# Claim Check-Worthiness Detection: How Well do LLMs Grasp Annotation Guidelines?

Anonymous ACL submission

#### Abstract

The rising threat of disinformation underscores the need to fully or partially automate the factchecking process. Identifying text segments requiring fact-checking is known as claim detection (CD) and claim check-worthiness detection (CW), the latter incorporating complex domainspecific criteria of worthiness and often framed as a ranking task. Zero- and few-shot LLM prompting is an attractive option for both tasks, as it bypasses the need for labeled datasets and allows verbalized claim and worthiness criteria 011 to be directly used for prompting. We evaluate the LLMs' predictive accuracy and accuracy on five CD/CW datasets from diverse domains, 014 each utilizing a different worthiness criterion. We examine two key aspects: (1) how to best distill factuality and worthiness criteria into a prompt, and (2) how much context to provide for each claim. To this end, we experiment with different levels of prompt verbosity and varying amounts of contextual information given to the model. We additionally eval-022 uate the top-performing models with ranking metrics, resembling prioritization done by factcheckers. Our results show that optimal prompt 026 verbosity varies, meta-data alone adds more performance boost than co-text, and confidence scores can be directly used to produce reliable check-worthiness rankings.

#### 1 Introduction

034

042

The vast amount of information spread globally, coupled with misinformation and disinformation, is intensifying the demand for fact-checking (News, 2022; Idrizi and Hanafin, 2023). As a result, there is a growing importance placed on automating the fact-checking process. However, complete automation may not necessarily be the solution. For example, PolitiFact, a fact-checking organization that utilized ChatGPT to verify claims already factchecked, noted several issues with automatic factchecking, such as inconsistency, knowledge limitations, and misleadingly high confidence (Abels,



Figure 1: Using annotation guidelines, we craft zeroand few-shot LLM prompts for claim and claim checkworthiness detection, varying the level of prompt verbosity and the amount of provided context. We evaluate the LLMs using classification and ranking metrics.

2023). However, they recognize the potential of using language models to assist fact-checkers, especially in identifying claims worthy of verification. Similarly, FullFact, another fact-checking organization, identified the lack of effective tools for selecting claims to check as a primary workflow challenge (FullFact, 2020).

To warrant fact-checking, a claim must be both *factual* (i.e., related to purported facts) and *check-worthy* (i.e., of interest to society). The NLP tasks of identifying factual and check-worthy claims are known as *claim detection* (CD) and *claim check-worthiness detection* (CW), respectively. The tasks make up the first component of the automatic fact-checking pipeline. While both tasks are typically defined as classification tasks, CW can also be framed as a ranking task, mimicking the prioritization process employed by fact-checking organizations (FullFact, 2020).

Both CD and CW are challenging for several

reasons. Firstly, the underlying concepts of factual 063 claims and check-worthiness resist straightforward 064 definitions. To grasp factuality, Konstantinovskiy 065 et al. (2021) presented a thorough categorization of factual claims, while Ni et al. (2024) provided a definition distinguishing opinions. Regardless of these variations, factual could be deemed universal and self-explanatory, unlike check-worthiness, a term frequently used in previous research. Defining check-worthiness is made more challenging by its subjective, context-dependent nature and temporal variability. Assessing it usually requires choosing more specific criteria, such as relevance to the general public (Hassan et al., 2017a) or policymakers, potential harm (Nakov et al., 2022), or alignment 077 with a particular topic (Stammbach et al., 2023; Gangi Reddy et al., 2022)). Another challenge is identifying the situational context (including previous discourse and speaker information) required to determine claim factuality and check-worthiness. For example, in CW annotation campaigns (Hassan et al., 2017a; Gangi Reddy et al., 2022), annotators are typically presented with surrounding sentences to aid their assessment. 086

> The CD and CW tasks have been approached using both traditional supervised machine learning and fine-tuning pre-trained language models, both of which depend on labeled data. However, obtaining such datasets can be challenging as they need to align with specific languages, domains, and genres and meet desired factuality and worthiness criteria. Moreover, dataset annotation is costly and requires redoing if criteria change. LLMs present a viable alternative to supervised methods owing to their strong zero- and few-shot performance (Kojima et al., 2022; Brown et al., 2020). Over time, fact-checking organizations have refined principles for claim prioritization, and zero- and few-shot prompting offers a seamless way to transfer this knowledge to the model. Thus, an effective strategy might entail zero- and few-shot prompting with check-worthiness criteria from annotation guidelines. The challenge, however, is that LLMs often exhibit sensitivity to variations in prompts (Mizrahi et al., 2024) and unreliability (Si et al., 2023).

097

101

103

104

106

107

108

109

110

111

112

113

114

In this paper, we study the predictive and calibration accuracy of zero- and few-shot LLM prompting for CD and CW. We experiment with five datasets, each with a different factuality or worthiness criterion outlined in the accompanying annotation guidelines. We investigate two key aspects: (1) how to best distill factuality and worthiness criteria from the annotation guidelines into the prompt 115 and (2) what amount of context to provide for each 116 claim. For (1), we experiment with varying the 117 level of prompt verbosity, starting from brief zero-118 shot prompts to more detailed few-shot prompts 119 that include examples. For (2), we expand the 120 prompt with co-text and other components of the 121 claim's situational context. Furthermore, inspired 122 by the fact-checker's prioritization process, we con-123 sider CW as a ranking task, using LLM confidence 124 scores as a proxy for determining priority. Figure 1 125 depicts the workflow of our experiments. We show 126 that prompting with worthiness criteria adopted 127 from annotation guidelines can yield accuracy and 128 ranking scores comparable to or surpassing exist-129 ing CD/CW methods. Although optimal prompt 130 verbosity varies across datasets, certain in-domain 131 trends can be observed across models. We also 132 find that the impact of adding context is greater 133 for lower verbosity levels, while meta-data is more 134 beneficial than co-text. Finally, we show that confi-135 dence scores can be directly used to produce reli-136 able check-worthiness rankings. 137

Our contributions include the analysis of LLM performance with respect to (1) level of prompt detail, (2) provided context and co-text, and (3) the differences across various domains and worthiness criteria.

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

## 2 Related Work

Developing a fully automated fact-checking system is appealing for both its applicability and the challenge it presents (Hassan et al., 2017c; Li et al., 2023). However, Glockner et al. (2022) question the purpose of such a system, pointing to its reliance on counter-evidence that may not be available for newly coined disinformation. This motivates a shift toward human-in-the-loop approaches and automating parts of the fact-checking pipeline.

The CD and CW tasks constitute the first part of the fact-checking pipeline and are meant for selecting parts of the input for which fact-checking is possible (CD) or deemed necessary (CW). Typically framed as classification tasks, the CD and CW tasks are handled using traditional supervised machine learning (Hassan et al., 2017b; Wright and Augenstein, 2020; Hassan et al., 2017a; Gencheva et al., 2017) or fine-tuning pre-trained language models (Stammbach et al., 2023; Sheikhi et al., 2023). Methods of solving include rich sentence and context-level features (Gencheva et al., 2017), speaker, object, and claim span identification (Gangi Reddy et al., 2022), or incorporating domain-specific knowledge by combining ontology and sentence embeddings (Hüsünbeyi and Scheffler, 2024). CW can also be framed as a ranking task (Jaradat et al., 2018; Gencheva et al., 2017), mimicking the prioritization of claims by fact-checking organizations.

165

166

167

168

169

170

171

172

173

174

175

176

178

179

180

182

183

184

186

187

190

191

192

193

194

195

196

207

208

209

Recently, the use of LLMs for CD and CW is starting to take on. Sawinski et al. (2023) and Hyben et al. (2023) compare the performance of fine-tuned language models such as BERT with LLMs using zero- and few-shot learning as well as fine-tuning. Although zero- and few-shot approaches for LLMs underperform, the authors note their reliance on internal definitions of worthiness and limited prompt testing. As part of the fully automated fact-checking system relying only on LLMs, Li et al. (2023) implement a CD module using a verbose few-shot prompt, yet they do not report performance metrics. Finally, Ni et al. (2024) tackle CD by proposing a three-step prompting approach to examine model consistency. However, neither Li et al. (2023) nor Ni et al. (2024) address the CW task. To our knowledge, there is no work on CW focused on describing specific worthiness criteria using verbose prompts.

#### **3** Datasets

Our experiments utilize five datasets in English covering diverse topics and genres. Examples from each dataset are presented in Table 1. We next describe each dataset in more detail, including the CD and CW criteria used.

**ClaimBuster (CB)** (Hassan et al., 2017a) is a widely used dataset of claims from USA presidential debates. It uses ternary labels (*non-factual*, *unimportant factual*, *check-worthy factual*), which allows one to distinguish between check-worthy and unimportant factual claims, therefore covering both the CD and CW tasks. The authors consider claims to be check-worthy if the general public would be interested in knowing their veracity. However, no specific definition of factuality is provided – unimportant factual claims are defined as those lacking check-worthiness.

210CLEF CheckThat!Lab 2022 (CLEF) (Alam211et al., 2021) is a dataset of tweets relating to212COVID-19. The dataset comprises two parts: a set213of tweets containing claims and a subset of those

containing check-worthy claims, thus covering both the CD and CW tasks. Check-worthiness is defined as the need for professional fact-checking, excluding claims that are jokes, not interesting or "too trivial to check". Factual claims are defined as sentences that assert something is true and can be verified using factual information, such as statistical data, specific examples, or personal testimony.

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

**EnvironmentalClaims (ENV)** (Stammbach et al., 2023) is a dataset compiled from environmental articles and reports. The dataset focuses on check-worthy environmental claims related to green-washing in marketing strategies. The authors define specific criteria for an environmental claim that extend beyond the topic itself (e.g., highlighting the positive environmental impact of a product, not being too technical). Furthermore, the annotators are instructed to label only the explicit claims, discouraging the selection of claims with inter-sentence coreferences.

**NewsClaims (NEWS)** (Gangi Reddy et al., 2022) is a dataset of sentences from news articles on COVID-19, with metadata available for positives (speaker, object, claim span). The annotators were asked to judge whether a claim falls into one of the four topic-specific categories, which essentially formed the worthiness criteria, even though check-worthiness was not explicitly mentioned in the annotation guidelines. The dataset includes both check-worthy and non-checkworthy claims with inter-sentence coreferences (e.g., *That's also false*), which typically require inspecting the surrounding context to determine their check-worthiness (we estimate this applies to about 10% of claims in the test set).

**PoliClaim (POLI)** (Ni et al., 2024) covers the same topic as ClaimBuster (politics, speeches of governors) but labels only verifiable claims, leaving out check-worthiness. The authors provided detailed guidelines on what constitutes verifiable claims, emphasizing the need for specificity and differentiation from opinions lacking factual basis. They also highlighted the importance of considering co-text in cases involving coreference. To handle ambiguous cases, they employed a ternary (*Yes, No, Maybe*) annotation scheme. *Maybe* indicates that a claim may contain factual information but does not fully meet all criteria. For claims labeled as *Maybe*, annotators answered a follow-up Yes-No question to determine whether the claim leans

Dataset	Label	Example
СВ	× 0 √	I would do the opposite in every respect. I have met with the heads of government bilaterally as well as multilaterally. Fifty percent of small business income taxes are paid by small businesses.
CLEF	× 0 √	If the vaccine was dangerous they would've given it to poor people first, not politicians and billionaires. Today, FDA approved the first COVID-19 vaccine for the prevention of #COVID19 disease in individuals 16 years of age and older. They said the vaccine stopped transmission. Now they are lying and saying they didn't. Video proof here
ENV	X V	We Love Green! The environment is at the heart of Parisian electro-pop music festival We Love Green. All pension fund clients have a target for carbon reduction of the equity investments.
NEWS	× ✓	In Germany, RT has also amplified voices questioning the threat of COVID-19, and calling testing and mask-wearing into question. "If you wash and dry a cloth face mask on high heat, then you should be good to go," according to professor Travis Glenn
POLI	X O	As I have said all along, the courts are where we will win this battle. I promised that our roads would be the envy of the nation.

Table 1: Examples from the datasets used. X = non-factual claim, O = factual claim,  $\checkmark =$  check-worthy claim

toward factual information or subjective opinion. As with NewsClaims, inter-sentence coreference was considered, necessitating a reliance on context. Since the claims are extracted from political speeches, many of them include personal pronouns (I, we), which necessitates coreference resolution to identify the claimant or subject.

264

270

271

276

277

278

279

284

288

290

291

292

295

We use the above five datasets because they disclose annotation guidelines in some detail and cover different topics, genres, and worthiness criteria. Table 2 summarizes the datasets' characteristics (cf. Appendix A for more details). The CB and CLEF datasets cover both the CD and CW tasks, with CB using ternary labels annotated together and CLEF employing binary labels with separate annotation questions for CD and CW. The five datasets were either originally annotated using a binary scheme (ENV), Likert scale (CLEF-CW), multi-class based on topic (NEWS), or a followup prompt for uncertain instances (POLI). All the datasets provided aggregated binary labels, except CB, where aggregation from ternary into binary CD and CW labels is straightforward. The reported inter-annotator agreement for POLI and CLEF is substantial (Landis and Koch, 1977), while the agreement for ENV and NEWS datasets is moderate, confirming the complexity of the domaindependant CW task.

#### 4 Experimental Setup

In our experiments, we use both closed-source and open-source LLMs. For closed source, we use OpenAI models *gpt-turbo-3.5* and *gpt-4-turbo*. For open-source models, due to hardware constraints, we opted for Llama 3 8B Instruct, we chose Llama 3 8B Instruct, which is the top performer in its parameter class. To ensure reproducibility and encourage deterministic behaviour, we prompt GPT models with the temperature setting of 0 along with a fixed seed parameter and use greedy sampling with top\_p=1 for open-source models. We also experimented with smaller open-source models. Mistral 7B Instruct v0.2 was not compliant with the provided labels, instead giving open-ended answers, even for less verbose prompts. See Appendix B for more detailed information on models.

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

#### 4.1 Prompt Verbosity

We first investigate how prompt verbosity affects LLMs' predictive accuracy. We hypothesize that the optimal verbosity level depends on the dataset, reflecting the factuality and worthiness criteria differences between the domains. While a brief prompt might lack essential details, a comprehensive prompt featuring extensive definitions and examples may make the task more difficult to solve. Across datasets and for each prompt level, we aim to preserve the original wording and typography of the annotation guidelines as much as possible since we aim to establish whether guidelines without much intervention can be used as prompts for up-to-par performance. We additionally instruct the model to reply using only the provided labels without additional explanation to increase compliance and streamline evaluation. For POLI, we use the same question structure as in the annotation – for instances where the model responded with Maybe,

	СВ	CLEF	ENV	NEWS	POLI
Task	CD+CWD	CD+CWD	CWD	CWD	CD
Labels	ternary	binary*	binary	binary	binary*
# instances	23,533	3,040	2,647	7,848	52 speeches
# instances used	1,032	251	570	6,129	816
Genre	debates	tweets	news articles	reports	speech transcripts
Topic	politics	healthcare	environment	healthcare	political
Co-text	4 preceding, on request	_	not available	inconclusive	1 preceding, 1 following
Agreement	*	0.75/0.7	0.47	0.405	0.69
Agreement metric	-	Fleiss- $\kappa$	Krippendorff- $\alpha$	Krippendorff- $\kappa$	Cohen- <i>k</i>

Table 2: Characteristics of the CD and CW datasets used in our experiments. \*CB reported no agreement evaluation, but the test set used is agreed upon by experts.

we prompt it again with the follow-up question, providing previous responses in the prompt.

329 330

331

334

335

336

340

341

342

343

345

347

349

353

354

Based on the content and style of annotation guidelines, we define the following four levels of verbosity (cf. Appendix D for full prompts for four verbosity levels across the five datasets):

- Level V0 serves as the baseline. We use a naive zero-shot prompt, relying on internal definitions of the model. For the CD task (for the CB, CLEF and POLI datasets), we use the following prompt: "Does the following sentence/statement/tweet contain a factual claim? Answer only with Yes or No." For the CW task (for the CB, CLEF, NEWS and ENV datasets) we use the following prompt: "Does the following sentence/statement/tweet contain a check-worthy claim? Answer only with Yes or No." As these prompts do not include the specific factuality or worthiness criteria from the guidelines, they serve as a domainagnostic baseline;
- Level V1 uses prompts that include the task definition and the set of possible labels but omit detailed explanations of the labels or principles. For example, for the CB dataset, the three categories of non-factual, unimportant factual, and check-worthy factual sentences are introduced but not explained;
- Level V2 expands on V1 by adding a more detailed explanation of the labels or general annotation principles (or both, in the case of PoliClaim). Some principles include avoiding implicit assumptions (ENV), defining checkworthiness criteria based on public interest (CB), and categorizing claims that can be verified by non-professionals as non-checkworthy (CLEF);

Level V3 builds on V2 by including examples from the original annotation guidelines. This level closely aligns with annotation guidelines, encompassing all or nearly all information the datasets' authors provide in their accompanying papers.<sup>1</sup> The examples are provided either along with the labels (CB), separately in a few-shot fashion (ENV), or both (POLI).

#### 4.2 Amount of Context

In real-world scenarios, claims are rarely evaluated in isolation. Accordingly, annotators working with CD and CW datasets were usually provided with some contextual information. However, the quantity of context varied between datasets (cf. Table 2), and furthermore, the role of context also differed among datasets - sometimes it was provided as guidance (CB), while in others, it was deemed crucial for assigning labels (POLI, NEWS). For NEWS, the amount of provided co-text is also inconclusive, so we decided to ommit it from co-text expansion. This difference highlights that co-text is both another undefined aspect of CD and CW, and that it can vary across domains. Another component of context is metadata such as speaker, affiliation, occasion, and date, revealed only during annotation for CLEF-CW, but not available in the dataset itself. However, metadata is available for the CB and POLI datasets, while NEWS provides metadata only for positives, making it unusable for our experiments.

We investigate how LLMs' predictive accuracy depends on the amount of situational context provided to the model. To this end, we leverage the context information available in the CB and POLI datasets and expand the prompts in three variants: 372

373

366

- 374
- 375 376 377 378
- 378 379 380 381

382

393

394

396

397

398

399

400

 $<sup>^{1}</sup>$ CB and ENV documented additional examples (typically 20–30 examples) provided to the annotators. We did not include these examples.

- Level C1 represents adding the co-text of the claim. The amount of co-text included in the 402 prompt is the same as what was originally 403 shown to the annotators – for CB, four pre-404 ceding statements (which were either by the 405 speaker, opposing speaker, or moderator), and 406 for POLI, one preceding and one following statement; 408
  - Level C2 expands the contextual information by adding metadata to the claim. In the case of POLI, the metadata is the speaker's identity and political party, whereas for CB it additionally contains the speaker's title and the sentiment of the statement, provided by the authors of the dataset;
  - Level C3 combines both C1 and C2 by providing both co-text and metadata.

We appended the contextual information to the user prompts, and only modified the system prompts of POLI slightly - adding guidance on how to handle context, ommited from the no-context variants (cf. Appendix A for a detailed description).

#### 5 **Results**

401

407

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

We present the results for prompt verbosity levels in Table 3 and for different context levels in Table 4. In Table 3, we also include the previous results reported by authors in the original papers introducing the datasets (note that some results are not directly comparable to ours, as we discuss below). We use a stratified random classifier as the baseline.

## 5.1 Prompt Verbosity

Table 3 shows the baselines and F1 scores by verbosity level for gpt-4-turbo, gpt-3.5-turbo, and Llama3 8B. Both performance and the optimal verbosity level is not consistent across datasets. The accuracy generally increases with verbosity levels for CB, but the trend is reversed for ENV. We observe no consistent trend for CLEF, POLI, and NEWS datasets. The most verbose prompts (V3) generally do not achieve the highest performance, except for the GPT models and CB. This highlights that providing detailed instructions and examples can be beneficial but potentially harm performance.

**Comparison to previous benchmarks.** For CB, 444 the authors evaluated used 4-fold cross-validation 445 on different-sized subsets (4,000, 8,000 ... 20,000), 446 all containing our chosen test set, annotated by 447

experts. The authors evaluated using weighted F1-448 score, achieving a maximum score of 0.818. Our 449 highest weighted F1-scores surpass this, reaching 450 0.933 for gpt-4-turbo and 0.906 for gpt-3.5-turbo. 451 On CLEF, the best-reported result is the accuracy 452 score of .761 for CD and the F1 score of .698 for 453 CW. While our approach underperforms for CW 454 (F1 of 0.583), it achieves higher accuracy for CD 455 (0.776 on Level V2). In the case of NEWS, the 456 authors reported an F1 score, but it remains unclear 457 whether it was evaluated based on binary or multi-458 class labels, given that annotators had to categorize 459 claims into different classes. They achieved the 460 highest F1 score of 0.309, which our approach ex-461 ceeds on the subset we selected, achieving an F1 462 score of 0.583. Our subset has a higher random 463 baseline due to a higher ratio of positive examples 464 and includes all positives from the original test set. 465 For POLI, the authors evaluated using accuracy. 466 They achieved an accuracy of 0.764 on the test set 467 using gpt-3.5 and 0.862 using gpt-4. Our approach 468 showed comparable performance with GPT-3.5-469 turbo using prompt Level V3, but resulted in lower 470 accuracy for GPT-4 (0.813 for naive V0, 0.670 for 471 V1) and Llama (0.784 on V1). Lastly, for ENV, the 472 metrics are directly comparable, and our approach 473 underperforms compared to previous results. 474

**CD vs. CW.** Generally, higher performance is achieved for the CD task, although the diverse domains of the datasets and differences in guidelines prevent definitive conclusions. Therefore, comparing performance on the two datasets that cover both tasks - CB and CLEF - is most straightforward. Interestingly, a reverse phenomenon is observed between these datasets-significantly higher performance is achieved for the CD task on CLEF, whereas on CB, CW performance is slightly higher. An important difference in the two datasets is precisely in the annotation styles – CB uses the same guidelines for both tasks and ternary annotation, while for CLEF the guidelines are different for the two tasks, originally using different labelling strategies (binary for CD and Likert scale for CW).

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

Closed-source vs. open-source. Although a wider span of both open- and closed-source models is necessary for broader conclusions, especially including open-sourced models greater in size, the performance of the Llama3 8B model is proximate to the GPT models, proving the potential of prompting open-source models using annotation guidelines. Furthermore, the results of both GPT

		C	B	CL	EF	ENV	NEWS	POLI
		CD	CW	CD	CW	CW	CW	CD
stratified random previous		.375 .818*	.452 .818*	.745 .761 <sup>a</sup>	.415 .698	.404 .849	.667 .309*	.779 .862 <sup>a</sup>
gpt-4	V0	.833	.805	.797	.467	.416	<b>.583</b>	<b>.844</b>
	V1	.883	.885	.799	.552	.773	.572	.679
	V2	.908	.889	<b>.806</b>	<b>.583</b>	.690	.480	.541
	V3	<b>.919</b>	<b>.927</b>	.781	.556	.596	.523	.563
gpt-3.5	V0	.853	.718	.656	<b>.496</b>	.484	<b>.531</b>	.707
	V1	.570	.739	.490	.438	<b>.710</b>	.371	.751
	V2	.774	.800	.650	.468	.701	.348	.657
	V3	<b>.872</b>	<b>.862</b>	<b>.757</b>	.446	.65	.206	<b>.803</b>
Llama3 8B	V0	.677	.743	.769	<b>.439</b>	.290	<b>.586</b>	.812
	V1	.478	.655	.803	.415	.755	.502	.827
	V2	<b>.742</b>	<b>.751</b>	<b>.807</b>	.433	.745	.466	.712
	V3	.702	.637	.790	.426	.742	.469	.651

Table 3: F1 scores across datasets and prompt verbosity levels (V1–V3). Level V0 corresponds to the naiveprompting baseline. For previous baselines:  $^{a}$  = accuracy,  $^{*}$  = not directly comparable

models on the CB dataset could indicate a potential data leakage (Balloccu et al., 2024) since the
performance of Llama3 8B is comparable in other
datasets but lags for CB.

503

504

505

507

510

511

512

513

514

515

516

517

518

519

520

521

522

523

Worst performance. The naive baseline prompt (V0) generally outperforms the prompts based on annotation guidelines on the CLEF CW and NEWS datasets, except for V2 for CLEF CW with gpt-4turbo. For CLEF CW, the annotation guidelines are adapted from the Likert scale, where multiple characteristics are attributed to negatives (e.g., not interesting, a joke, not containing claims, or too trivial to be checked by a professional). In our prompts, we converted the Likert scale to binary, where the already diverse and vaguely defined criteria were binned in a single label, increasing complexity. For NEWS, although the dataset's purpose is claim check-worthiness detection, check-worthiness as a concept is not mentioned in the annotation guidelines. Positives are merely selected by containing claims falling into four predefined categories relating to the COVID-19 virus, and check-worthiness is assumed implicitly. This, along with the presence of inter-sentence coreference in the positive instances, might be the cause of poor performance.

#### 5.2 Amount of Context

525Table 4 shows the F1 scores by verbosity and con-526text level for all models. The benefit of including527context varies across models – there is a bigger per-528formance increase for the Llama model with added529contextual information, topping the performance530for CB in both tasks as opposed to prompts with

no context. For the GPT models, there is some positive impact of meta-data (C2). The least beneficial is the addition of co-text with no added metadata (C1), including speaker information, which is vital when given previous responses. Concerning prompt verbosity levels, context's impact is higher on less verbose prompts, showing contextual information is complimentary to brief definitions.

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

#### 5.3 Rank-Based Evaluation

In light of resource constraints, fact-checking organizations have devised principles to prioritize claims based on their check-worthiness. This invites the question of whether zero- and few-shot LLM prompting could be used for that purpose. To investigate this, we frame CW as a ranking task and rank the claims based on the LLM's confidence for the positive class. We used token likelihood of the positive class as a measure of confidence. The quality of the so-obtained ranking will depend on how well the LLM is calibrated. Thus, we first evaluate the LLMs' calibration accuracy using the expected calibration error (ECE). Figure 2 shows the predictive accuracy (F1 score) against calibration accuracy (1 - ECE) across datasets and prompt verbosity levels (we only use prompts at context level C0, i.e., we add no context information).

Per model and dataset, we select the prompt that scores high on both predictive and calibration accuracy. The prompts with the highest F1 scores are usually also the best-calibrated ones, except for NEWS, where we select level V1 as Paretooptimal.

				С	B				POLI	
			CD			CW			CD	
		V1	V2	V3	V1	V2	V3	V1	V2	V3
	C0	.883	.908	.919	.885	.889	.927	.619	.541	.563
	C1	.806	.849	.862	.803	.847	.872	.722	.650	.727
gpt-4-turbo	C2	.879	.908	.913	.880	.901	.916	.707	.470	.592
	C3	.794	.857	.877	.791	.854	.885	.692	.632	.732
	C0	.570	.774	.872	.739	.800	.862	.751	.657	.803
( ) 5 ( 1	C1	.461	.299	.513	.517	.301	.528	.790	.688	.794
gpt-3.5-turbo	C2	.560	.801	.836	.747	.826	.832	.730	.523	.704
	C3	.474	.724	.758	.643	.716	.749	.794	.754	.800
	C0	.478	.742	.702	.655	.751	.637	.827	.712	.651
11 200	C1	.460	.591	.614	.531	.528	.552	.799	.789	.803
Liama3 8B	C2	.483	.773	.764	.727	.819	.736	.807	.703	.628
	C3	.468	.610	.601	.506	.618	.556	.806	.798	.805

Table 4: F1 scores by level of context information (C1–C3) added to the prompt ranging in verbosity (V1–V3). Level C0 corresponds to the prompt level with no context information. The best scores across verbosity levels are shown in bold, and the best scores per model and dataset are highlighted in green.



Figure 2: F1 scores and calibration accuracy (1 - ECE) for the CW task, across datasets and models

		СВ	CLEF	ENV	NEWS
	AP	.951	.552	.767	.67
gpt-4	P10	1	.9	.9	1
	PR	.924	.615	.761	1
	AP	.934	.464	.796	.669
gpt-3.5	P10	1	.6	.9	.7
	PR	.919	.436	.772	.700
	AP	.878	.350	.794	.688
Llama3 8B	P10	1	.2	.1	1
	PR	.823	.282	.762	1

Table 5: Rank-based CW performance scores

Table 5 shows the rank-based performance scores for the selected prompts: average precision (AP), precision-at-10 (P@10), and precision-at-R, where R equals the total number of positives in the dataset. The rank-based performance scores mirror the classification accuracy scores: they are high for datasets with high predictive accuracy (CB and

563

564

565

566

568

569

ENV) and lower for datasets with lower predictive accuracy (NEWS and CLEF). Our results suggest that LLM models with high predictive accuracy also produce well-calibrated scores using ECE and may be readily used as check-worthiness rankers. 570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

#### 6 Conclusion

We tackled claim detection and check-worthiness tasks using zero- and few-shot LLM prompting based on existing annotation guidelines. The optimal level of prompt verbosity, from minimal prompts to detailed prompts that include criteria and examples, varies depending on the domain and guidelines style. Adding claim context (co-text and speaker information) does not improve performance. Models with high predictive accuracy can directly utilize confidence scores to produce reliable check-worthiness rankings.

## Limitations

587

614

617

619

Datasets. In our experiments, we do not use datasets created by fact-checking organizations. 589 While the datasets were created specifically for 590 the tasks of CD and CW, and most were annotated by experts, the datasets were constructed for research purposes. To most accurately evaluate the potential of using our approach in fact-checking organizations, a dataset annotated according to of-595 ficial factuality or check-worthiness criteria with 596 appropriate annotation guidelines should be used.

Models. Due to hardware constraints, no open-598 source LLMs greater than 8B parameters were used in our experiments. We acknowledge the importance of relying on open-source models in the research community and the lack of insight that results from disregarding larger open-source models. Using closed-source models has the additional 604 caveat of possible leakage of the dataset, which is a growing concern in the community (Balloccu et al., 2024). We also note that the outstanding results on the ClaimBuster dataset (CB) could be due to data leakage, considering the dataset was published sev-610 eral years ago and has a wide reach in the research of automatic fact-checking. 611

612 Languages. In this work, we only do experiments on datasets in English. This is for two rea-613 sons: (1) the necessity to understand the annotation guidelines to draft prompts using them and (2) 615 the lack of datasets in other languages. However, we acknowledge that disinformation is a global problem and that tackling it requires working with 618 multiple languages.

Lack of prompt engineering experiments In 620 this work, we do minimal prompt engineering in-621 terventions beyond merely adapting the level of detail in annotation guidelines and appending con-623 textual information. We opted for this approach instead of drafting prompts ourselves to investigate 625 how original wording, definitions, and examples given to annotators could fare with LLMs. We realize weak performance in some cases (e.g., CLEF, for the naive aggregation from the Likert scale to binary labels) and performance variations could be due to the models' sensitivity to prompt struc-632 ture, wording and examples. However, translating the complex criteria of worthiness in such a streamlined way could benefit fact-checkers. Furthermore, prompt design should be adapted for each dataset, significantly expanding the scope of this research 636

(since five datasets are used). We leave experiments regarding prompt design for future work.

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

#### Risks

Although we intend to combat the spread of disinformation with this work, there is still a potential for misuse. The prompts and insights reported in this work could potentially be used to create disinformative claims adapted to make their detection more difficult. A big challenge of disinformation detection is the growing use of generative models for creating disinformative claims. The prompts provided in this work could be reverted for generative purposes, achieving the exact opposite effect than what our work aims to achieve.

#### References

- Grace Abels. 2023. Can ChatGPT fact-check? Politi-Fact tested. Accessed: 2024-3-18.
- Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 611-649, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closedsource LLMs. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 67-93, St. Julian's, Malta. Association for Computational Linguistics.
- 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, pages 1877–1901. Curran Associates, Inc.

FullFact. 2020. The challenges of online fact checking.

797

798

799

746

747

- Revanth Gangi Reddy, Sai Chetan Chinthakindi, Zhenhailong Wang, Yi Fung, Kathryn Conger, Ahmed ELsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small, and Heng Ji. 2022. NewsClaims: A new benchmark for claim detection from news with attribute knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6002–6018, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
  - Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 267–276, Varna, Bulgaria. INCOMA Ltd.

700

701

704

710

711

712

713

714

715

716

717

718

719

720

721

723 724

727

731

732

734

736

737

738

740

741

742

743

744

745

- Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017a. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1803–1812.
- Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017b. Toward automated factchecking: Detecting check-worthy factual claims by claimbuster. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*
- Naeemul Hassan, Anil Nayak, Vikas Sable, Chengkai Li, Mark Tremayne, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, and Aaditya Kulkarni. 2017c. Claimbuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10:1945–1948.
- Martin Hyben, Sebastian Kula, Ivan Srba, Robert Moro, and Jakub Simko. 2023. Is it indeed bigger better? the comprehensive study of claim detection lms applied for disinformation tackling.
- Zehra Melce Hüsünbeyi and Tatjana Scheffler. 2024. Ontology enhanced claim detection.
- Zana Idrizi and Niamh Hanafin. 2023. Navigating the landscape of increased disinformation in europe and central asia.
- Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. Claim-Rank: Detecting check-worthy claims in Arabic and English. In Proceedings of the 2018 Conference of the North American Chapter of the Association for

*Computational Linguistics: Demonstrations*, pages 26–30, New Orleans, Louisiana. Association for Computational Linguistics.

- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats*, 2(2).
- J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.
- Miaoran Li, Baolin Peng, and Zhu Zhang. 2023. Selfchecker: Plug-and-play modules for fact-checking with large language models.
- Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation.
- Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Hamdy Mubarak, Alex Nikolov, and Yavuz Selim Kartal. 2022. Overview of the clef-2022 checkthat! lab task 1 on identifying relevant claims in tweets. In *CLEF 2022: Conference and Labs of the Evaluation Forum*, volume 3180 of *CEUR Workshop Proceedings*, pages 368–392. CEUR Workshop Proceedings (CEUR-WS.org).
- UN News. 2022. Rise of disinformation a symptom of 'global diseases' undermining public trust: Bachelet.
- Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. Afacta: Assisting the annotation of factual claim detection with reliable llm annotators.
- Marcin Sawinski, Krzysztof Wecel, Ewelina Ksiezniak, Milena Strózyna, Wlodzimierz Lewoniewski, Piotr Stolarski, and Witold Abramowicz. 2023. Openfact at checkthat!-2023: Head-to-head GPT vs. BERT - A comparative study of transformers language models for the detection of check-worthy claims. In Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of CEUR Workshop Proceedings, pages 453–472. CEUR-WS.org.
- Ghazaal Sheikhi, Samia Touileb, and Sohail Khan. 2023. Automated claim detection for fact-checking: A case study using Norwegian pre-trained language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 1–9, Tórshavn, Faroe Islands. University of Tartu Library.

Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III au2, and Jordan Boyd-Graber. 2023. Large language models help humans verify truthfulness – except when they are convincingly wrong.

803

810

811

812

813

815

816

817

818

820

821

822

826

828

830

836

841

- Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.
- Dustin Wright and Isabelle Augenstein. 2020. Claim check-worthiness detection as positive unlabelled learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 476–488, Online. Association for Computational Linguistics.

#### **A** Dataset Information

In this section, we provide details on the datasets used in our experiments.

#### A.1 Test set selection

Here, we provide details on the test set selection for each dataset. Furthermore, we state which set the authors used for evaluation and whether the results can be comparable.

**ClaimBuster.** The dataset does not have an explicit test set. The authors instead used 4-fold cross-validation on different-sized subsets during their experiments (4,000, 8,000 ... 20,000). However, a high-quality *groundtruth* set is available in the dataset. It contains 1,032 samples that experts agreed on and was used for screening during annotation. Also, all the test sets the authors used contain the screening sentences. For the quality of labels and to have somewhat comparable results to the authors, we selected the *groundtruth* set for experiments.

**CLEF.** The dataset consists of both a *dev* and a *test* set. Since the *test* set was used to evaluate teams participating in the CLEF CheckThat! the challenge, we opted to do our experiments on this set to compare to the metrics of the best-submitted solution.

EnvironmentalClaims. The dataset contains
both a *dev* and *test* set of equal size, whereas the
original work publishes metrics on both sets separately. We selected the *test* set for our experiments.

NewsClaims. The dataset provides both a *dev* and a test set; however, the disclosed sets contain only positive instances. The complete dataset consists of around 10% of positive instances, with a high number of low-quality negative instances created by errors in sentencizing and filtering - instances containing only names, dates, links. The dataset also contains duplicate instances, also in the set of positives. To create a viable subset and avoid high costs during inference, we sampled the negative instances from a normal distribution with the parameters fitted to the length of the instances. We chose to sample the same number of instances as there are positives without duplicates, creating a higher baseline.

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

**PoliClaim.** The dataset provides an explicit *test* set consisting of both gold labels and labels resulting from inference on 4 political speeches. To be able to compare results, we opted to use the complete *test* set.

## A.2 Context information

**ClaimBuster.** During the annotation of the ClaimBuster dataset, 4 preceding statements could be viewed with an extra button, which was used in 14% of all cases. Since the dataset covers presidential debates with multiple speakers, including the moderator and audience questioners, it is not completely clear how the speakers were differentiated in the provided preceding sentences. Therefore, we selected the method of differentiating the speakers arbitrarily – 'A' was used for the speaker of the statement that is meant to be annotated, and 'B' for the opposing speaker.

**EnvironmentalClaims.** No additional contextual or co-textual information was provided in the dataset. The annotators were not shown any cotext during annotation due to budget. The authors considered annotating whole paragraphs instead of sentence-level annotation but decided against it due to time and budget constraints.

**PoliClaim.** The annotators were provided with the preceding and following sentences of the one they are annotating. Since there is only one speaker (as opposed to ClaimBuster, which covers debates), there is no need for denoting the speaker, minimizing confusion in prompts. In annotation guidelines, context was explicitly mentioned, as well as clarified in examples. In our experiments, we used two versions of the prompts – one mentioning context

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

945

for experiments with co-text expansion and one
without the mention of context used when only one
sentence from the speech is provided. The two
alternatives are shown in D.

**CLEF.** The dataset consists of tweets covering 900 COVID-19 topics. For the check-worthiness task, 901 annotators were shown metadata such as time, ac-902 count, number of likes and reposts. However, this 903 information is not readily available in the dataset 904 and requires crawling the tweets to obtain it. It was 905 also not available in the dataset of the CLEF2022 906 CheckThat! Challenge, which was derived from 907 the original dataset. Since we wanted to make our 908 effort comparable to alternative methods used in 909 the competition, we did not opt for crawling the 910 911 tweets to acquire metadata.

912 **NewsClaims.** The research paper introducing the dataset has inconsistencies regarding the co-text 913 provided to annotators. While it is stated in the 914 paper that whole articles are provided for co-text, 915 in the screenshot of the annotation platform, only 916 three preceding and following sentences were pro-917 vided. Regarding context, the work emphasizes 918 the importance of metadata such as claim object, 919 speaker and span, and provides that data for positive instances (sentences containing claims related 921 to 4 specified COVID-19 subtopics). The effort of annotating the claims with metadata is worth-923 while, however we decided against using it in in-924 ference since no such data is available for negative instances.

#### **B** Model Information

927

929

931

932

933

937

938

941

For OpenAI models, we use *gpt-3.5-turbo-0125* and *gpt-4-0125-preview*. We use a temperature of 0 for all experiments. To get confidence, we use *logprobs* and *n\_probs=5*, to account for the target labels ending up as less probable tokens. We use a random seed of 42 in all experiments, to avoid stochastic answers as much as possible. The run was executed once per model and prompt variant. Inference was done through the OpenAI API. GPU hours are hard to estimate.

We use Llama3 8B Instruct for experiments on open-source models. It is the only smaller opensource model from the ones we tested compliant with provided labels. The experiments took 10 GPU hours on 2x GeForce RTX 2080 Ti. We use greedy decoding and run once per model and prompt variant. Onitial experiments were done on *neural-chat:7b-v3.3-q5\_K\_M* and *mistral:7b-instruct-v0.2-q5\_K\_M*. A total of 5 GPU hours was used.

## **C** Calibration

In this section, the ECE per prompt verbosity level is shown for all models in Table 6. The ECE is calculated with the parameters  $n_b ins = 10$  and norm = l1.

## **D** Complete prompts

This section provides the complete prompts used in our experiments. The instructions were given in system prompts, while the instances were in user prompts. The added context information is also appended to user prompts.

For each dataset, the three prompt levels are shown, with the content expanded in relation to the previous level highlighted. To visually separate the levels, Level V2 is highlighted in yellow, while Level V3 is highlighted in pink.

For CLEF, two alternative prompts are given, since for CD and CW different annotation guidelines were used. For POLI, parts of the Level V2 and Level V3 prompts regarding surrounding sentences are either provided or not, based on whether context expansion is used (surrounding sentences are given in prompts C1 and C3). Those parts are highlighted in blue.

**Baseline prompts.** Two naive zero-shot prompts serve as a baseline. For the claim detection task, the baseline is:

Does the following <sentence> contain a factual claim? Answer only with Yes or No.

For the claim check-worthiness detection task, the baseline is:

Does the following <sentence> contain a checkworthy claim? Answer only with Yes or No.

**User prompts.** The user prompts were based on how the instance was reffered to in the corresponding annotation gudelines. The instances are surrounded with HTML tags. The same is done for context expansion on CB and POLI.

		С	B	CL	ÆF	ENV	NEWS	POLI
		CD	CW	CD	CW	CD	CW	CW
	V0	.094	.068	.259	.601	.231	.322	.142
	V1	.050	.047	.196	.391	.119	.210	.271
gpt-4-turbo	V2	.043	.039	.194	.352	.127	.277	.373
	V3	.039	.032	.222	.367	.150	.194	.348
	V0	.033	.068	.212	.359	.189	.246	.257
	V1	.323	.085	.386	.609	.088	.260	.229
gpt-3.5-turbo	V2	.103	.071	.279	.560	.097	.280	.327
	V3	.061	.050	.285	.646	.100	.379	.196
	V0	.218	.126	.307	.611	.286	.314	.223
Llama 2 9D	V1	.607	.218	.244	.723	.114	.228	.172
Liamas 8B	V2	.184	.135	.241	.687	.102	.229	.321
	V3	.231	.259	.241	.686	.134	.214	.379

Table 6: ECE score by prompt level per dataset for *gpt-4-turbo*. 'CD' and 'CW' mark claim detection and claim check-worthiness detection, respectively, while 'V0' marks the score for the naive baseline

Level	Prompt
V1	Categorize the <sentence> spoken in the presidential debates into one of three categories: Non-Factual Sentence (NFS), Unimportant Factual Sentence (UFS) or Check-worthy Factual Sentence (CFS). Use only one of the three labels (NFS, UFS or CFS), do not provide any additional explanation.</sentence>
V2	Categorize the <sentence> spoken in the presidential debates into three categories: Non-Factual Sentence (NFS): Subjective sentences (opinions, beliefs, declarations) and many questions fall under this category. These sentences do not contain any fac tual claim. Unimportant Factual Sentence (UFS): These are factual claims but not check-worthy. The general public will not be interested in knowing whether these sentences are true or false. Fact-checkers do not find these sentences as important for checking. Check-worthy Factual Sentence (CFS): They contain factual claims and the general public will be interested in knowing whether the claims are true. Journalists look for these type of claims for fact-checking. Use only one of the three labels (NFS, UFS and CFS), do not provide any additional explanation.</sentence>
V3	Categorize the <sentence> spoken in the presidential debates into three categories: Non-Factual Sentence (NFS): Subjective sentences (opinions, beliefs, declarations) and many questions fall under this category. These sentences do not contain any factual claim. Here are two such examples. "But I think it's time to talk about the future." "You remember the last time you said that?" Unimportant Factual Sentence (UFS): These are factual claims but not check-worthy. The general public will not be interested in knowing whether these sentences are true or false. Fact-checkers do not find these sentences as important for checking. Some examples are as fol lows. "Next Tuesday is Election day." "Two days ago we ate lunch at a restaurant." Check-worthy Factual Sentence (CFS): They contain factual claims and the general public will be interested in knowing whether the claims are true. Journalists look for these type of claims for fact-checking. Some examples are: "He voted against the first Gulf War." "Over a million and a quarter Americans are HIV-positive." Use only one of the three labels (NFS, UFS and CFS), do not provide any additional explanation.</sentence>

Table 7: System prompts used for inference on the ClaimBuster dataset.

Level	Prompt
V1	Your task is to label the <sentence>. The information I need is whether it is an environmental claim. A broad definition for such a claim is given by the European Commission: Environmental claims refer to the practice of suggesting or otherwise creating the impression that a product or a service is environmentally friendly (i.e., it has a positive impact on the environment) or is less damaging to the environment than competing goods or services. Answer only with Yes or No.</sentence>
V2	Your task is to label the <sentence>. The information I need is whether it is an environmental claim. A broad definition for such a claim is given by the European Commission: Environmental claims refer to the practice of suggesting or otherwise creating the impression that a product or a service is environmentally friendly (i.e., it has a positive impact on the environment) or is less damaging to the environment than competing goods or services. General principles: You will be pre sented with a <sentence> and have to decide whether the <sentence> contains an ex plicit environmental claim. Do not rely on implicit assumptions when you decide on the label. Base your decision on the information that is available within the sentence. However, if a sentence contains an abbreviation, you could consider the meaning of the abbreviation before assigning the label. In case a sentence is too technical/complicated and thus not easily understandable, it usually does not sug gest to the average consumer that a product or a service is environmentally friendly and thus can be rejected. Likewise, if a sentence is not specific about having an environmental impact for a product or service, it can be rejected. Answer only with Yes or No.</sentence></sentence></sentence>
V3	Your task is to label the <sentence>. The information I need is whether it is an environmental claim. A broad definition for such a claim is given by the European Commission: Environmental claims refer to the practice of suggesting or otherwise creating the impression that a product or a service is environmentally friendly (i.e., it has a positive impact on the environment) or is less damaging to the environment than competing goods or services. General principles: You will be presented with a sentence and have to decide whether the sentence contains an explicit environmental claim. Do not rely on implicit assumptions when you decide on the label. Base your decision on the information that is available within the sentence. However, if a sentence contains an abbreviation, you could consider the meaning of the abbreviation before assigning the label. In case a sentence is too technical/complicated and thus not easily understandable, it usually does not suggest to the average consumer that a product or a service is environmentally friendly and thus can be rejected. Likewise, if a sentence is not specific about having an environmental impact for a product or service, it can be rejected. Examples: <sentence>: Farmers who operate under this scheme are required to dedicate 10% of their land to wildlife preservation. Label: Yes Explanation: Environmental scheme with details on implementation. <sentence>: UPM Biofuels is developing a new feedstock concept by growing Brassica Carinata as a sequential crop in South America. Label: No Explanation: Sentence con text would be required to understand whether it is a claim. Answer only with Yes or No. don't provide any additional explanation.</sentence></sentence></sentence>

Table 8: System prompts used for inference on the EnvironmentalClaims dataset.

Level	Prompt
V1	A verifiable factual claim is a sentence claiming that something is true, and this can be verified using factual, verifiable information such as statistics, specific examples, or personal testimony. Does the <tweet> contain a verifiable factual claim? Answer only with Yes or No, don't provide any additional explanation.</tweet>
V2	A verifiable factual claim is a sentence claiming that something is true, and this can be verified using factual, verifiable information such as statistics, specific examples, or personal testimony. Factual claims include the following: Stating a definition; Mentioning quantity in the present or the past; Making a verifiable prediction about the future; Reference to laws, procedures, and rules of operation; References to images or videos (e.g., "This is a video showing a hospital in Spain."); Statements about correlations or causations. Such correlation and causation needs to be explicit, i.e., sentences like "This is why the beaches haven't closed in Florida." is not a claim because it does not say why explicitly, thus it is not verifiable. Tweets containing personal opinions and preferences are not factual claims. Note: if a tweet is composed of multiple sentences or clauses, at least one full sentence or clause needs to be a claim in order for the tweet to contain a factual claim. If a claim exist in a sub-sentence or sub-clause then tweet is not considered to have a factual claim. For example, "My new favorite thing is Italian mayors and regional presidents LOSING IT at people violating quarantine" is not a claim, how ever, it is an opinion. Moreover, if we consider "Italian mayors and regional presi idents LOSING IT at people violating quarantine" it would be a claim. In addition, when answering this question, annotator should not open the tweet URL. Does the <tweet> contain a verifiable factual claim? Answer only with Yes or No.</tweet>
V3	A verifiable factual claim is a sentence claiming that something is true, and this can be verified using factual, verifiable information such as statistics, specific examples, or personal testimony. Factual claims include the following: Stating a definition; Mentioning quantity in the present or the past; Making a verifiable prediction about the future; Reference to laws, procedures, and rules of operation; References to images or videos (e.g., "This is a video showing a hospital in Spain."); Statements about correlations or causations. Such correlation and causation needs to be explicit, i.e., sentences like "This is why the beaches haven't closed in Florida." is not a claim because it does not say why explicitly, thus it is not verifiable. Tweets containing personal opinions and preferences are not factual claims. Note: if a tweet is composed of multiple sentences or clauses, at least one full sentence or clause needs to be a claim in order for the tweet to contain a factual claim. If a claim exist in a sub-sentence or sub-clause then tweet is not considered to have a factual claim. For example, "My new favorite thing is Italian mayors and regional presidents LOSING IT at people violating quarantine" is not a claim, however, it is an opinion. Moreover, if we consider "Italian mayors and regional presidents LOSING IT at people violating quarantine" it would be a claim. In addition, when answering this question, annotator should not open the tweet URL. Does the <tweet> contain a verifiable factual claim? Answer only with Yes or No. Examples: Tweet: Please don't take hydroxychloroquine (Plaquenil) plus Azithromycin for #COVID19 UNLESS your doctor prescribes it. Both drugs affect the QT interval of your heart and can lead to arrhythmias and sudden death, especially if you are tak ing other meds or have a heart condition. Label: Yes Explanation: There is a claim in the text. Tweet: Saw this on Facebook today and it's a must read for all those idiots clearing the shelves #coronavirus #toiletpapercrisis #auspol Labe</tweet>

Table 9: System prompts used for inference on the CLEF dataset for claim detection.

Level	Prompt
V1	It is important that a verifiable factual check-worthy claim be verified by a professional fact-checker, as the claim may cause harm to society, specific person(s), company(s), product(s), or some government entities. However, not all factual claims are important or worth fact-checking by a professional fact-checker, as this very time-consuming. Do you think that a professional fact-checker should verify the claim in the <tweet>? Labels: No, no need to check; No, too trivial to check; Yes, not urgent; Yes, very urgent. Decide on one label. Then, answer only with Yes or No.</tweet>
V2	It is important that a verifiable factual check-worthy claim be verified by a professional fact-checker, as the claim may cause harm to society, specific person(s), company(s), product(s), or some government entities. However, not all factual claims are important or worth fact-checking by a professional fact-checker, as this very time-consuming. Do you think that a professional fact-checker should verify the claim in the <tweet>? Labels: No, no need to check: the tweet does not need to be fact-checked, e.g., be- cause it is not interesting, a joke, or does not contain any claim. No, too trivial to check: the tweet is worth fact-checking, how ever, this does not require a professional fact-checker, i.e., a non-expert might be able to fact-check the claim. For example, one can verify the information using reli able sources such as the official website of the WHO, etc. An example of a claim is as follows: "The GDP of the USA grew by 50% last year." Yes, not urgent: the tweet should be fact-checked by a professional fact-checker, however, this is not urgent or critical; Yes, very urgent: the tweet can cause immediate harm to a large number of people; therefore, it should be verified as soon as possible by a professional fact-checker; Decide on one label. Then, answer only with Yes or No.</tweet>
V3	It is important to verify a factual claim by a professional fact-checker, which can cause harm to the society, specific person(s), company(s), product(s) or government entities. However, not all factual claims are important or worthwhile to be fact-checked by a professional fact-checker as it is a time-consuming procedure. Do you think that a professional fact-checker should verify the claim in the <tweet>? Labels: No, no need to check: the tweet does not need to be fact-checked, e.g., be- cause it is not interesting, a joke, or does not contain any claim. No, too trivial to check: the tweet is worth fact-checking, however, this does not require a professional fact-checker, i.e., a non-expert might be able to fact-check the claim. For example, one can verify the information using reliable sources such as the official website of the WHO, etc. An example of a claim is as follows: "The GDP of the USA grew by 50Yes, not urgent: the tweet should be fact-checked by a professional fact-checker, however, this is not urgent or critical; Yes, very urgent: the tweet can cause immediate harm to a large number of people; therefore, it should be verified as soon as possible by a professional fact-checker; Examples: Tweet: Wash your hands like you've been chopping jalapeños and need to change a contact lens" says BC Public Health Officer Dr. Bonnie Henry re. ways to protect against #coronavirus #Covid_19 Label: Yes, not urgent Explanation: Overall it is less important for a professional fact-checker to verify this information. The statement does not harm anyone. The truth value of whether the official said the statement is not important. Also it appears that washing hands is very important to protect onself from the virus. Tweet: ALERT! The corona virus can be spread through internationaly printed albums. If you have any albums at home, put on some gloves, put all the albums in a box and put it outside the front door tonight. I'm collect ing all the boxes tonight for safety. Think of your health. Label: No, no need to check E</tweet>

Table 10: System prompts used for inference on the CLEF dataset for claim check-worthiness detection.

Level	Prompt
V1	The task is to select verifiable statements from political speeches for fact-checking. Given a <statement> from a political speech, answer the question. Does the <statement> explicitly present any verifiable factual information? Answer with A, B or C only. A - Yes, B - Maybe, C - No.</statement></statement>
V2	The task is to select verifiable statements from political speeches for fact-checking. Given a <statement> from a political speech, answer the question following the guidelines. Definitions and guidelines: Fact: A fact is a statement or assertion that can be objectively verified as true or false based on empirical evidence or reality. Opinion: An opinion is a judgment based on facts, an attempt to draw a reasonable conclusion from factual evidence. While the underlying facts can be verified, the derived opinion remains subjective and is not universally verifi able. Context: Make sure to consider a small context of the target statement (the previous and next sentence) when annotating. Some statements require context to un derstand the meaning. Factual claim: A factual claim is a statement that explicitly presents some verifiable facts. Statements with subjective components like opinions can also be factual claims if they explicitly present objectively verifiable facts. Opinion with Facts: Opinions can also be based on factual information. When does an opinion explicitly present a fact: Many opinions are more or less based on some factual information. However, some facts are explicitly presented by the speakers, while others are not. What is verifiable: The verifiability of the factual inform tion depends on how specific it is. If there is enough specific information to guide a general fact-checker in checking it, the factual information is verifiable. Other wise, it is not verifiable. The question: Does the <statement> explicitly present any verifiable factual information? Answer with A, B or C only. A - Yes, the statement contains factual information with enough specific details that a fact-checker knows how to verify it. E.g., Birmingham is small in population compared to London. B - Maybe, the statement seems to contain some factual information. However, there are certain ambiguities (e.g., lack of specificity) making it hard to determine the verifiability. E.g., Birmingham is small compared to London. (lack</statement></statement>

Table 11: System prompts of Level V1 and Level V2 used for inference on the PoliClaim dataset for claim checkworthiness detection. The blue highlight shows instructions for regarding context.

#### Level

V3

#### Prompt

The task is to select verifiable statements from political speeches for fact-checking. Given a statement from a political speech and its context, answer the question following the quidelines. Definitions and quidelines: Fact: A fact is a statement or assertion that can be objectively verified as true or false based on empirical evidence or reality. Opinion: An opinion is a judgment based on facts, an attempt to draw a reasonable conclusion from factual evidence. While the underlying facts can be verified, the derived opinion remains subjective and is not universally verifiable. Factual claim: A factual claim is a statement that explicitly presents some verifiable facts. Statements with subjective components like opinions can also be factual claims if they explicitly present objectively verifiable facts. Context: Make sure to consider a small context of the target statement (the previous and next sentence) when annotating. Some statements require context to understand the meaning. For example: E1. "... Just consider what we did last year for the middle class in California, sending 12 billion dollars back - the largest state tax rebate in American history. <statement> But we didn't stop there. <> We raised the mini mum wage. We increased paid sick leave. Provided more paid family leave. Expanded child care to help working parents..." Without the context, the sentence marked wi <statement> seems an incomplete sentence. With the context, we know the speaker is claiming a bunch of verifiable achievements of their administration. E2. "... Wher I first stood before this chamber three years ago, I declared war on criminals and asked for the Legislature to repeal and replace the catch-and-release policies in  $\overline{\mathrm{SI}}$ 91. <statement> With the help of many of you, we got it done. <> Policies do matter We've seen our overall crime rate decline by 10 percent in 2019 and another 18.5 pe cent in 2020!  $\ldots$  " The part marked with <statement> claims that the policies against crimes have been "done", which is verifiable. It needs context to understand it. Opinion with Facts: Opinions can also be based on factual information. For example: E1. "I am proud to report that on top of the local improvements, the state has ad ministered projects in almost all 67 counties already, and like I said, we've only just begun." The speaker's "proud of" is a subjective opinion. However, the content of pride (administered projects) is factual information. E2. "I first want to thank my wife of 34 years, First Lady Rose Dunleavy." The speaker expresses their thank fulness to their wife. However, there is factual information about the first lady's name and the length of their marriage. When does an opinion explicitly present a fact: Many opinions are more or less based on some factual information. However, some facts are explicitly presented by the speakers, while others are not. Explicit presentation means the fact is directly entailed by the opinion without extrapolation: E1. "The pizza is delicious." This opinion seems to be based on the fact that "pizza is a kind of food". However, this fact is not explicitly presented. E2. "I first want to thank my wife of 34 years, First Lady Rose Dunleavy." The name of the speaker's wife and their year of marriage are explicitly presented. What is verifiable: The verifiability of the factual information depends on how spe cific it is. If there is enough specific information to guide a general fact-checker in checking it, the factual information is verifiable. Otherwise, it is not veri fiable. E1. "Birmingham is small." is not verifiable because it lacks any specific information for determining veracity. It leans more toward subjective opinion. E2. "Birmingham is small, compared to London" is more verifiable than E1. A fact-checker can retrieve the city size, population size ...etc., of London and Birmingham to com pare them. However, what to compare to prove Birmingham's "small" is not specific enough. E3. "Birmingham is small in population size, compared to London" is more ver ifiable than E1 and E2. A fact-checker now knows it is exactly the population size to be compared. The question: Does the <statement> explicitly present any verifiable factual information? Answer with A, B or C only. A - Yes, the statement contains factual information with enough specific details that a fact-checker knows how to verify it. E.g., Birmingham is small in population compared to London. B - Maybe, Maybe, the statement seems to contain some factual information. However, there are certain ambiguities (e.g., lack of specificity) making it hard to determine the verifiability. E.g., Birmingham is small compared to London. (lack of details about what standard Birmingham is small) C - No, the statement contains no verifiable factual information. Even if there is some, it is clearly unverifiable. E.g., Birmingham is small.

Table 12: System prompts of Level V3 used for inference on the PoliClaim dataset for claim check-worthiness detection. The blue highlight shows instructions for regarding context.