framing, or hosting group community, improves model performances (Park et al., 2021; Bourgeade et al., 2024; Nouri et al., 2025).

We therefore extend our baseline with two contextual variants commonly used in context-aware encoder architectures: (i) Context Concat, where contextual text (post message, article title, or a concatenation of all contextual elements) is appended to the input comment using the `[SEP]` token; and (ii) Context Embed, where all contextual elements (post, article title, page/group name, and parent comment) are first concatenated into a single text sequence to produce one context embedding. This embedding is then concatenated with the comment embedding and projected back to a 768-dimensional space before classification.

We define the contextual text as the concatenation of all available elements of the publication context, each preceded by a tag indicating its type (`[PARENT]` parent comment `[SEP]` `[ARTICLE]` article title (in green in Figure 1a) `[SEP]` `[PAGE]` page name (in orange in Figure 1a), etc.). To fit

within CamemBERT's 512-token window, each element is truncated according to predefined contextual limits based on its mean length (see Appendix B for details). This setup enables a controlled evaluation of how contextual cues contribute to detection performance.

Large Language Models. Recent advances in *Large Language Models* (LLMs) have renewed interest in using instruction-tuned models for CSS tasks without task-specific fine-tuning (Mu et al., 2024). LLMs such as Llama-3, Mistral, and Qwen can perform text classification via prompt adaptation, allowing flexible task formulation through natural language instructions. However, as emphasized by Ziems et al. (2024), their performance on social media data—especially for non-English content (Mohta et al., 2023)—remains uneven and often below that of fine-tuned encoders.

To situate stopping point detection within this emerging paradigm, we evaluate a set of state-of-the-art open-weights instruction-tuned LLMs—Llama 3.2 3B Instruct (Grattafiori et al., 2024), Mistral 7B Instruct-v0.2 (Jiang et al., 2023), and Qwen2.5 7B Instruct (Qwen et al., 2025). In addition, we include GPT-4o-mini, a recent proprietary model from OpenAI, as a closed-source reference of comparable size (approximately 8B parameters (Zeff, 2024)). This inclusion enables a direct comparison between open and commercial instruction-tuned models under identical prompting conditions, allowing us to assess the extent to which open-weights LLMs approach the performance of state-of-the-art closed systems on multilingual and noisy social media data.

All models are evaluated under *zero-shot* (Kojima et al., 2022), *few-shot* (Brown et al., 2020), and *chain-of-thought* (CoT) (Wei et al., 2022) prompting strategies. Annotation guidelines are reformulated as natural language prompts following recent practices in CSS research (Ziems et al., 2024), and we systematically experiment with prompts in both French and English, as well as with alternative label formulations (*Yes/No*, *1/0*, *Stop/No stop*). Consistent with prior findings (Mu et al., 2024; Ziems et al., 2024), we observe that seemingly minor prompt variations can lead to substantial and unpredictable differences in model behavior. For the CoT setting, we transform the annotation guidelines into a structured decision tree with illustrative examples, guiding the model through reasoning steps analogous to those followed by human annotators.

As with encoder-based models, we also evaluate the inclusion of publication context—article, post, parent comment, domain, and page or group name—directly within the prompts, using the same tags and truncation limits for comparability.

Overall, we test six prompt configurations: zero-shot and few-shot with and without context, and CoT with context. All prompts are available for reference in Appendix C. Fine-tuning of LLMs was not attempted due to computational costs limitations; future work may explore fine-tuned or adapter-based approaches. All LLM experiments were run with temperature $T{=}0$ for reproducibility.

4.2. Experiments and Results

Model Comparison. We compare the performance of the model architectures introduced in Section 4.1 on the stopping point classification task using the test set. Since stopping points constitute approximately 10% of the annotated corpus, and the train/validation/test splits were sampled to preserve this class distribution, the random baseline, reported for reference, samples from $P(y{=}1){=}0.1$. All encoder-based models were fine-tuned on the training set following the procedure described in Appendix D. Table 2 reports the mean F_1 scores (%) for all model configurations. For encoder-based models, we include 95% confidence intervals computed over five independent runs. For LLM-based experiments, the temperature was set to $T = 0$ to ensure reproducible outputs; therefore, confidence intervals are not reported for these results. For conciseness, we display results only for the best-performing open-weights model (Qwen2.5-7B-Instruct, as Qwen) and the best-performing proprietary model (GPT-4o-mini as GPT). Results for all evaluated LLMs are provided in Appendix E.

Model	F_1 **(%)**
Random	16.4
Qwen Zero-Shot (No Context)	39.23
Qwen Zero-Shot (Context)	45.59
Qwen Few-Shot (No Context)	53.62
Qwen Few-Shot (Context)	42.52
Qwen CoT (Context)	45.57
GPT Zero-Shot (No Context)	53.39
GPT Zero-Shot (Context)	55.94
GPT Few-Shot (No Context)	62.94
GPT Few-Shot (Context)	55.57
GPT CoT (Context)	54.33
CamemBERT (No Context)	74.67 $\pm$ 0.77
CamemBERT Context Concat	76.96 $\pm$ 1.39
CamemBERT Context Embed	**78.09** $\pm$ **0.84**

Table 2: Mean F_1 scores (%) and 95% confidence intervals over five runs.

Overall, encoder-based models substantially outperform LLMs across all settings, achieving F_1 scores more than 10 percentage points higher on average. Despite using detailed prompts—explicitly defining the task, providing in-context examples, or even chain-of-thought instructions—LLMs remain far behind the finetuned encoders. This suggests that prompting instruction-tuned LLMs remains in-

sufficient for new, complex, and context-sensitive social media analysis tasks in non-English settings. In particular, as detailed in Section 3.3, stopping point detection requires understanding subtle conversational cues and pragmatic shifts that may not be captured without task-specific supervision. This gap aligns with recent findings showing that even instruction-tuned LLMs struggle to generalize to fine-grained, context-dependent classification tasks (Ziems et al., 2024; Mu et al., 2024), especially in non-English settings where pretraining data coverage is lower (Mohta et al., 2023). Interestingly, the performance of LLMs varies considerably across prompting strategies. In our experiments, few-shot prompting without context consistently yields the highest F_1 scores for both QWEN and GPT models, outperforming few-shot, and chain-of-thought with context variants. This suggests that including additional contextual elements or more complex instructions may dilute relevant information or introduce patterns that are too complex to be effectively learned without supervision, preventing LLMs from leveraging the extra context for stopping point detection.

Within the encoder-based models, incorporating conversational context improves performance, confirming prior findings in related social media comment classification tasks such as contextual hate speech and rule-violation detection (Park et al., 2021; Nouri et al., 2025). Among the two integration strategies, the CONTEXT EMBED model achieves the best performance, outperforming both the context-free and CONTEXT CONCAT variants. This supports the view that embedding the conversational context separately preserves the salience of the comment's own linguistic features, while concatenation tends to dilute them within a longer input string. Together, these results highlight the continued relevance of finetuned, encoder-based architectures for nuanced conversational modeling tasks, where context-integration mechanisms play a decisive role in capturing discourse-level dependencies.

Context Contribution Analysis. To better understand which contextual components contribute most to stopping point prediction, we conducted a series of controlled experiments inspired by ablation studies. Since many contextual fields (parent comment or post message) are not available in all instances, we used the CONTEXT CONCAT architecture to maintain consistency across models. In each run, we provided only one contextual element at a time—*article text*, *post message*, *domain name*, *page or group name*, or *parent comment*—alongside the target comment, and evaluated performance under the same conditions as before.

Results, displayed in Table 3, indicate that the *article text* provides the most informative context, yielding the largest performance gain over the no-context baseline. This is consistent with the intuition that stopping points often contain reactions or critiques targeting the shared article. The *post message* also improves classification, as it frequently aligns with or paraphrases the article's content. Notably, *domain name* and *page or group name* also contribute positively, suggesting that the media source and the social page name carry relevant cues for how ordinary users express criticism in different online contexts. These findings align with Park et al. (2021), who similarly observed that including the community name (in their case, the subreddit) through a CONTEXT CONCAT architecture improved the prediction of moderation rule violations, highlighting the value of contextual information for understanding social media discourse.

Model Configuration	F_1 **(%)**
NO CONTEXT	74.67 ± 0.77
PARENT COMMENT CONCAT	73.97 ± 1.09
PAGE/GROUP CONCAT	76.15 ± 1.35
DOMAIN CONCAT	77.15 ± 0.65
POST CONCAT	77.40 ± 1.51
ARTICLE CONCAT	78.03 ± 1.40
CONTEXT CONCAT	76.96 ± 1.39
CONTEXT EMBED	**78.09 ± 0.84**

Table 3: Performance of fine-tuned CAMEMBERT classifiers in different context configuration. Mean F_1 scores (%) with 95% confidence intervals over five runs are shown.

5. Error Analysis

To better understand the limitations of our best-performing model (CONTEXT EMBED), we conducted a manual error analysis on all 360 misclassified instances from the test set.

Each error was first categorized as a *top-level* comment or a *reply* (see Figure 1b), and then annotated as *simple* (decidable from the comment alone) or *complex* (requiring context). Complex cases were further categorized according to the typology described in the Annotation Guidelines (provided in our code repository) (explicit markers, reported speech, reply-dependency, irony, short fragments, multi-turn phenomena). Categories are not mutually exclusive, as a single instance may belong to multiple sources of ambiguity.

False positives analysis. Among the 203 false positives, 85.2% are labelled *complex*. Many stem from comments that contain explicit critical markers (77.3%) while reacting to posts that report or quote third-party claims (30.0%). In these situations the user's reaction typically attacks the quoted claim or

source rather than performing a critical intervention on the act of reposting. For example, in response to a post claiming *"President Erdogan encourages a Turkish girl to die as a martyr"*, the comment *"Nonsense! He should drop the mic and go die as a martyr himself"* is an emotional reaction aimed at the Erdogan and his rhetoric. It expresses outrage but does not invite readers to verify or report the post, and so should not be annotated as a stopping point. Reply-specific phenomena also matter: 36.5% of errors involve reply dynamics and 11.8% match a "reply to a stopping point" pattern in which the model mistakes a reactive turn for a critical intervention. Additionally, irony accounts for 16.3% of false positive cases, and short comments for 5.4%. Together, these observations reveal a clear pattern: the context-aware encoder still over-weights surface cues ("fake", "montage", URLs, numerical claims) and contextual signals associated with controversy, which leads it to misread phatic, ironic or meta-discursive reactions as stopping points.

False negative analysis. Out of 157 false negatives, 79% correspond to complex cases, confirming that most model errors occur in linguistically or contextually ambiguous situations. Replies represent 45% of false negatives, compared to 34.3% in the overall dataset, suggesting that replies constitute a particularly challenging structure. Among these, 38% are replies to stopping points, a type of interaction that is also difficult for human annotators. The most frequent source of false negative errors is the absence of explicit refutation markers (51.6%), showing that the model tends to rely on surface lexical cues rather than pragmatic or discourse-level information. Irony and humor (12.7%) also account for a notable share of errors, illustrating the limitations of encoder models in capturing implicit stance or socio-cultural nuances.

Category	Count	Percent (%)
False Positives (N = 203)		
No refutation with markers	157	77.3
Reported speech	61	30.0
Irony / humor	33	16.3
Reply to a stopping point	24	11.8
Short or fragmentary	11	5.4
False Negatives (N = 157)		
Refutation without markers	81	51.6
Reply to a stopping point	27	17.2
Irony or humor	20	12.7

Table 4: Error types for the CONTEXT EMBED model. Percentages are within each class (FP/FN); only categories with $>$10 cases are shown.

In summary, the error analysis shows that context-aware encoders over-rely on lexical markers, producing false positives when explicit cues of criticism appear in supportive comments, and false negatives when criticism is implicit or ironic. Stopping points are also detected more reliably in top-level comments than in replies, highlighting the need to model conversational structure or use separate models for different comment types.

6. Conclusion and Future Directions

We present SPOT, the first large-scale corpus of 43,305 French Facebook comments manually annotated to capture *stopping points*. SPOT extends beyond conventional notions of user corrections or fact-checking, revealing diverse everyday critical interventions that pause, question, or redirect online discourse. The corpus includes detailed contextual metadata (post, article, domain, page or group, and parent comment) and is accompanied by comprehensive annotation guidelines to ensure transparency and reproducibility.

Using SPOT, we benchmarked fine-tuned encoders and instruction-tuned LLMs across multiple prompting strategies. Supervised encoders outperform prompted LLMs by over 10 F_1 points, showing that for nuanced phenomena like stopping points—where meaning depends on social and contextual cues—models benefit more from explicit supervision than from general-purpose instructions, particularly in non-English settings. Adding contextual metadata further improves F_1 scores from 0.75 to 0.78, emphasizing the importance of analyzing comments within their broader publication context. Error analysis shows that encoders still struggle when lexical markers contradict intent, such as irony or emotionally charged supportive comments, indicating overreliance on surface signals. Stopping points are detected more reliably in top-level comments than in replies, highlighting the need to model conversational structure or design separate models for different comment types.

Future work will improve encoder architectures to better capture conversational and social context, moving beyond linear concatenation toward graph-based or hierarchical models and integrating multimodal signals (images, videos). We will also extend the current binary formulation to a multi-label classification task that can automatically distinguish different types of stopping points, and we will separate detection of top-level comments from nested replies (e.g., distinct models or pipelines) to better capture hierarchical conversational dynamics. Finally, we plan to study critical interventions across platforms (Reddit, YouTube) and languages, while iteratively refining our annotation guide to enhance label quality and model robustness. Collectively, these directions aim to advance computational sociology and NLP by modeling online interventions as socially situated, context-dependent phenomena.

7. Ethical Considerations and Limitations

The SPOT corpus contains user-generated content from public Facebook pages and groups, including comments on posts flagged as potentially misleading. Although the posts and comments were public at the time of collection, some content may later be deleted or restricted by its authors or communities. Additionally, the dataset includes user reactions in potentially sensitive contexts, such as disagreements, critiques, or emotionally charged responses.

To protect user privacy and follow established guidelines for ethical social media research (Townsend and Wallace, 2016), we applied several precautions: (i) all user identifiers and profile names were anonymized; and (ii) we do not distribute the dataset publicly to avoid preserving or republishing sensitive material.

Access to SPOT is granted only for academic research upon request through a secure institutional repository. Each request is evaluated to ensure that the proposed use aligns with ethical guidelines and that the data will be handled responsibly. This controlled-access model balances reproducibility and research transparency with the protection of individual privacy and community norms.

8. Acknowledgements

We would like to thank Dominique Cardon, Salim Hafid, Sofia Imbert de Trémiolles, Aina Garí Soler, Théophile Pénigaud de Mourgues, Paul Lerner, and Carlo Romano Marcello Alessandro Santagiustina for their careful reading of the manuscript and for their thoughtful comments and suggestions.

This work was partially funded by the "AI For Democracy Democratic Commons" project (Bpifrance's 'Digital Commons for Generative AI', France 2030), and the French National Research Agency (ANR) under the SINNet project (ANR-23-CE23-0033-01), the France 2030 program PRAIRIE (ANR-23-IACL-0008), PostGenAI@Paris (ANR-23-IACL-0007), and TIERED (ANR-22-EXES-0014).

References

Jacob Andreas, Sara Rosenthal, and Kathleen McKeown. 2012. Annotating agreement and disagreement in threaded discussion. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, page 818–822, Istanbul, Turkey. European Language Resources Association (ELRA).

Wissam Antoun, Francis Kulumba, Rian Touchent, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2024. Camembert 2.0: A smarter french language model aged to perfection. (arXiv:2411.08868). ArXiv:2411.08868 [cs].

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, page 48–57, Portland, Oregon. Association for Computational Linguistics.

Manon Berriche. 2024. *Tu crois que c'est vrai ? : diversité des régimes d'énonciation face aux fake news et mécanismes d'autorégulation conversationnelle*. Theses, Université Paris Cité.

Leticia Bode, Emily K. Vraga, and Rongwei Tang. 2024. User correction. *Current Opinion in Psychology*, 56:101786.

Luc Boltanski. 2011. *On Critique: A Sociology of Emancipation*. Polity Press.

Luc Boltanski and Laurent Thévenot. 2006. *On Justification: Economies of Worth*. Princeton University Press.

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. Nlp for counter-speech against hate: A survey and how-to guide. In *Findings of the Association for Computational Linguistics: NAACL 2024*, page 3480–3499, Mexico City, Mexico. Association for Computational Linguistics.

Tom Bourgeade, Zongmin Li, Farah Benamara, Véronique Moriceau, Jian Su, and Aixin Sun. 2024. Humans need context, what about machines? investigating conversational context in abusive language detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8438–8452, Torino, Italia. ELRA and ICCL.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are

few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, page 1877–1901. Curran Associates, Inc.

Catherine Buerger and Lucas Wright. 2019. Counterspeech: A literature review. (3829816).

Michael Castelle. 2018. The linguistic ideologies of deep abusive language classification. In *Proceedings of the Workshop on Abusive Language Online (ALW)*, pages 160–170, Melbourne, Australia. Association for Computational Linguistics.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. Conan - counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 2819–2829, Florence, Italy. Association for Computational Linguistics.

Christine De Kock and Andreas Vlachos. 2021. I beg to differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, page 2017–2027, Online. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, page 69–76, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, page 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Neele Falk, Eva Vecchi, Iman Jundi, and Gabriella Lapesa. 2024. Moderation in the wild: Investigating user-driven moderation in online discussions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 992–1013, St. Julian's, Malta. Association for Computational Linguistics.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, page 3226–3240. ArXiv:2107.08720 [cs].

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 1163–1168, San Diego, California. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Lucie Flek. 2020. Returning the n to nlp: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7828–7838, Online. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. Semeval-2019 task 7: Rumoureval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, page 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin,

Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva,

Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. (arXiv:2407.21783). ArXiv:2407.21783 [cs].

Stuart Hall. 1980. *Encoding/Decoding*. Hutchinson.

John Heritage. 1984. *A Change-of-State Token and Aspects of Its Sequential Placement*. Cambridge University Press.

Ian Hutchby and Robin Wooffitt. 2008. *Conversation Analysis*, 2nd edition. Polity Press.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. (arXiv:2310.06825). ArXiv:2310.06825 [cs].

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, page 22199–22213, Red Hook, NY, USA. Curran Associates Inc.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2nd edition. SAGE Publications, Thousand Oaks, CA. Comprehensive methodology for content analysis, including Krippendorff's α.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1):12:1–12:37.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

James Lull. 1995. *Media, Communication, Culture: A Global Approach*. Columbia University Press.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 708–717, Vancouver, Canada. Association for Computational Linguistics.

Yingchen Ma, Bing He, Nathan Subrahmanian, and Srijan Kumar. 2023. Characterizing and predicting social correction on twitter. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23, page 86–95, New York, NY, USA. Association for Computing Machinery.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7203–7219. ArXiv:1911.03894 [cs].

Solomon Messing, Christina DeGregorio, Bennett Hillenbrand, Gary King, Saurav Mahanti, Zagreb Mukerjee, Chaya Nayak, Nate Persily, Bogdan State, and Arjun Wilkins. 2023. Facebook privacy-protected full urls data set.

Jay Mohta, Kenan Ak, Yan Xu, and Mingwei Shen. 2023. Are large language models good annotators? In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 38–48. PMLR.

Yida Mu, Ben P. Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2024. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. (arXiv:2305.14310). ArXiv:2305.14310 [cs].

Célia Nouri, Jean-Philippe Cointet, and Chloé Clavel. 2025. Graphically speaking: Unmasking abuse in social media with conversation insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 18271–18286, Vienna, Austria. Association for Computational Linguistics.

Chan Young Park, Julia Mendelsohn, Karthik Radhakrishnan, Kinjal Jain, Tushar Kanakagiri, David Jurgens, and Yulia Tsvetkov. 2021. Detecting community sensitive norm violations in online conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, page 3386–3397, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guillaume Plique, Pauline Breteau, Jules Farjas, Héloïse Théro, Jean Descamps, Amélie Pellé, and Laura Miguel. 2019. Minet, a webmining CLI tool & library for python.

Pablo Porten-Cheé, Marlene Kunst, and Martin Emmer. 2020. Online civic intervention: A new form of political participation under conditions of a disruptive online discourse. *International Journal of Communication*, 14:21–21.

Aashima Poudhar, Ioannis Konstas, and Gavin Abercrombie. 2024. A strategy labelled dataset of counterspeech. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, page 256–265, Mexico City, Mexico. Association for Computational Linguistics.

John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and Doyne Farmer. 2021. Debagreement: A comment-reply dataset for (dis)agreement detection in online debates. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. (arXiv:2412.15115). ArXiv:2412.15115 [cs].

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking in conversation. *Language*, 50(4):696–735.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Leanne Townsend and Claire Wallace. 2016. Social media research: A guide to ethics. Technical Report 16, University of Aberdeen.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, page 812–817, Istanbul, Turkey. European Language Resources Association (ELRA).

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, page 138–142, Austin, Texas. Association for Computational Linguistics.

Brendan R. Watson, Zhao Peng, and Seth C. Lewis. 2019. Who will intervene to save news comments? deviance and social control in communities of news commenters. *New Media & Society*, 21(8):1840–1858.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, page 24824–24837, Red Hook, NY, USA. Curran Associates Inc.

Maxwell Zeff. 2024. OpenAI unveils GPT-4o mini, a smaller and cheaper AI model. TechCrunch. Accessed: 2026-02-20.

Amy Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):357–366.

Marc Ziegele, Teresa K. Naab, and Pablo Jost. 2020. Lonely together? identifying the determinants of collective corrective action against uncivil comments. *New Media & Society*, 22(5):731–751.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, and Rob Procter. 2016. PHEME dataset of rumours and non-rumours. *arXiv preprint*.

A. Annotation Details

Annotator Demographics The main annotator is a female researcher in Sociology from France, aged 25–35. Two additional annotators participated in calibration: (1) a male researcher in Sociology from France, aged 35–45, and (2) a female researcher in NLP and Computational Social Science from France, aged 25–35.

Pilot Testing and Guideline Calibration To assess clarity and applicability of the annotation guidelines, a pilot study was conducted on deliberately challenging sets of ten comments selected by the main annotator. Annotators independently labeled each item and provided brief justification notes. Disagreements and the underlying rationales were discussed during group calibration sessions. This iterative process led to targeted refinements of the guidelines, including explicit instructions for handling irony and parody, rules for interpreting very short text fragments in context, and a conservative default rule when contextual evidence was insufficient. The process was repeated twice, until annotators reached a high level of consistency in applying the rules. Detailed information about the calibration sets are provided in the Annotation Guidelines.

B. Context Concatenation

To integrate contextual information, all available elements of the publication for each comment were concatenated into a single input string. Each element was prepended with a descriptive tag indicating its type: `[PARENT]` for the parent comment, `[POST]` for the post message, `[ARTICLE]` for the article title and description, `[ACCOUNT]` for the page or group name, and `[DOMAIN]` for the media source. Elements were joined using the model's `[SEP]` token.

Before concatenation, each field was truncated according to empirically defined limits based on its typical length: comment text to 600 tokens, post title to 200 tokens, article title and description to 200 tokens each, parent comment to 300 tokens, account name to 50 tokens, and domain name to 50 tokens. Empty fields were omitted from the concatenated string to reduce noise. The final sequence was truncated to fit within CamemBERT's 512-token input window, allowing the model to leverage rich contextual cues while respecting input size constraints. We use the same truncation rules when providing contextual elements to the LLMs prompts to ensure comparability.

C. Prompts Used for LLM Experiments

This section presents the prompts employed in our LLM experiments.

C.1. Prompt 1: Stopping Point Detection Based on the Annotation Guidelines

```
Rôle et tâche :
Vous êtes un annotateur. Votre tâche
    est de déterminer si un commentaire
    Facebook est un point d'arrêt (Oui)
    ou non (Non).
Chaque commentaire est associé à un
    post contenant un lien ou un
    article signalé pour fausse
    information. Le commentaire peut ê
    tre directement sous le post ou ê
    tre une réponse à un autre
    commentaire. Pour cette tâche, vous
    n'avez pas accès à ce contexte de
    publication.

Définition :
Un point d'arrêt est une intervention
    critique dans une conversation en
    ligne.
Cela peut aller d'une simple expression
    de doute à une réfutation ou un
    appel à vérification.
Un commentaire peut être un point
    d'arrêt s'il critique, corrige, é
    met un doute ou demande une
    vérification sur :
- la crédibilité ou la fiabilité du
    contenu,
- la pertinence ou la forme (texte,
    image, vidéo),
- la source ou l'auteur du post,
- les autres utilisateurs.

Cas particuliers :
- Réfutation implicite (sans mots-clés)
    : Oui si opposition, critique ou
    avertissement ; Non si simple ajout
    d'info.
- Alignement incrédule ("c'est
    peut-être faux mais je m'en fiche")
    : Non si émotion/adhésion seulement
    ; Oui si demande de vérification ou
    doute explicite.
- Ironie/parodie/sarcasme : Non si
    purement humoristique ou phatique ;
    Oui si utilisé pour réfuter,
    corriger ou exprimer une
    incrédulité critique.

Instructions :
- Déterminez si le commentaire suivant
    est un point d'arrêt (Oui) ou non
    (Non).
- Répondez uniquement par Oui ou Non.

Commentaire : « {text} »
Réponse :
```

C.2. Prompt 2: Annotation Guidelines with Few-Shot Examples

```
Rôle et tâche :
Vous êtes un annotateur. Votre tâche
    est de déterminer si un commentaire
    Facebook est un point d'arrêt (Oui)
    ou non (Non).
Chaque commentaire est associé à un
    post contenant un lien ou un
    article signalé pour fausse
    information. Le commentaire peut ê
    tre directement sous le post ou ê
    tre une réponse à un autre
    commentaire. Pour cette tâche, vous
    n'avez pas accès à ce contexte de
    publication.

Définition :
Un point d'arrêt est une intervention
    critique dans une conversation en
    ligne.
Cela peut aller d'une simple expression
    de doute à une réfutation ou un
    appel à vérification.
Un commentaire peut être un point
    d'arrêt s'il critique, corrige, é
    met un doute ou demande une
    vérification sur :
- la crédibilité ou la fiabilité du
    contenu,
- la pertinence ou la forme (texte,
    image, vidéo),
- la source ou l'auteur du post,
- les autres utilisateurs.

Exemples :
« Ah oui, bien sûr on vous croit... » →
     Oui (ironie exprimant le doute)
« Encore une rumeur Twitter » → Oui
    (réfutation implicite)
« Si seulement c'était vrai... » → Non
    (souhait sans critique)
« C'est dégueulasse » → Non (émotion
    ou indignation sans remise en cause)

Commentaire : « C'est complètement faux
    ! / fake news »
Réponse : Oui
Commentaire : « Tellement drôle haha ! »
Réponse : Non
Commentaire : « Arrêtez d'inventer des
    trucs pareils »
Réponse : Oui
Commentaire : « J'y crois pas une
    seconde »
Réponse : Oui
Commentaire : « Haha les gens sont fous
    »
Réponse : Non
Commentaire : « T'as vérifié avant de
    poster ? »
Réponse : Oui
Commentaire : « Grave, c'est choquant !
    »
Réponse : Non
Commentaire : « Ce site raconte
    toujours n'importe quoi »
Réponse : Oui
Commentaire : « Et sinon, il fait beau
    chez vous ? »
Réponse : Non
Commentaire : « Encore une intox,
    sérieux... »
Réponse : Oui

Instructions :
- Déterminez si le commentaire suivant
    est un point d'arrêt (Oui) ou non
    (Non).
- Répondez uniquement par Oui ou Non.

Commentaire : « {text} »
Réponse :
```

C.3. Prompt 3: Annotation Guidelines with Publication Context

```
Rôle et tâche :
Vous êtes un annotateur. Votre tâche
    est de déterminer si un commentaire
    Facebook est un point d'arrêt (Oui)
    ou non (Non), en vous appuyant sur
    le contexte de publication.
Chaque commentaire est associé à un
    post contenant un lien ou un
    article signalé pour fausse
    information. Le commentaire peut ê
    tre directement sous le post ou ê
    tre une réponse à un autre
    commentaire. Pour cette tâche, vous
    avez accès au post, lien ou article
    partagé, à la source médiatique, au
    commentaire parent, et la page ou
    le compte ayant publié le contenu.

Définition :
Un point d'arrêt est une intervention
    critique dans une conversation en
    ligne.
Cela peut aller d'une simple expression
    de doute à une réfutation ou un
    appel à vérification.
Un commentaire peut être un point
    d'arrêt s'il critique, corrige, é
    met un doute ou demande une
    vérification sur :
- la crédibilité ou la fiabilité du
    contenu,
- la pertinence ou la forme (texte,
    image, vidéo),
- la source ou l'auteur du post,
- les autres utilisateurs.

Cas particuliers :
- Réfutation implicite (sans mots-clés)
    : Oui si opposition, critique ou
    avertissement ; Non si simple ajout
    d'info.
- Alignement incrédule (« c'est
    peut-être faux mais je m'en fiche »
    ) : Non si émotion/adhésion
    seulement ; Oui si demande de
    vérification ou doute explicite.
- Ironie/parodie/sarcasme : Non si
    purement humoristique ou phatique ;
    Oui si utilisé pour réfuter,
```

```
    corriger ou exprimer une
    incrédulité critique.

Instructions :
- Déterminez si le commentaire suivant
    est un point d'arrêt (Oui) ou non
    (Non), en tenant compte du contexte
    de publication (post, article,
    commentaire parent).
- Répondez uniquement par Oui ou Non.

Post : « {account} : {title} »
Article partagé : « {domain} :
    {url_title} {description} »
Commentaire parent : « {parent_comment}
    »
Commentaire : « {text} »
Réponse :
```

C.4. Prompt 4: Annotation Guidelines with Publication Context and Few-Shot Examples

```
Rôle et tâche :
Vous êtes un annotateur. Votre tâche
    est de déterminer si un commentaire
    Facebook est un point d'arrêt (Oui)
    ou non (Non), en vous appuyant sur
    le contexte de publication.
Chaque commentaire est associé à un
    post contenant un lien ou un
    article signalé pour fausse
    information. Le commentaire peut ê
    tre directement sous le post ou ê
    tre une réponse à un autre
    commentaire. Pour cette tâche, vous
    avez accès au post, lien ou article
    partagé, à la source médiatique, au
    commentaire parent, et la page ou
    le compte ayant publié le contenu.

Définition :
Un point d'arrêt est une intervention
    critique dans une conversation en
    ligne.
Cela peut aller d'une simple expression
    de doute à une réfutation ou un
    appel à vérification.
Un commentaire peut être un point
    d'arrêt s'il critique, corrige, é
    met un doute ou demande une
    vérification sur :
- la crédibilité ou la fiabilité du
    contenu,
- la pertinence ou la forme (texte,
    image, vidéo),
- la source ou l'auteur du post,
- les autres utilisateurs.

Exemples :
« Ah oui, bien sûr on vous croit ... »
    → Oui (ironie exprimant le doute)
« Encore une rumeur Twitter » → Oui
    (réfutation implicite)
« Si seulement c'était vrai... » → Non
    (souhait, pas critique)
« C'est dégueulasse » → Non (outrage
    ou émotion sans critique)

Post : « Info Vaccins France : une
    nouvelle victime des industries
    pharmaceutiques... RIP »
Article partagé : «
    lesmoutonsrebelles.com : Encore un
    bébé de deux mois qui décède 48H
    après avoir reçu 8 vaccins »
Commentaire parent : « Mais sérieux?!
    Ils comprennent rien de rien! Il en
    faudra combien de victimes sérieux! »

Commentaire : « Avant les vaccins c é
    tait des milliers de victimes et
    d'enfants qui mouraient.. faudra
    pas venir pleurer après! »
Réponse : Oui

Post : « LOSC : Le monde du foot est en
    deuil ! »
Article partagé : « losc.fr : Un grand
    joueur s'éteint, le monde du
    football en pleure »
Commentaire parent : «»
Commentaire : « Toutes ces pages de pub
    pour annoncer (en retard) que
    Stéphane Paille est décédé ! Le 27
    juin.»
Réponse : Oui

Post : « 1 Million de J'aime Contre
    Emmanuel Macron : »
Article partagé : «
    valeursactuelles.com : Parlement
    européen : Bayer-Monsanto finance
    bien le parti de Macron »
Commentaire parent : « @USER et oui ...
    »
Commentaire : « On comprend mieux
    pourquoi nous mangeons encore du
    cancer »
Réponse : Non

Post : « Force gilet jaune 31 : »
Article partagé : «planetes360.fr : «
    Je demande des efforts aux Français »
    ... 14 chauffeurs, 60 cuisiniers et
    hôteliers : les chiffres sur le
    cabinet d'Édouard Philippe -
    PLANETES360 »
Commentaire parent : «»
Commentaire : « Honteux... D'émission
    !!! »
Réponse : Non

Instructions :
- Déterminez si le commentaire suivant
    est un point d'arrêt (Oui) ou non
    (Non), en tenant compte du contexte
    de publication (post, article,
    commentaire parent).
- Répondez uniquement par Oui ou Non.

Post : « {account} : {title} »
Article partagé : « {domain} :
    {url_title} {description} »
Commentaire parent : « {parent_comment}
    »
```

```
Commentaire : « {text} »
Réponse :
```

C.5. Prompt 5: Annotation Guide with Publication Context and Chain-of-Thought Reasoning

```
Rôle et tâche :
Vous êtes un annotateur. Votre tâche
    est de déterminer si un commentaire
    Facebook est un point d'arrêt (Oui)
    ou non (Non), en vous appuyant sur
    le contexte de publication.
Chaque commentaire est associé à un
    post contenant un lien ou un
    article signalé pour fausse
    information. Le commentaire peut ê
    tre directement sous le post ou ê
    tre une réponse à un autre
    commentaire. Pour cette tâche, vous
    avez accès au post, lien ou article
    partagé, à la source médiatique, au
    commentaire parent, et la page ou
    le compte ayant publié le contenu.

Définition :
Un point d'arrêt est une intervention
    critique dans une conversation en
    ligne.
Cela peut aller d'une simple expression
    de doute à une réfutation ou un
    appel à vérification.
Un commentaire peut être un point
    d'arrêt s'il critique, corrige, é
    met un doute ou demande une
    vérification sur :
- la crédibilité ou la fiabilité du
    contenu,
- la pertinence ou la forme (texte,
    image, vidéo),
- la source ou l'auteur du post,
- les autres utilisateurs.

Chaîne de raisonnement :
Étape 1 - Identifier le niveau du
    commentaire
- Parent vide → commentaire de premier
    niveau (Étape 2A).
- Parent présent → commentaire de
    second niveau (Étape 2B).

Étape 2A - Commentaire de premier niveau
- Si le post et l'article se
    contredisent :
  Oui → si le commentaire apporte un é
    lément de critique, de preuve, de
    doute ou de signalement sur le
    contenu partagé ou sa lecture.
  Non → si le commentaire approuve
    simplement le post (ex. "Exactement
    !").
  - Si le commentaire met en doute,
    corrige ou critique le contenu, la
    source ou la page → Oui.
- Si le commentaire réfute
    implicitement la position du post
    (ex. post anti-vax / commentaire
    pro-vax) → Oui.
- Si le commentaire mentionne la source
    pour en questionner la fiabilité →
    Oui.
- Si le commentaire se moque du site ou
    de la page (ex. "Encore une fake
    news de...") → Oui.
- Sinon, ou si le commentaire approuve,
    exprime une émotion ou est hors
    sujet → Non.

Étape 2B - Commentaire de second niveau
- Si le parent soutient le post → Oui
    si le commentaire le contredit ou
    le critique.
- Si le parent critique le post → Oui
    si le commentaire l'appuie avec un
    nouvel argument ; Non s'il se
    contente d'approuver.
- Sinon, ou si le commentaire attaque
    le parent critique → Non
    (contre-stop).

Instructions :
- Déterminez si le commentaire suivant
    est un point d'arrêt (Oui) ou non
    (Non), en tenant compte du contexte
    de publication (post, article,
    commentaire parent).
- Suivez rigoureusement la chaîne de
    raisonnement ci-dessus avant de
    répondre.
- Répondez uniquement par Oui ou Non.

Post : « {account} : {title} »
Article partagé : « {domain} :
    {url_title} {description} »
Commentaire parent : « {parent_comment}
    »
Commentaire : « {text} »
Réponse :
```

D. Training Setup and Context Concatenation

For all encoder-based experiments, the dataset was divided into training, validation, and test splits following an 80/20 train–test ratio, and an additional 80/20 split on the training portion to create the validation set. Since stopping points constitute approximately 10% of the annotated corpus, all splits were stratified to preserve this class distribution.

Models were trained using a weighted cross-entropy loss to compensate for class imbalance (approximately 10% positive samples). Training was performed for 25 epochs using the AdamW optimizer with a learning rate of 2×10^{-5}, weight decay of 0.01, and a batch size of 4 with gradient accumulation over 8 steps to yield an effective batch size of 32. All experiments were conducted on distributed multi-GPU setups (NVIDIA RTX 8000 and H100), with fixed random seeds to ensure full reproducibility.

E. LLM results

Model	z|f|c	ctx?	F_1	Prec.	Recall	Inval
QWEN	z	n	39.2	67.3	27.7	0
QWEN	f	n	53.6	54.2	53.1	2
QWEN	z	y	45.6	57.0	38.0	0
QWEN	f	y	42.5	34.5	55.4	6
QWEN	c	y	45.6	49.1	42.5	0
MISTRAL	z	n	36.1	55.6	26.8	385
MISTRAL	f	n	44.0	33.9	62.7	1848
MISTRAL	f	n	43.0	32.0	65.3	1637
MISTRAL	z	y	27.8	21.8	38.2	692
MISTRAL	f	y	25.7	16.0	64.8	898
MISTRAL	c	y	31.8	23.0	51.6	963
LLAMA	z	n	24.1	15.3	56.6	0
LLAMA	f	n	22.4	12.8	88.5	2
LLAMA	z	y	16.4	9.0	93.6	1656
LLAMA	f	y	20.2	11.4	90.6	9
LLAMA	c	y	15.1	8.5	65.3	5180
GPT	z	n	53.4	57.8	49.6	0
GPT	f	n	**62.9**	60.5	65.6	0
GPT	z	y	55.9	55.6	56.2	0
GPT	f	y	55.6	45.7	**71.0**	0
GPT	c	y	54.3	**61.3**	48.8	0

Table 5: Extended results of LLMs under different prompting strategies ('z' for zero-shot, 'f' for few-shot, and 'c' for chain-of-thought) and contextual settings ('y' for 'with context', and 'n' for 'without context') . Invalid outputs indicate the number of cases where the model failed to produce a valid label.