Par-ITA: Benchmarking Seq2Seq and LLMs on a Human-Supervised Parallel Corpus for Italian Hyperpartisan Neutralization

Michele Joshua Maggini[1], Søren Fomsgaard[2], Michele Maestroni, Gaël Dias[2], Pablo Gamallo[1]
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes da USC,
Universidade de Santiago de Compostela,
[2]UNICAEN, ENSICAEN, CNRS, GREYC, Normandie Univ,
GREYC UMR 6072, F-14000 Caen, France
Correspondence: name.surname@usc.es

Abstract

Neutralizing hyperpartisan content is essential for mitigating online polarization, yet research has largely focused on English. We present Par-ITA, a curated subset from Semeval 2023 task 3, consisting in the first human-supervised parallel corpus for Italian hyperpartisan neutralization of 2,475 paragraph pairs. The dataset is constructed using a rigorous three-stage pipeline: (1) expert-led preliminary selection of LLMs for high-quality generation, (2) human-supervised data production with high editing rates (32–68%), and (3) post-hoc human validation. We establish extensive benchmarks for this task across seq2seq and decoder-only architectures, evaluating standard fine-tuning, Direct Preference Optimization (DPO), and in-context learning. Our analysis highlights that while DPO effectively maximizes neutrality scores in seq2seq models, automated evaluators like GPT-4o-mini exhibit systematic biases, specifically over-penalizing sensitive political topics compared to human experts. Par-ITA provides a foundational resource for non-English neutralization and a reproducible framework for developing high-quality datasets in subjective domains.

1 Introduction

The spread of hyperpartisan content threatens democratic societies by polarizing public debate (Brüggemann and Meyer, 2023). This is particularly acute in discussions on immigration and climate change, where biased framing shapes public opinion and deepens social divisions (Marino et al., 2024; Iannelli et al., 2021). Hyperpartisan news reinforces echo chambers through emotional language and extreme viewpoints (Ross Arguedas et al., 2022). Despite its impact, research gaps remain: most datasets are English-centric (Maggini et al., 2025c), and work on *debiasing* hyperpartisan content in low-resource languages like Italian is scarce. While studies exist for general sentence debiasing (Pryzant et al., 2020) or text detoxification (Logacheva et al., 2022), none address Italian hyperpartisan neutralization. To bridge this gap, we introduce PAR-ITA, a parallel corpus of 2,475 Italian paragraph pairs for hyperpartisan neutralization entirely curated by humans. We develop the dataset through a rigorous three-stage pipeline (Figure 1): **(1) Preliminary Selection:** evaluating open-source LLMs across syntax (SY) and semantics (SE) to identify the most reliable generator; **(2) Supervised Generation:** using the selected model to generate neutralized and hyperpartisan variants with human-in-the-loop editing; and **(3) Post-hoc Validation:** verifying LLMs neutralization capabilities via expert evaluation. Throughout this process, we also assess the viability of GPT-4o-mini as an automated judge during the three-steps annotation pipeline (prompts are available in Appendix F.1). We establish comprehensive benchmarks for Italian neutralization, evaluating both seq2seq architectures (e.g., IT5, BART) and LLMs. Our experiments compare standard fine-tuning (FT), Direct Preference Optimization (DPO) (Rafailov et al., 2024), and In-Context Learning (ICL). Our human evaluation protocol assesses both linguistic quality and the alignment between human judgments and automated metrics. Our main contributions are: **(i)** PAR-ITA, the first parallel corpus for Italian hyperpartisan neutralization; **(ii)** a three-stage annotation methodology integrating preliminary selection, human-supervised generation, and post-hoc validation; and **(iii)** a systematic benchmark of training strategies (FT, DPO, ICL) providing insights into architecture-specific trade-offs. Our files are publicly available[1]

[1]The SemEval 2023 Task 3 dataset is available at https://propaganda.math.unipd.it/semeval2023task3/. Due to copyright restrictions and the Terms of Service of the original news outlets, our neutralized versions are available upon request for research purposes only. Our code is available at https://github.com/MichJoM/italian-hyperpartisan-neutralization/

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5592–5615
July 2-7, 2026

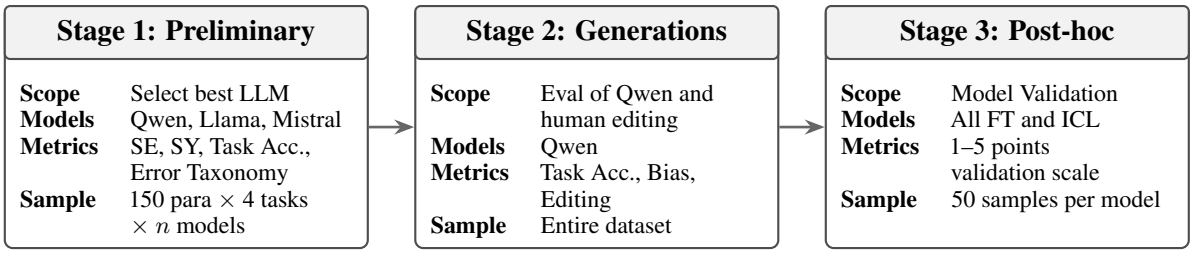

Figure 1: The three-stage annotation and evaluation pipeline.

2 Related Work

Neutralization of Biased Texts Text debiasing and neutralization strategies range from fine-grained classification (Maggini et al., 2025a) and counter-example generation (Sermsri and Panboonyuen, 2025) to reinforcement learning using embedding or classifier rewards (Liu et al., 2021a,b). Recently, LLM-based approaches have dominated the field (Fernández et al., 2024; Juvino Santos et al., 2025; Dale et al., 2021; Sermsri and Panboonyuen, 2025). Political neutralization is closely related to Text-Style Transfer (TST), typically involving word removal or synonym replacement (De Ruvo et al., 2025). However, political discourse involves complex linguistic layers—including semantics, syntax, and pragmatics (Maggini et al., 2025a)—and rhetorical strategies like irony or cherry-picking (Maggini et al., 2025b; de León et al., 2024). Despite progress in detoxification and depolarization (Toshevska and Gievska, 2025; Dementieva et al., 2024b), the task is hindered by resource scarcity. Existing parallel datasets include the Wiki Neutrality Corpus (Pryzant et al., 2020), PARADETOX for detoxification (Logacheva et al., 2022), and Reddit-based offensive language corpora (Atwell et al., 2022). While multilingual efforts like MultiParaDetox (Dementieva et al., 2024a) aim to automate collection, Italian remains under-represented. Aside from DETOXIFY-IT for toxic comments (De Ruvo et al., 2025), Par-ITA is the first humanly curated parallel corpus for Italian hyperpartisan neutralization.

LLMs as Annotators In political science, LLMs are increasingly used to scale supervised machine learning while reducing human labeling costs (Horych et al., 2025; Stromer-Galley et al., 2025). Törnberg (2024) found that LLMs can outperform experts in classifying political affiliations across 11 languages, though performance may be inflated by data leakage (Sainz et al., 2023). Nevertheless, GPT-4 still struggles with complex ideological detection and long-form articles, particularly in non-English contexts (Heseltine and von Hohenberg, 2024). Furthermore, deployment choices can introduce systematic biases or inaccuracies (Baumann et al., 2025), suggesting that integrating sociodemographic details may be necessary to improve performance in interpretive NLP tasks (Beck et al., 2024).

3 Dataset Construction

3.1 Data Collection and Preparation

In order to build a hyperpartisan parallel corpus, we selected Semeval 2023 task 3 dataset (Piskorski et al., 2023) tailored for detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup.

3.2 Annotation Protocol

After filtering Italian data from the corpus, two native speakers (a Ph.D. student in Disinformation and an expert journalist) annotated paragraphs as hyperpartisan or neutral. Hyperpartisanship indicates one-sided or ideologically extreme views expressed through irony, sarcasm, framing bias, or cherry-picking. The annotation process involved training on guidelines, pilot annotations, and interactive sessions to refine edge cases. Independent annotation achieved 0.85 Cohen's κ, with disagreements resolved through discussion and third-party mediation, yielding perfect final agreement on 825 hyperpartisan paragraphs. The entire process required around 300 hours across three custom web interfaces (see Figures 7, 8, 9 in Appendix A.2).

3.3 Three-Stage Annotation Pipeline

Figure 1 illustrates our three-stage approach: (1) preliminary LLM selection, (2) full dataset generation with human supervision, and (3) post-hoc model validation.

tree/master.

Type	Task	Text ITA	Text ENG
HP	O	L'Italia si barcamena con Di Maio...	Italy is muddling through with Di Maio...
HP	Rep	l'Italia cammina sul filo del rasoio con Di Maio...	Italy is walking on the razor's edge with Di Maio...
HP	Per	l'Italia naviga a vista con Di Maio...	Italy is navigating by sight with Di Maio...
N	Rew	L'Italia fatica a gestire la situazione con Di Maio.	Italy is struggling to manage the situation with Di Maio.

Table 1: Examples of model-generated and human-supervised neutral rewrites for hyperpartisan input phrases. O: original text, Rep: Rephrased, Per: Perturbed, Rew: Rewritten

3.3.1 Stage 1: Preliminary LLM Selection

To select the optimal model for data augmentation, we evaluated three LLMs: Mistral-Nemo-2407 (Mistral, 2024), Llama-3.1-8b-Instruct (Touvron et al., 2023), and Qwen2.5-14B-Instruct (Yang et al., 2025), chosen for their open availability and balanced efficiency. These models generated three variants per paragraph via backtranslation (ita→eng→ita): one neutralized (*Rewritten*) and two hyperpartisan versions (*Rephrased* with synonym substitution, *Perturbed* focusing on changing only loaded language and rhetoric).

Evaluation Protocol Annotators independently assessed 150 sampled paragraphs on 3-point scales for **SY** (1: Major Errors, 2: Minor Errors, 3: Correct) and **SE** (1: Poor Preservation, 2: Partial Preservation, 3: Full Preservation), plus binary neutrality classification. We applied a multi-dimensional error taxonomy (Appendix A.3) categorizing failures in neutralization, factuality, fluency, and structure. Additionally, GPT-4o-mini served as an automated judge being prompted with the guidelines and the taxonomy.

Results Human annotators achieved high agreement (Table 2): Krippendorff's α of 0.890 (SY), 0.863 (SE), and 0.815 (Neutrality). Model performance varied significantly (for more details see Appendix A.4). Llama-3.1-8b-Instruct excelled in SY for paraphrasing tasks (2.88 Rephrased, 2.54 Perturbed) but frequently generated English text, yielding poor semantic scores. Mistral-Nemo-Instruct-2407 struggled with neutralization, retaining hyperpartisan cues in 70% of outputs. Qwen-2.5-14B-Instruct demonstrated consistent quality across all tasks (Rewritten: 2.90 SY, 2.73 SE) with only 13% neutralization failures, achieving the following human alignment (Cohen's κ = 0.801, MCC = 0.806).

GPT-4o-mini as Judge GPT-4o-mini showed moderate correlation with human consensus on quality scores (Figure 2), particularly diverging on Rewritten texts where humans assigned higher ratings. Considering the agreement between humans and GPT on the neutralization task (Figure 3), it ranged from poor to fair (Cohen's κ: -0.039 to 0.454), with systematic over-classification of hyperpartisanship on sensitive topics (COVID, immigration, climate change). This aligns with documented "safety over-refusal" in RLHF-tuned models (Röttger et al., 2024; Santurkar et al., 2023), which rely on lexical triggers rather than nuanced framing analysis.

Justification for LLM-Based Data Generation. While comprehensive benchmarking (Section 4)

	α	κ_w	r	ρ	κ_c	MCC
Syntax	.890	.890	.891	.902	-	-
Semantics	.863	.863	.865	.871	-	-
Neutrality	.815	.815	.815	.812	-	-
Best Model	-	-	-	-	.871	-
Llama Rew.	-	-	-	-	.834	.840
Mistral Rew.	-	-	-	-	.762	.764
Qwen Rew.	-	-	-	-	.801	.806

Table 2: Inter-annotator agreement for human experts and model-human alignment on neutrality classification. α: Krippendorff's α, κ_w: weighted Cohen's κ, r: Pearson correlation, ρ: Spearman correlation, κ_c: Cohen's κ, MCC: Matthews Correlation Coefficient.

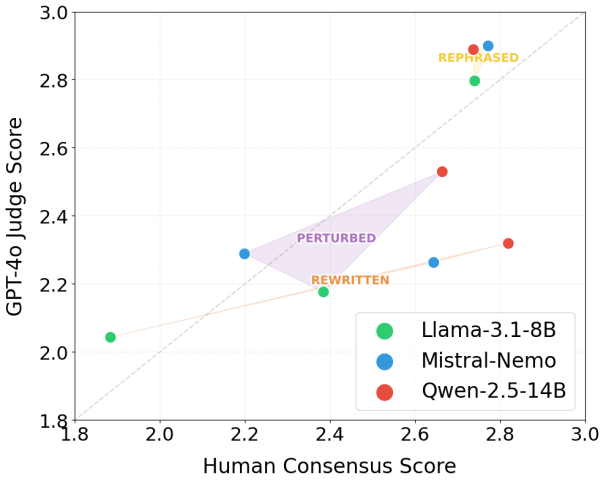

Figure 2: Correlation between human consensus and GPT-4o-mini scores (averaged SY and SE).

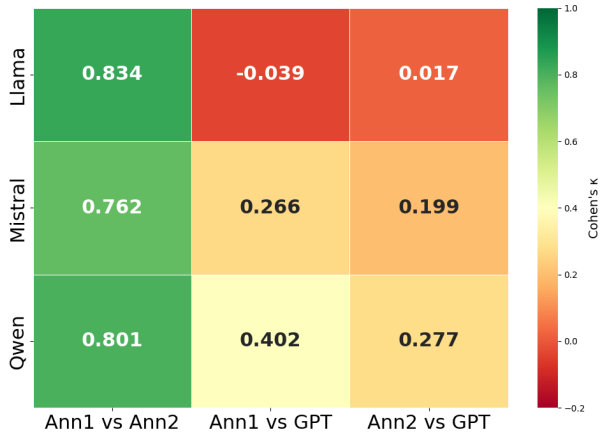

Figure 3: Cohen's k between human experts and GPT-4o-mini on binary label for the Rewritten task across models.

reveals seq2seq models achieve superior performance, we selected Qwen-2.5-14B-Instruct for data generation based on Stage 1 evaluation comparing only LLMs. This decision reflects three considerations: (1) Stage 1 established Qwen as best among LLMs for generation quality and human alignment; (2) High editing rates (32-68%, Table 3) mean final quality depends on expert supervision rather than raw model output: Levenshtein distances of 25-32% confirm substantial human refinement; (3) At experimentation time, we prioritized selecting a versatile model capable of both neutralization and bias amplification tasks.

3.3.2 Stage 2: Full Dataset Generation with Human Supervision

Based on Stage 1 results, we selected Qwen-2.5-14B-Instruct to generate variants for all 825 hyperpartisan paragraphs. At the time of experimentation, we wanted to assess the capabilities of the selected model to accomplish the task and use it as a tool to speed up the writing process for the data augmentation. Annotators evaluated each output using a 3-level task accomplishment scale (No/Partially/Yes) and performed necessary editing. Indeed, Qwen successfully generated hyperpartisan variants in most cases (Table 3), but achieved only 44.4% complete success on Rewritten. Consequently, annotators edited 32.8% of Rephrased texts, 41.9% of Perturbed texts, and 67.8% of Rewritten texts, with Avg. Levenshtein distances of 29%, 25%, and 32% respectively. Failed generations were collaboratively rewritten to ensure corpus quality.

Task	Edited	Lev.	N	P	Y
Rephrased	32.8%	29%	12.4%	11.6%	76.0%
Perturbed	41.9%	25%	17.5%	8.1%	74.4%
Rewritten	67.8%	32%	27.9%	27.8%	44.4%

Table 3: Generation quality. Lev.: Avg. Levenshtein distance from original text. N: No, P: Partially, Y:Yes.

Human-GPT Agreement on Qwen's Rewritten Task For Rewritten texts, human annotators achieved substantial agreement (Cohen's κ = 0.694), while GPT-4o-mini agreement remained fair (κ = 0.302–0.454), confirming its tendency to over-classify contentious topics as hyperpartisan regardless of actual framing bias.

Common Error Patterns *Rephrased*: Difficulty replacing neologisms and managing noisy elements (calls-to-action, malformed punctuation). *Perturbed*: While successfully amplifying bias markers, the model occasionally introduced grammatical errors, nonsense neologisms, or output placeholder text instead of content. *Rewritten*: Most failures involved incomplete neutralization (residual loaded terms, preserved biased framing) or difficulty detecting subtle irony. To preserve semantic meaning, we avoided complete context removal except when sentences contained only insults or slurs, prioritizing reduction of loaded language over topic elimination. Additional analysis on the generated vs. edited texts can be found in Appendix B.

The final corpus comprises 1,986 training pairs and 489 test pairs (see Table 4), ensuring high-quality parallel data for neutralization tasks.

Metric	Train	Test	Total
# Pairs	1,986	489	2,475
# Unique Originals	662	163	825
Avg. input chars	365.7	377.5	368.0
Avg. input words	55.3	57.2	55.6
Avg. output chars	352.9	402.0	362.6
Avg. output words	51.3	59.2	52.9
Vocabulary size	15,639	6,760	18,032

Table 4: Statistics of the human-curated Par-ITA corpus.

3.3.3 Stage 3: Post-hoc Model Validation

Following model training, we conducted human evaluation to assess generation quality beyond automated metrics. Two expert annotators independently rated 50 randomly sampled outputs per

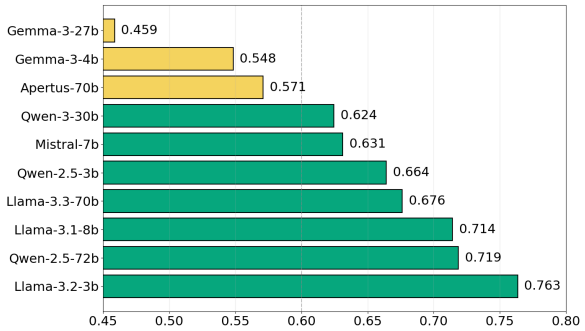

Figure 4: Inter-annotator agreement (Cohen's k) between humansacross evaluated LLMs. Higher values indicate more consistent human judgments. Phi-3.5 has not been included in the plot since the agreement was perfect because the model outputted no-sense text.

model on a 5-point scale (Figure 13), evaluating both neutralization success and semantic preservation. GPT-4o-mini provided parallel judgments while being prompted with the task and the taxonomy.

Inter-Annotator Agreement on LLMs Generations Post-Hoc Human annotators achieved substantial-to-strong agreement across models (Figure 4), with Cohen's κ ranging from 0.459 (gemma-3-27b) to 0.763 (llama-3.2-3b). Higher agreement on models like llama-3.2-3b (κ=0.763) and qwen-25-72b (κ=0.719) suggests these models produce more consistent, interpretable outputs. Lower agreement on larger models like gemma-3-27b and apertus-70b indicates greater output variability (e.g.: instructions partially followed, repeated outputs).

Human vs. GPT-4o-mini Rating Alignment Figure 13 in Appendix D reveals systematic differences in rating patterns. GPT-4o-mini consistently assigned higher scores (concentrated in ratings 3-5), while human annotators distributed ratings more evenly, particularly using lower scores (1-2) to penalize subtle neutralization failures, semantic drift, or hallucinations. This pattern confirms GPT-4o-mini's tendency toward leniency and insufficient sensitivity to residual bias or factual inconsistencies that humans readily detect.

Implications for Automated Evaluation These findings reinforce conclusions from earlier stages: while GPT-4o-mini provides useful signal for high-level quality assessment, it cannot replace human judgment for nuanced neutralization tasks. The divergence between high automated semantic similarity scores (see Appendix B) and lower human ratings highlights that lexical/embedding similarity does not capture subtle hyperpartisan cues, framing bias, or hallucinated content. We therefore prioritize human evaluation as the gold standard for model comparison.

4 Benchmark Experiments: Setup and Results

4.1 Models

Our experimental setup combines various seq2seq models in both base and large sizes: BART (Lewis et al., 2020), IT5 (Sarti and Nissim, 2024), FLAN-T5, (Chung et al., 2024), and mT5 (Xue et al., 2021); and instructed LLMs loaded through `unsloth` library: Gemma-3-4b-it (Team et al., 2025), Phi-3.5-mini-instruct (Abdin et al., 2024), Qwen2.5-3b-instruct, Mistral-7b-it-v0.3, Llama-3.2-3b-instruct/-3.1-8b-instruct, and with `vllm`: Apertus-70B-2509 (Apertus et al., 2025), Gemma-3-27b-Instruct/-2.5-72B-Instruct, Qwen-3-30B-Instruct.

4.2 Guided SFT Classifier

In this approach, we augment standard maximum likelihood estimation (MLE) with an auxiliary loss derived from a fine-tuned binary neutrality classifier for italian hyperpartisan paragraph detection on Maggini et al. (2025a)'s dataset, consisting of 1,010 paragraphs. During training, the seq2seq model generates outputs via standard teacher forcing, which are then periodically (every n steps) decoded to text and scored by the frozen classifier. The classifier produces neutrality probabilities p_{neutral} for each generated sequence, which we convert into a guidance loss term:

$$\mathcal{L}_{\text{guide}} = \lambda \cdot \max(0, \tau - p_{\text{neutral}}) \quad (1)$$

where τ is a target neutrality rate and λ is a weighting coefficient. The total training objective then becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MLE}} + \mathcal{L}_{\text{guide}}. \quad (2)$$

4.3 DPO

Since for each task we had a preferred and optimal variant, the direct and valid output of Qwen, or the human rewritten version, and a underoptimal one, namely the failed generation or a corrupted version of it, we crafted a DPO dataset, too. We prioritized the human edited texts as the preferred

version. While when we kept the Qwen's generated text as the preferred option, we corrupted it to generate the suboptimal choice with the following methods: (i) swapping the order of the words, (ii) inserting loaded adverbs like "ovviamente", "incredibilmente", (iii) random truncation, (iv) word repetition, (v) random masking and (v) random word deletion.

4.4 Seq2Seq Model Results

We evaluate four seq2seq model families, IT5, mT5, FLAN-T5, and BART-IT across three training configurations: standard supervised fine-tuning (SFT), classifier-guided fine-tuning (SFT-CLASS), and Direct Preference Optimization (DPO). Table 5 presents comprehensive results across all metrics (further analysis is available in Appendix E).

Training Configuration Analysis. We observe three distinct patterns. Firstly, **SFT** provides balanced performance with high semantic preservation ($\text{SBERT}_{\text{src-gen}} > 0.83$) but moderate BLEU scores (14–19). Moreover, **SFT-CLASS** marginally improves neutrality (+2–5%) at slight cost to fluency (–1–3% BLEU), confirming that classifier-guided decoding offers modest but consistent neutrality gains. Lastly, **DPO** achieves the highest BLEU scores (+23% average improvement) but with variable neutrality outcomes.

DPO Performance. DPO significantly improves generation quality for seq2seq models. BART-base DPO achieves 29.1 BLEU—the highest across all configurations—while FLAN-T5-large DPO maintains both high quality (24.1 BLEU) and acceptable neutrality (0.73). However, DPO shows higher variance in neutrality outcomes: while some models maintain high neutrality (IT5-large DPO: 0.73, mT5-base DPO: 0.72), others exhibit significant degradation (BART-base DPO: 0.13).

Model Family Comparison. Among model families, FLAN-T5 demonstrates the most consistent performance across all configurations, benefiting from its instruction-tuning pretraining. BART-IT achieves the highest peak BLEU but with unstable neutrality. mT5 provides a balanced middle ground. IT5 shows the weakest overall performance despite being Italian-native: IT5-base achieves only 11.9–12.2 BLEU with lower neutrality (0.52–0.53), while IT5-large performs slightly better (13.6–13.7 BLEU). This suggests that domain-specific pretraining does not guarantee task-specific performance, and cross-lingual models (mT5, FLAN-T5) may better generalize to the neutralization task.

Semantic Preservation. The $\text{SBERT}_{\text{src-gen}}$ metric reveals an important trade-off: SFT models preserve source semantics best (0.83–0.91), while DPO models show moderate semantic drift (0.62–0.75). This suggests that preference optimization encourages more extensive rewriting, which increases BLEU but may alter original meaning beyond neutralization requirements.

4.5 LLM Results

We evaluate 11 LLMs across two configurations: zero-shot prompting for large models (27B–72B parameters) and supervised fine-tuning (FT) using LoRA and unsloth for smaller models (3B–8B). See Section F for detailed configuration descriptions. Table 6 summarizes performance across both automatic metrics and human evaluation.

Zero-shot Performance LLMs demonstrate zero-shot neutralization capabilities, achieving near-perfect neutrality rates (0.94–1.00) with simple prompting, at least according to our classifier. Qwen2.5-72B leads in automatic metrics with 12.8 BLEU, followed by Qwen3-30B (9.4) and Gemma3-27B (8.1). However, human evaluation reveals a different picture: Gemma3-27B ranks highest (3.96/5) despite lower BLEU, suggesting qualitative factors beyond surface-level similarity matter for perceived neutralization quality.

Fine-tuning Fine-tuning smaller models (3B–8B) with LoRA and unsloth substantially improves alignment with reference neutralizations. Llama3.1-8B FT achieves the highest BLEU (16.6) and maintains excellent neutrality (0.92), representing a +29% improvement over the best zero-shot model in BLEU. Mistral-7B FT performs comparably (16.3 BLEU, 0.91 neutrality) and achieves the best human rating among FT models (3.24/5). The fine-tuning process improves both fluency metrics and semantic preservation ($\text{SBERT}_{\text{src-gen}}$: 0.70–0.76) while incurring only modest neutrality degradation (–8% average vs. zero-shot).

Notable Failure Case. Phi3.5-Mini FT represents a complete failure case, producing incoherent outputs with near-zero BLEU (0.3) and negligible neutrality (0.03). All 50 human annotated samples received the lowest possible rating (1/5). This failure is attributed to the model's limited capacity and potential instruction-following limitations,

Model	Config	BLEU	ROUGE-L	BERTScore	Neutrality	$\text{SBERT}_{\text{src-gen}}$	$\text{SBERT}_{\text{gen-ref}}$
IT5-base	SFT	11.9	0.32	**0.74**	0.52	0.83	0.69
	SFT-CLASS	12.2	0.32	**0.74**	0.53	0.83	0.69
	DPO	14.0	0.39	0.35	0.57	0.62	0.64
IT5-large	SFT	13.7	0.33	0.35	0.59	0.83	0.70
	SFT-CLASS	13.6	0.34	0.35	0.62	0.84	0.70
	DPO	4.3	0.15	-0.01	0.73	0.55	0.43
mT5-base	SFT	14.3	0.34	0.39	0.72	0.85	0.70
	SFT-CLASS	14.5	0.34	0.39	0.71	0.86	0.70
	DPO	23.0	0.39	0.41	0.72	0.72	0.68
mT5-large	SFT	14.9	0.35	0.41	0.67	0.88	0.72
	SFT-CLASS	15.8	0.36	0.41	0.69	0.88	0.72
	DPO	21.1	0.37	0.36	0.57	0.75	0.69
FLAN-T5-base	SFT	18.4	0.38	0.42	0.83	0.91	0.73
	SFT-CLASS	17.5	0.44	0.26	**0.85**	0.79	0.63
	DPO	22.7	0.38	0.41	0.71	0.68	0.65
FLAN-T5-large	SFT	19.3	0.38	0.43	0.81	0.91	0.73
	SFT-CLASS	18.8	0.38	0.43	0.79	0.90	0.74
	DPO	24.1	0.43	0.47	0.73	0.69	0.73
BART-base	SFT	18.1	0.37	0.73	0.33	0.80	0.62
	SFT-CLASS	16.4	0.36	0.20	0.73	0.81	0.62
	DPO	**29.1**	**0.44**	0.44	0.13	0.75	0.73
BART-large	SFT	18.1	0.37	0.40	0.75	0.91	0.72
	SFT-CLASS	18.0	0.38	0.44	**0.85**	**0.92**	**0.75**
	DPO	24.0	0.39	0.45	0.36	0.62	0.70

Table 5: Seq2seq model performance on Par-ITA test set. Best values per column are **bolded**. $\text{SBERT}_{\text{src-gen}}$ measures semantic preservation from source (*how much the model preserved the original meaning*); $\text{SBERT}_{\text{gen-ref}}$ measures alignment with neutral reference (*how close the model's output is to the human-written neutral gold standard*). Neutrality is measured by our fine-tuned BERT classifier.

making it unsuitable for the neutralization task. Based on our comprehensive evaluation combining automatic metrics and human judgment: Gemma3-27B 0-shot is the best overall according to human preference, Mistral-7B FT or Llama3.1-8B FT (high BLEU >16, 0.91–0.92 neutrality, 3.2/5 human) provided the highest quality-neutrality balance, while Qwen2.5-72B 0-shot (12.8 BLEU, 0.94 neutrality) though ranked 6th by humans scored the highest automatic metrics.

4.6 Human Evaluation and Metric Alignment

To validate automatic metrics and gain deeper insight into model behavior, we conducted human evaluation on a stratified sample of 50 examples across all 11 LLM configurations. This section analyzes the correlation between human judgments and automatic metrics, revealing both alignment and notable discrepancies.

Correlation Analysis Table 7 shows that semantic similarity best predicts human judgment: $\text{SBERT}_{\text{gen-ref}}$ achieves the strongest correlation ($r = 0.84$), suggesting humans value outputs that closely match reference neutralizations. Overlap metrics like ROUGE-L ($r = 0.82$) and BLEU ($r = 0.75$) show strong linear correlation with human ratings, validating their use as primary evaluation metrics. Conversely, neutrality rate is weakly aligned. Indeed, despite being the target objective, automatic neutrality classification shows only moderate correlation ($r = 0.70$) and poor rank correlation ($\rho = 0.35$, $p = 0.30$). However, when comparing automatic neutrality rates with human preferences, models achieve near-perfect neutrality (>0.98) but receive mediocre human ratings (see Table 8). The manual inspection of low-rated zero-shot outputs reveals several failure modes that the neutrality classifier does not capture: **Over-neutralization**: Models like Apertus-70B and Llama3.3-70B sometimes remove too much content, producing outputs that are technically neutral but lack the informational substance of the original. **Generic phrasing**: Zero-shot models often produce safe but uninformative rewrites that avoid partisan language by avoiding specificity entirely. **Language quality issues**: Some models (particularly Apertus-70B) generated the same neutral paragraph multiple time, which annotators penalized heavily. **Gemma3-27B** succeeds at zero-shot neutralization, achieving both high neutrality (1.00)

Model	Config	BLEU	ROUGE-L	BERTScore	Neutrality	$\text{SBERT}_{\text{src-gen}}$	Human	Rank
Large Models (0-shot prompting)								
Gemma3-27B	0-shot	8.1	0.25	0.34	**1.00**	0.62	**3.96±1.56**	**1**
Qwen3-30B	0-shot	9.4	0.26	0.33	0.98	0.63	3.42±1.69	2
Qwen2.5-72B	0-shot	**12.8**	**0.31**	**0.34**	0.94	**0.68**	3.10±1.61	6
Llama3.3-70B	0-shot	6.8	0.18	0.23	0.98	0.60	2.52±1.73	9
Apertus-70B	0-shot	1.5	0.06	0.01	**0.99**	0.51	1.64±1.21	10
Small Models (LoRA Fine-tuned)								
Mistral-7B	FT	16.3	0.37	0.48	0.91	0.76	3.24±1.59	3
Llama3.1-8B	FT	**16.6**	**0.38**	**0.48**	0.92	**0.76**	3.20±1.54	4
Llama3.2-3B	FT	13.3	0.33	0.42	0.91	0.72	3.10±1.56	5
Gemma3-4B	FT	11.0	0.26	−0.16	0.89	0.56	3.00±1.60	7
Qwen2.5-3B	FT	10.8	0.29	0.39	0.92	0.70	2.84±1.43	8
Phi3.5-Mini	FT	0.3	0.03	−0.33	0.03	0.25	1.00±0.00	11

Table 6: LLM performance on Par-ITA test set. **Human** column shows mean±std human ratings (1–5 scale, n=50 per model). **Rank** is determined by human evaluation. Best values per column within each configuration group are **bolded**. Phi3.5-Mini produces degenerate outputs.

Metric	Pearson r	p	Spearman ρ	p
$\text{SBERT}_{\text{gen-ref}}$	+0.84	.001	+0.61	.048
ROUGE-L	+0.82	.002	+0.53	.091
$\text{SBERT}_{\text{src-gen}}$	+0.79	.004	+0.59	.055
BLEU	+0.75	.008	+0.54	.085
BERTScore	+0.75	.009	+0.57	.065
Neutrality	+0.70	.016	+0.35	.299

Table 7: Correlation between mean human ratings and automatic metrics across 11 LLM configurations. All Pearson correlations are significant at $p < 0.02$. Phi3.5-Mini was excluded from the correlation computation due to the degenerate outputs.

Model	Neutrality	Human	Gap
Gemma3-27B	1.00	3.96	aligned
Llama3.3-70B	0.98	2.52	−1.44
Apertus-70B	0.99	1.64	−2.32

Table 8: Discrepancy between neutrality rate and human rating for selected 0-shot models.

and the highest human rating (3.96/5). Analysis suggests Gemma3's instruction-following training and multilingual capabilities enable more nuanced rewrites that preserve content while removing bias.

4.7 Implications for Metric Selection

Our analysis suggests the following hierarchy for metric reliability in predicting human preferences: **Best predictors**: $\text{SBERT}_{\text{gen-ref}}$, ROUGE-L (strong linear and rank correlation); **Good predictors**: BLEU, BERTScore, $\text{SBERT}_{\text{src-gen}}$ (strong linear, moderate rank); **Weak predictor**: Neutrality rate alone (confounded by over-neutralization). For practical evaluation, we recommend using **BLEU + Neutrality jointly**, as high performance on both metrics is required for human acceptance. Models achieving >10 BLEU and >0.90 neutrality consistently receive human ratings ≥3.0/5.

Metric	Value
Cohen's κ (unweighted)	0.69
Cohen's κ (quadratic weighted)	0.71
Krippendorff's α	0.51
Pearson r	0.71
Spearman ρ	0.72
Mean Absolute Error	0.57

Table 9: Inter-annotator agreement between human annotator and GPT-4o-mini on LLM output evaluation (1–5 rating scale).

Human vs. GPT-4o-mini Agreement on LLMs Post-Hoc Evaluation We also collected GPT-4o-mini ratings for all samples to assess LLM-as-judge reliability, which we found being substantial. Table 9 presents inter-annotator agreement metrics between human and GPT-4o-mini annotations.

5 Conclusions and Future Work

This study introduces Par-ITA, the first human-curated dataset for italian hyperpartisan neutralization. Our work establishes a high-quality benchmark and demonstrates that FT and DPO enhance the ability of both seq2seq and LLM architectures to mitigate bias. However, our evaluation also reveals a significant gap between automated metrics and human judgment, particularly regarding the preservation of semantic nuances in complex political discourse. We further demonstrate that while small LLM-as-a-judge frameworks provide a scalable alternative to human annotation, they remain

susceptible to safety-driven over-refusals on controversial topics. We plan to extend DPO with LLMs.

6 Limitations

Despite the contributions of this work, several limitations remain. **Domain Specificity:** PAR-ITA focuses primarily on the domains of migration and climate change. Despite international organizations and institutions being cited, the main perspective of the news is from an italian point of view. While these are critical areas of hyperpartisanship in Italy, the linguistic strategies for neutralization may differ in other domains like healthcare or economics and countries. **Language Constraint:** Our study is limited to the Italian language. Although our three-stage methodology is transferable, the specific findings regarding model performance (e.g., the superiority of Qwen over Llama for Italian) may not generalize to other low-resource or Romance languages. **Metric Sensitivity:** automated metrics like BLEU and Neutrality scores do not fully capture the qualitative nuances of human preference. Our post-hoc human evaluation was conducted on a subset of the data (50 samples per model), which, while standard for the field, limits the statistical power of some qualitative comparisons. **LLM-as-a-Judge Bias:** For budget constraint we limited our experiment to GPT-4o-mini and open models. We also found that GPT-4o-mini exhibits systematic biases when evaluating sensitive political topics. Relying solely on automated judges for this task remains risky without human oversight. **Model Scale Constraints**: Our largest fine-tuned model was 8B parameters due to GPU memory constraints. Models >27B were evaluated only zero-shot, preventing fair comparison across all configurations. Future work with distributed training infrastructure could assess whether fine-tuning larger models closes the performance gap. **Dataset Size**: At 2,475 pairs, Par-ITA is relatively small compared to English resources (WNC: 180K pairs). While our high-quality human supervision ensures reliability, model performance might improve with larger training sets, particularly for data-hungry transformer architectures.

Ethical Statement

The development of neutralization tools carries inherent ethical risks. **Dual Use:** While our goal is to reduce polarization, the same techniques could theoretically be inverted to "partisan-ize" neutral text or to sanitize extremist rhetoric in a way that makes it more palatable to a mainstream audience. We release our dataset under a license that encourages research into transparency and depolarization. This paper presents a neutralization task intended as an invitation to journalists to adopt a professional language aligned with their deontology, we absolutely do not promote the misleading usage of this dataset to foster polarization or censorship. **Annotator Well-being:** Our annotators were exposed to hyperpartisan and potentially inflammatory content regarding migration and climate change. To mitigate harm, we utilized expert annotators who were briefed on the nature of the content, and we ensured the annotation sessions were structured to avoid fatigue and emotional distress. **Algorithmic Bias:** Neutralization is a subjective task. What one observer considers "neutral," another may consider "biased toward the center." We have addressed this by using a multi-stage validation process and providing a clear error taxonomy, but we acknowledge that our models may still reflect the subjective viewpoints of the experts involved in the supervision process. **Environmental Impact:** Training and fine-tuning the LLMs (ranging from 1B to 70B parameters) used in this study involves significant computational resources and carbon footprint. We have mitigated this by using parameter-efficient fine-tuning (PEFT) where possible and utilizing pre-trained open-source weights.

7 Acknowledgements

This paper was funded from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351 and from the UK Research and Innovation (UKRI) Horizon Europe funding guarantee - Grant Number: EP/X036758/1: HYBRIDS Project. It was also funded by MCIU/AEI (PID2024-161928OB-I00 and AIA2025-163322-C62) and by the Galician Government (Research Center of Galicia accreditation 2024-2027 ED431G-2023/04 and GPC ED431B 2025/16).

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110

others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, and 1 others. 2025. Apertus: Democratizing open and compliant llms for global language environments. *arXiv preprint arXiv:2509.14233*.

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. APPDIA: A discourse-aware transformer-based style transfer model for offensive social media conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6063–6074, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using llms for text annotation. *Preprint*, arXiv:2509.08825.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Michael Brüggemann and Hendrik Meyer. 2023. When debates break apart: Discursive polarization as a multi-dimensional divergence emerging in and through communication. *Communication Theory*, 33(2-3):132–142.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1).

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ernesto de León, Mykola Makhortykh, and Silke Adam. 2024. Hyperpartisan, alternative, and conspiracy media users: An anti-establishment portrait. *Political Communication*, 41(6):877–902.

Viola De Ruvo, Arianna Muti, Daryna Dementieva, and Debora Nozza. 2025. Detoxify-IT: An Italian parallel dataset for text detoxification. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 267–275, Vienna, Austria. Association for Computational Linguistics.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2024a. Multiparadetox: Extending text detoxification with parallel data to new languages. *arXiv preprint arXiv:2404.02037*.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Frolian Schneider, Xintog Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, and 1 others. 2024b. Overview of the multilingual text detoxification task at pan 2024. *Working Notes of CLEF*.

Miguel Fernández, Maximiliano Ojeda, Lilly Guevara, Diego Varela, Marcelo Mendoza, and Alberto Barrón-Cedeño. 2024. Victor vectors@ dipromats 2024: Propaganda detection with llm paraphrasing and machine translation. In *Proceedings of the Iberian languages evaluation forum (IberLEF 2024) co-located with the conference of the Spanish society for natural language processing (SEPLN 2024), Valladolid, Spain*, page 3756.

Michael Heseltine and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1):20531680241236239.

Tomáš Horych, Christoph Mandl, Terry Ruas, Andre Greiner-Petter, Bela Gipp, Akiko Aizawa, and Timo Spinde. 2025. The promises and pitfalls of LLM annotations in dataset labeling: a case study on media bias detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1370–1386, Albuquerque, New Mexico. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Laura Iannelli, Bianca Biagi, and Marta Meleddu. 2021. Public opinion polarization on immigration in italy: the role of traditional and digital news media practices. *The Communication Review*, 24(3):244–274.

Lucas Raniére Juvino Santos, Leandro Balby Marinho, Claudio Elizio Calazans Campelo, Filippo Menczer, and Alessandro Flammini. 2025. Can large language models effectively mitigate polarization in social media text? In *Proceedings of the 17th ACM Web Science Conference 2025*, Websci '25, page 348–357, New York, NY, USA. Association for Computing Machinery.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training

for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021a. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866.

Ruibo Liu, Lili Wang, Chenyan Jia, and Soroush Vosoughi. 2021b. Political depolarization of news articles using attribute-aware word embeddings. *CoRR*, abs/2101.01391.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Michele Joshua Maggini, Davide Bassi, and Pablo Gamallo. 2025a. Detecting hyperpartisanship and rhetorical bias in climate journalism: A sentence-level Italian dataset. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 168–187, Vienna, Austria. Association for Computational Linguistics.

Michele Joshua Maggini, Davide Bassi, Paloma Piot, Gaël Dias, and Pablo Gamallo Otero. 2025b. A systematic review of automated hyperpartisan news detection. *PloS one*, 20(2):e0316989.

Michele Joshua Maggini, Davide Bassi, Paloma Piot, Gaël Dias, and Pablo Gamallo Otero. 2025c. A systematic review of automated hyperpartisan news detection. *PLOS ONE*, 20(2):1–39.

Erik Bran Marino, Jesus M Benitez-Baleato, and Ana Sofia Ribeiro. 2024. The polarization loop: How emotions drive propagation of disinformation in online media—the case of conspiracy theories and extreme right movements in southern europe. *Social Sciences*, 13(11):603.

Mistral. 2024. Mistral nemo: Collaborative innovation with nvidia.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Amy Ross Arguedas, Craig Robertson, Richard Fletcher, and Rasmus Nielsen. 2022. Echo chambers, filter bubbles, and polarisation: A literature review.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Gabriele Sarti and Malvina Nissim. 2024. IT5: Text-to-text pretraining for Italian language understanding and generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9422–9433, Torino, Italia. ELRA and ICCL.

Kasidit Sermsri and Teerapong Panboonyuen. 2025. Debiasing large language models in thai political stance detection via counterfactual calibration. In *Proceedings of the 9th Widening NLP Workshop*, pages 56–64.

Jennifer Stromer-Galley, Brian McKernan, Saklain Zaman, Chinmay Maganur, and Sampada Regmi. 2025. The efficacy of large language models and crowd annotation for accurate content analysis of political social media messages. *Social Science Computer Review*, 0(0):08944393251334977.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Martina Toshevska and Sonja Gievska. 2025. Llm-based text style transfer: Have we taken a step forward? *IEEE Access*, 13:44707–44721.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Petter Törnberg. 2024. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 0(2024):08944393241286471.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 483–498.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Annotation and Dataset Details

A.1 Dataset Statistics

Figure 5 shows the token length and the TTR for each task. The token length remained almost the same across the task, while the Token Type Ratio (TTR) increased in the text variants, showing higher variability because the editing introduced different types of the tokens.

The dataset presents the following topic shown in Figure 6: Ukraine War, Immigration, Italian and European Politics, Health. The dataset has been split into 80%/20% for train and test with stratified sampling on the topics.

A.2 Custom Platforms

All the annotations except for the application of the Error Taxonomy with 1-5 point scales, have been conducted with Custom Platforms created with Claude Opus, ensuring they fit our guidelines and annotation scopes. The first platform, used to validate the best LLM for generation, masked model names to avoid source bias. The second platform allowed annotators to read and edit the generated text by Qwen and select the hyperpartisan label. The third platform The final dataset consists in a parallel corpus of 2,475 pairs of paragraphs for hyperpartisan neutralization in italian news articles.

A.3 Error Taxonomy Rewritten Task

Our taxonomy categorizes errors into four primary dimensions: (1) Neutralization errors, encompassing under-neutralization (biased language retained), over-neutralization (excessive removal of contextual information), tone shift failures (emotional language persists), and framing preservation (biased narrative structure maintained); (2) Factuality errors, including fact omission (information loss during rewriting), fact addition (hallucinated content), fact distortion (meaning alteration), and entity errors (incorrect references to people, organizations, or locations); (3) Fluency errors, covering grammatical mistakes, wrong language, coherence issues, unnatural phrasing, and redundant expressions; and (4) Structural errors, such as inappropriate length changes, disrupted information ordering, and incomplete outputs. We applied this taxonomy to the outputs of the three LLMs.

A.4 Analysis of SE and SY during Preliminary Annotation

Figure 10 illustrates that generally all the selected models for the preliminary generation of the pairs had consist results for SY (>2.0) (Mean: 2.64) across all the tasks, while the SE suffers from degrading depending on the model (Mean: 2.44). Particularly, Llama-3.1-8B-instruct received low scores depending on the task's complexity. Qwen-2.5-14B-Instructu obtained the highest scores in all the categories except for SE Rephrased, outperformed by Mistral-Nemo-2407-Instruct.

Analysis Rewritten Task during Preliminary Annotation Figure 11 shows that Mistral retained tracks of hyperpartisan language in the outputs 70% of the times, meaning that was not able to completely neutralize it, 6% deleted the context, 48% failed to remove emotional/loaded language, 43% preserved the framing completely hyperpartisan; lastly, 7% added facts and 9% provided empty outputs. Llama's performance was quite better compared to Mistral's but presented the same critical points. Conversely, Qwen only 13% left cues of hyperpartisan dialogue in the output and by far, according to our taxonomy, provided the best re-

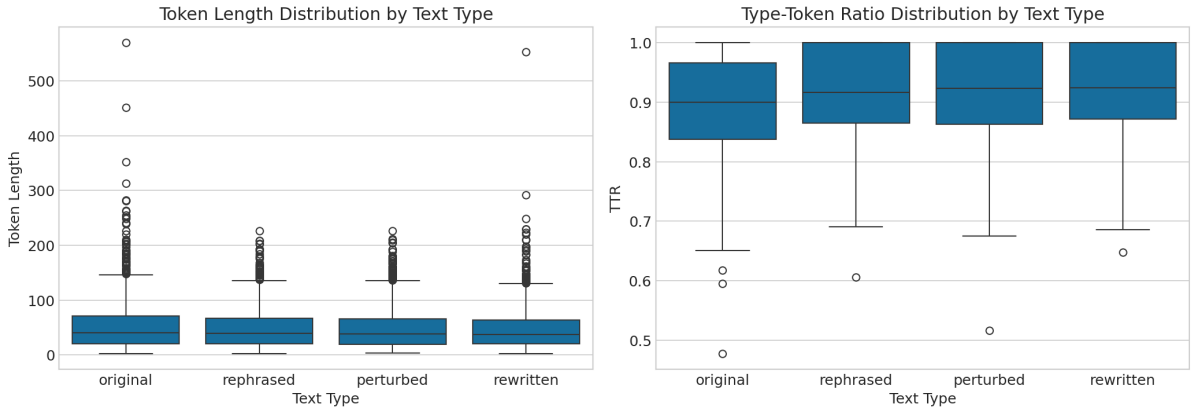

Figure 5: Token Length and TTR across tasks.

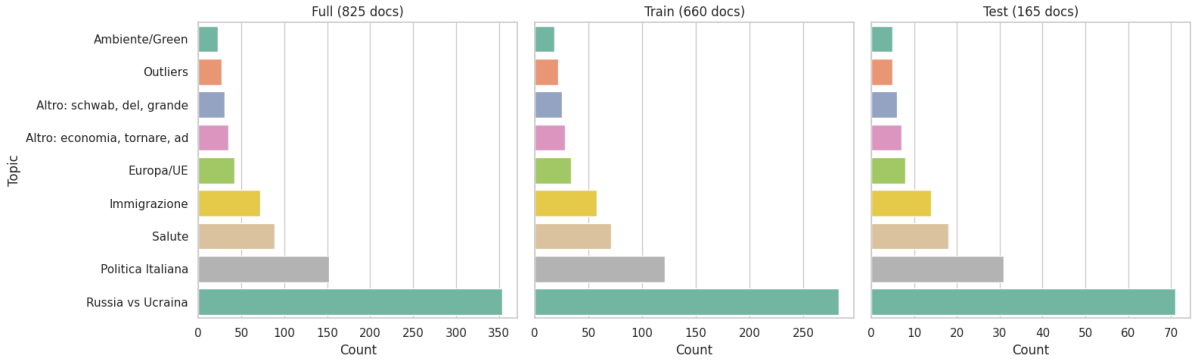

Figure 6: Topic Distribution of unique entries.

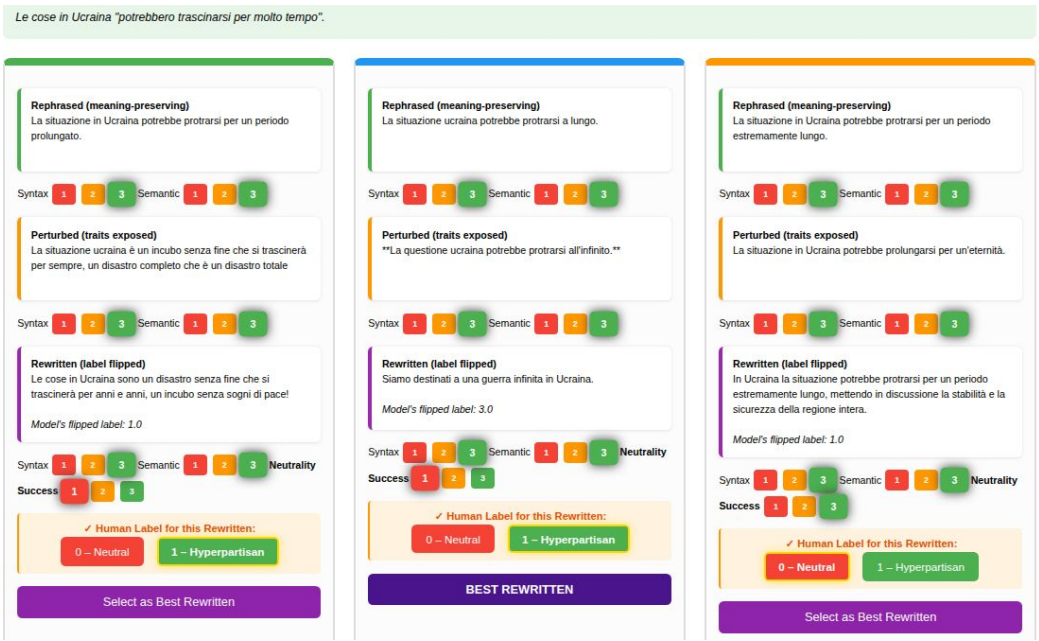

Figure 7: Screenshot of the platform for the preliminary evaluation of the LLMs.

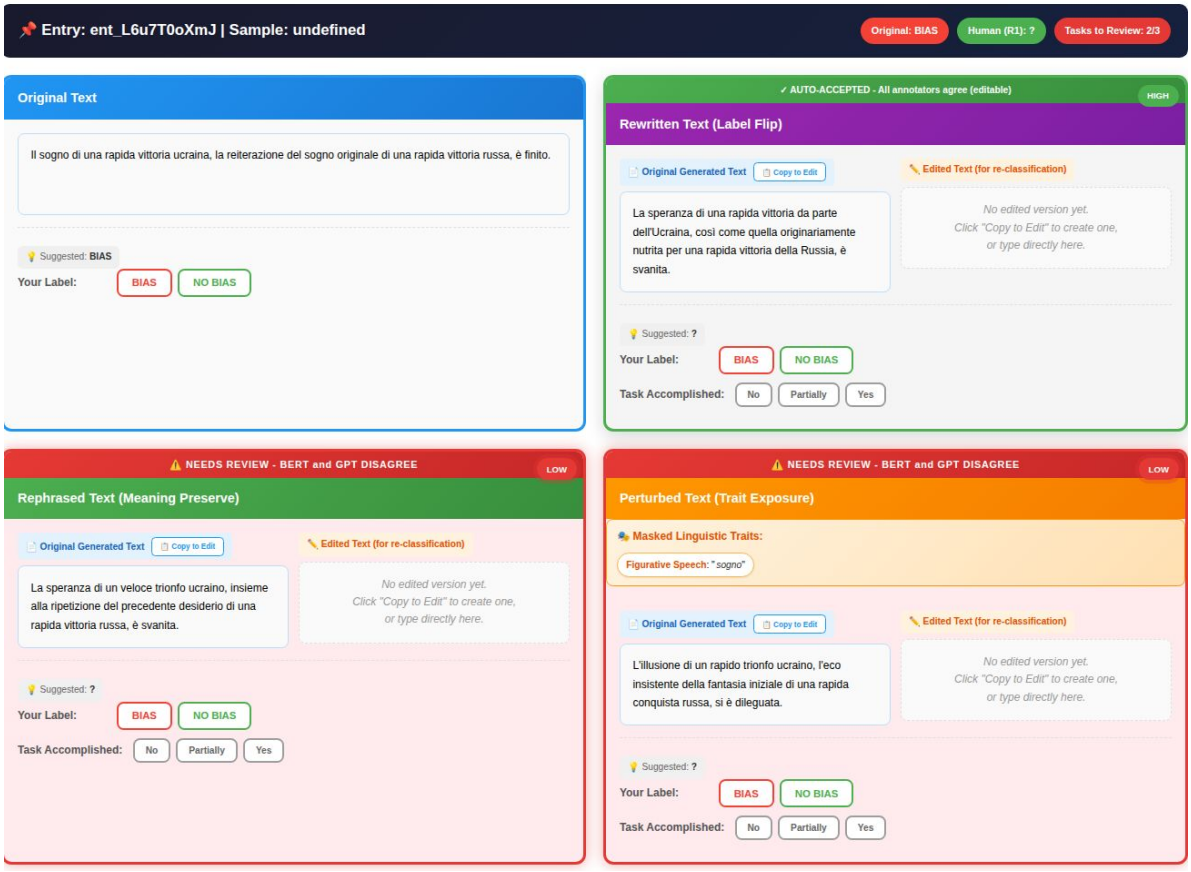

Figure 8: Screenshot of the platform for the annotation of the Qwen's generations.

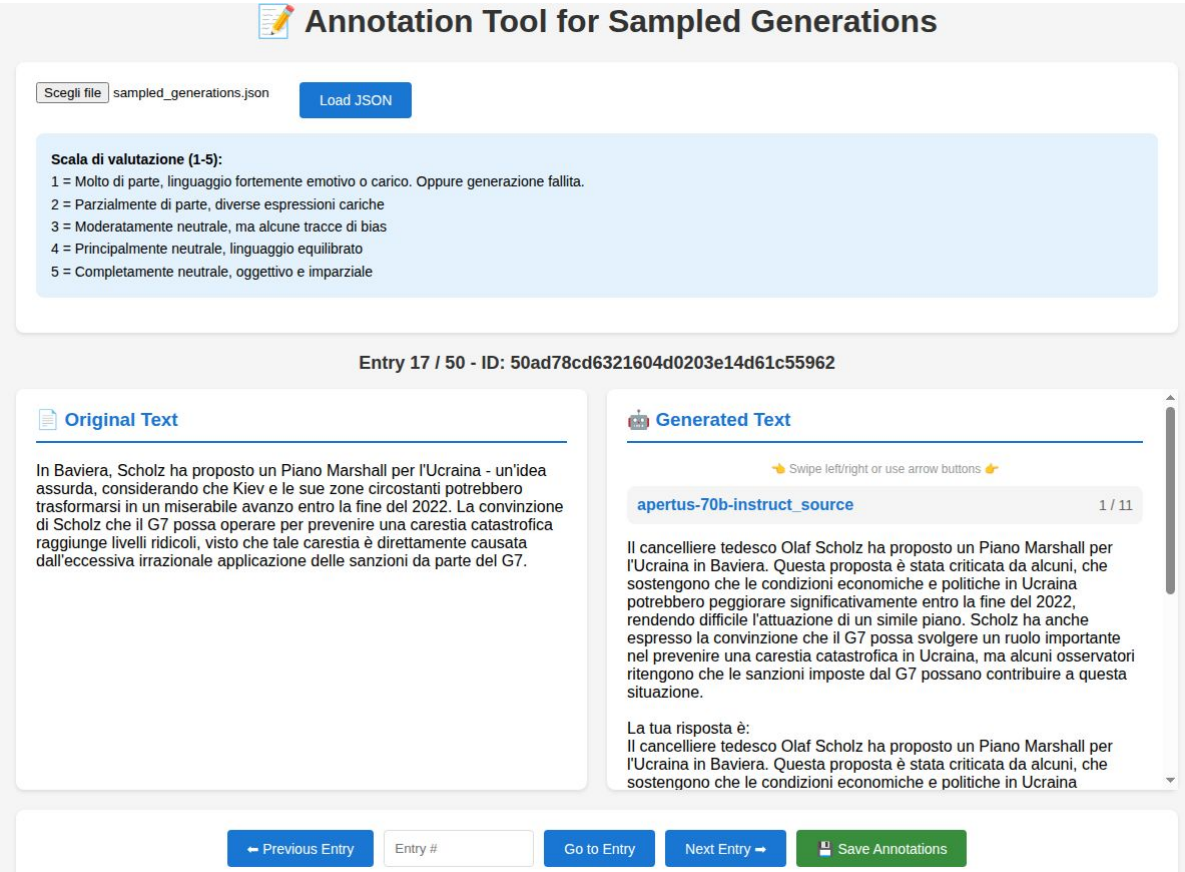

Figure 9: Screenshot of the platform for the annotation of the FT and 0shot LLMs generations.

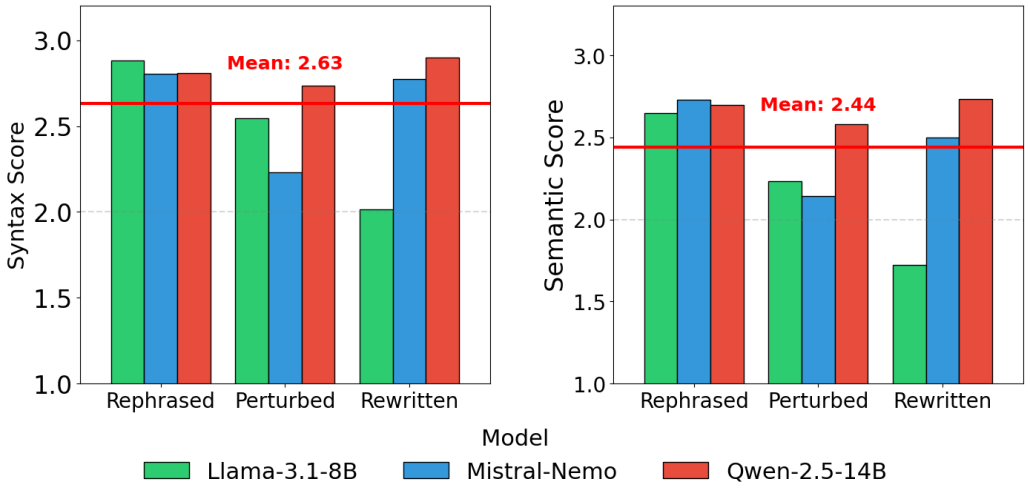

Figure 10: Syntax and semantic scores across generation tasks for evaluated LLMs.

sults. The extensive comparison illustrate a ierarchy amongst the selected models in terms of performance and tasks: Qwen>Llama>Mistral.

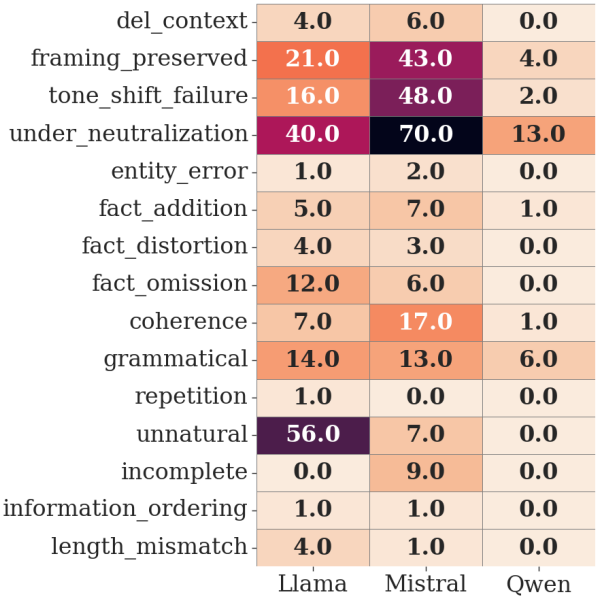

Figure 11: Percentage of the error types by LLMs during the neutralization.

B Additional Quantitative and Qualitative Analysis of Generated Texts by Qwen-2.5-14B-Instruct

To further evaluate the quality of the generated texts, we computed pairwise cosine similarities using the Italian SBERT model (nickprock/sentence-bert-base-italian-xxl-uncased) 10. Across tasks, mean similarities between original and generated texts ranged from 0.787 (rewritten) to 0.835 (rephrased) and 0.821 (perturbed), with a Kruskal-Wallis test confirming highly significant differences in similarity distributions H statistics = 227.11, p < 4.8359e-50). 2. Paired t-test: Token Length (Original vs Rewritten) t-statistic: 9.6456, p-value: 6.2318e-21 Original mean: 59.6, Rewritten mean: 52.9

A paired t-test also revealed that rewritten texts were significantly shorter than originals (tokens in original = 59.6 tokens, tokens in rewritten = 52.9 tokens; t = 9.65, p= 6.2e-11). Complementing these quantitative findings, t-SNE visualization of the SBERT embeddings (12) shows strong overall overlap between original and generated texts, with all categories forming a single dense cluster. This indicates robust semantic preservation, while subtle shifts—particularly for rewritten texts toward the cluster periphery—align with the observed lower similarity scores.

Notice that the mean similarity for Rewritten (Original vs Edited) is the lowest (0.7476). This shows that the human experts had to make the most significant semantic changes to achieve neutrality in that task.

LLMs struggle with Neutralization: The similarity between Generated vs Edited for Rewritten text (0.8579) shows that the human editors often had to move quite far away from what the LLM (Qwen) initially proposed to get a truly neutral result.

C LLM-as-a-Judge Agreement Details

Table 11 shows that GPT-4o-mini reached fair (0.302) agreement with annotator 2 and moderate (0.454) with annotator 1. However, these values fall in the low end of the ranges, posing questions

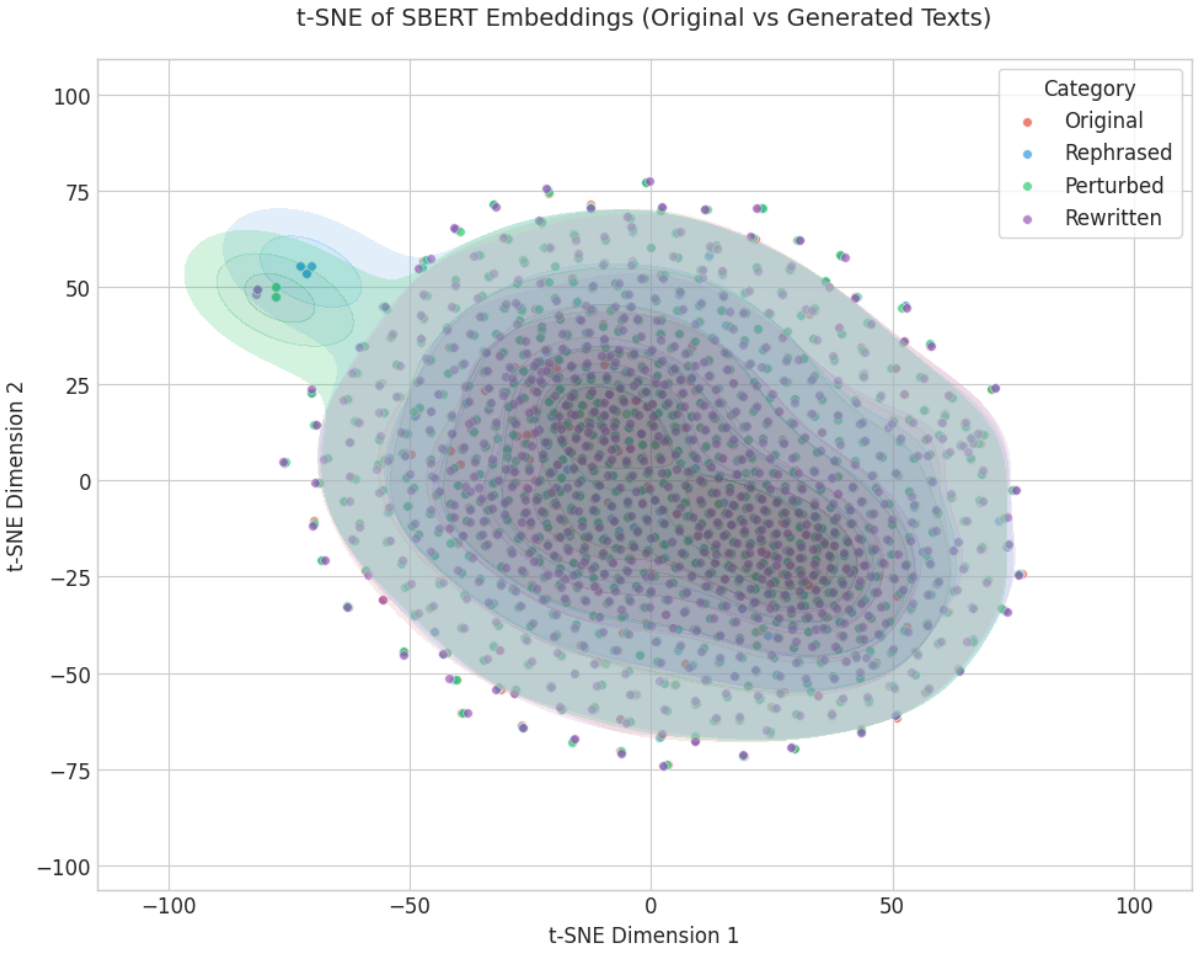

Figure 12: T-SNE.

Task	Comparison	n	Mean	Std
Rephrased	original vs generated	825	0.8352	0.1566
	original vs edited	271	**0.9207**	**0.0733**
	generated vs edited	271	0.8158	0.2562
Perturbed	original vs generated	825	0.8209	0.1731
	original vs edited	346	**0.8918**	**0.0819**
	generated vs edited	346	0.8289	0.2620
Rewritten	original vs generated	825	0.7869	0.1183
	original vs edited	559	0.7476	0.1267
	generated vs edited	559	**0.8579**	0.1559

Table 10: Semantic Similarity (Cosine) between text versions computed with Italian SBERT. Higher values indicate greater semantic similarity.

Ann 1	1	-	-
Ann 2	0.694	1	-
GPT-4o-mini	0.454	0.302	1
	Ann 1	Ann 2	GPT

Table 11: Cohen's K: IAA Human vs. GPT-4o-mini on Qwen's Rewritten Accomplishment.

about the application of small LLMs for this evaluation task of the Rewritten texts. Conversely, the agreement between humans is substancial (0.694), confirming that human revision for this subjective task is essential.

D LLMs FT and 0-shot Evaluation

To situate these findings within the broader context of model reliability, we must address the discrepancy between the automated semantic metrics and the human-centric evaluations. While fine-tuned (FT) models generally outperformed 0-shot configurations in lexical alignment (BLEU/ROUGE), the FT-sent-BERT scores remained remarkably high across nearly all models, often exceeding 0.90. This suggests a high degree of semantic preservation; however, a granular analysis using LLM-as-a-judge (GPT-4o-mini) and Human Evaluation reveals a more nuanced reality. The high FT-sent-BERT scores, which contrast with the lower qualitative ratings assigned by humans and GPT-4o-mini, may stem from two primary factors: i) Overestimation of Neutrality: ii) Sensitivity to Hallucinations: human judges penalize "added context" or "semantic inconsistencies" (Scores 1–2). Automated semantic metrics often fail to distinguish between a "neutralized" fact and a "hallucinated" fact if they share a similar embedding space, whereas human judges and GPT-4o-mini are more sensitive to these referential errors.

E Seq2seq results

In this section we provide more details for the seq2seq results.

We employ a fine-tuned BERT classifier to evaluate the neutrality of generated outputs. The classifier, trained on our corrected annotation dataset, provides binary predictions (neutral vs. hyperpartisan) as well as continuous probability scores.

E.0.1 Neutrality Rates by Configuration

Table 12 presents the neutrality rates achieved by each training configuration.

Metric	SFT	SFT-CLASS	DPO
Mean Neutrality	0.73	0.75	0.56
Std Dev	0.09	0.09	0.22
Min	0.59	0.62	0.13
Max	0.83	0.85	0.73

Table 12: Neutrality rate statistics by training configuration across all model families.

SFT-CLASS achieves highest mean neutrality (0.75) with consistent performance across models (std=0.09), validating the classifier-guided training approach. **Standard SFT provides competitive neutrality** (0.73) without requiring classifier integration during training. **DPO shows high variance** (std=0.22), with some models maintaining neutrality (mT5-base DPO: 0.72) while others experience catastrophic degradation (BART-base DPO: 0.13).

E.0.2 Model Family Comparison

Table 13 breaks down neutrality performance by model family.

Model Family	SFT	SFT-CLASS	DPO
FLAN-T5	**0.82**	**0.82**	0.72
BART	0.75	0.79	0.24
mT5	0.69	0.70	0.65
IT5	0.59	0.62	0.65

Table 13: Mean neutrality rate by model family and training configuration.

FLAN-T5 consistently achieves the highest neutrality rates across all configurations, likely benefiting from its instruction-tuning pretraining which aligns well with the neutralization task framing. Particularly: **FLAN-T5-base SFT-CLASS** achieves the highest overall neutrality (0.85), tied with BART-large SFT-CLASS; **BART models show extreme DPO sensitivity**: BART-base drops from 0.73 (SFT-CLASS) to 0.13 (DPO), a 82%

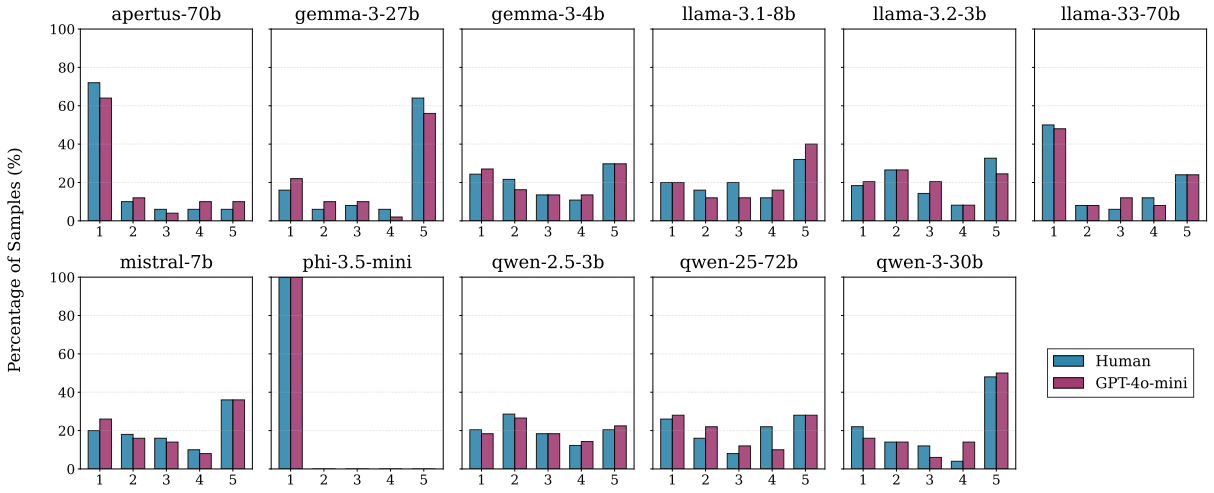

Figure 13: Ratings distribution.

relative decrease. Lastly, **IT5 underperforms** despite being Italian-native, suggesting that domain-specific pretraining does not guarantee task-specific performance.

E.0.3 Top Performers by Neutrality

The five highest-neutrality configurations are:

1. FLAN-T5-base SFT-CLASS: 0.853
2. BART-large SFT-CLASS: 0.853
3. FLAN-T5-base SFT: 0.832
4. FLAN-T5-large SFT: 0.812
5. FLAN-T5-large SFT-CLASS: 0.788

This ranking demonstrates FLAN-T5's dominance in neutrality and the effectiveness of classifier-guided training (SFT-CLASS) for maximizing neutrality.

E.0.4 Neutrality-BLEU Trade-off

We observe a negative correlation between BLEU and neutrality rate ($r = -0.42$, $p = 0.056$), suggesting a quality-neutrality trade-off. DPO models exemplify this tension:

- BART-base DPO achieves the highest BLEU (29.1) but the lowest neutrality (0.13)
- FLAN-T5-large DPO balances both: 24.1 BLEU with 0.73 neutrality
- mT5-base DPO maintains neutrality (0.72) while improving BLEU (+8.7 over SFT)

E.1 Semantic Similarity Analysis

We employ SBERT (sentence-BERT: nickprock/sentence-bert-base-italian-xxl-uncased) embeddings to measure semantic similarity across three dimensions: **SBERT$_{\text{src-gen}}$**: Source-to-generation similarity (semantic preservation); **SBERT$_{\text{gen-ref}}$**: Generation-to-reference similarity (alignment with gold standard); **SBERT$_{\text{src-ref}}$**: Source-to-reference similarity (baseline similarity).

E.1.1 Semantic Preservation (SBERT$_{\text{src-gen}}$)

Table 14 presents SBERT$_{\text{src-gen}}$ scores measuring how well models preserve source semantics.

Metric	SFT	SFT-CLASS	DPO
Mean	**0.88**	0.86	0.67
Std Dev	0.04	0.05	0.07

Table 14: SBERT$_{\text{src-gen}}$ statistics by training configuration.

Critical finding: DPO significantly reduces semantic preservation compared to SFT ($t = 6.40$, $p < 0.0001$, Mann-Whitney $p = 0.0007$). This is the only metric showing statistically significant difference between configurations. The 0.21-point drop (24% relative) indicates that DPO's BLEU improvements come at the cost of semantic fidelity to the original text.

E.1.2 Reference Alignment (SBERT$_{\text{gen-ref}}$)

SBERT$_{\text{gen-ref}}$ measures how closely generations match reference neutralizations:

SFT achieves the highest reference alignment with minimal variance, indicating consistent learning of neutralization patterns. DPO shows higher

Metric	SFT	SFT-CLASS	DPO
Mean	0.72	0.69	0.66
Std Dev	0.02	0.05	0.10

Table 15: $\text{SBERT}_{\text{gen-ref}}$ statistics by training configuration.

variance (std=0.10), reflecting its model-dependent behavior.

E.1.3 Model-Level Semantic Analysis

Table 16 provides per-model semantic metrics.

The Δ column reveals an interesting pattern: SFT models preserve source semantics more than they align with references ($\Delta > 0$), while some DPO models show the reverse pattern ($\Delta < 0$), indicating they diverge from sources while coincidentally matching references.

E.1.4 Metric Correlations

Table 17 presents Pearson correlations between key metrics.

BERTScore strongly correlates with $\text{SBERT}_{\text{gen-ref}}$ ($r = 0.94$), validating both as measures of reference alignment. **BLEU correlates with $\text{SBERT}_{\text{gen-ref}}$** ($r = 0.59$) but not with $\text{SBERT}_{\text{src-gen}}$ ($r = -0.01$), suggesting BLEU measures reference matching rather than semantic preservation. **Neutrality shows weak correlations** with all other metrics, indicating it captures an independent dimension of quality. **Semantic preservation and reference alignment are moderately correlated** ($r = 0.66$), suggesting models that preserve source meaning also tend to match references.

E.2 Configuration Impact Analysis

E.2.1 SFT to DPO Transition

Table 18 quantifies the impact of transitioning from SFT to DPO for each model.

We found that mT5-base is the optimal DPO candidate: +8.7 BLEU with no neutrality loss and minimal semantic drift (-0.13). BART-large shows severe degradation. Despite +5.9 BLEU, it loses 39 percentage points of neutrality. IT5-large fails entirely under DPO: -9.4 BLEU indicates catastrophic forgetting. Mean semantic preservation loss of 0.21 confirms DPO systematically trades semantic fidelity for surface-level improvements.

E.2.2 SFT vs. SFT-CLASS Comparison

Statistical testing reveals no significant differences between SFT and SFT-CLASS:

While SFT-CLASS shows a trend toward improved neutrality (+0.02), this difference is not statistically significant ($p = 0.66$). The classifier-guided approach may benefit from larger training sets or stronger guidance signals.

F Models and Training Configurations

We evaluate 11 decoder-only LLMs spanning three model families and two configuration strategies: **zero-shot prompting** for large-scale models and **fine-tuning** for smaller models. Table 20 summarizes the configurations.

Zero-shot Prompting (Large Models). For models with 7B or more parameters, full fine-tuning would require multiple high-memory GPUs exceeding our computational budget. Instead, we employ zero-shot prompting with the following Italian-language instruction and unsloth for finetuning.

These models leverage their extensive pretraining and instruction-following capabilities to perform neutralization without task-specific training. The zero-shot approach evaluates the models' inherent understanding of neutrality and their ability to generalize from instructions alone.

LoRA Fine-tuning (Small Models). For models with 8B or fewer parameters, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2021) and unsloth to efficiently fine-tune on our training set. LoRA introduces trainable low-rank matrices into the attention layers while keeping the pretrained weights frozen, enabling fine-tuning on a single 48GB GPU. Our LoRA configuration uses:

- Rank $r = 16$
- Alpha $\alpha = 32$ (scaling factor)
- Dropout = 0.1
- Target modules: query, key, value projection layers

For seq2seq we trained them for 10 epochs. To FT LLMs, training proceeds for 1 epochs with learning rate 2×10^{-4} and batch size 4 (with gradient accumulation to effective batch size 16). Fine-tuned models learn task-specific patterns from the training examples, including proper handling of

Model	Config	$\text{SBERT}_{\text{src-gen}}$	$\text{SBERT}_{\text{gen-ref}}$	Δ
BART-large	SFT-CLASS	**0.925**	0.747	0.178
BART-large	SFT	0.913	0.722	0.191
FLAN-T5-large	SFT	0.909	0.735	0.174
FLAN-T5-base	SFT	0.908	0.727	0.181
FLAN-T5-large	SFT-CLASS	0.905	0.737	0.168
IT5-large	DPO	0.548	0.427	0.121
BART-large	DPO	0.618	0.701	−0.083
IT5-base	DPO	0.625	0.640	−0.015

Table 16: Top 5 and bottom 3 models by $\text{SBERT}_{\text{src-gen}}$. Δ = $\text{SBERT}_{\text{src-gen}}$ − $\text{SBERT}_{\text{gen-ref}}$.

	BLEU	Neutrality	$\text{SBERT}_{\text{src}}$	$\text{SBERT}_{\text{ref}}$
BLEU	—	−0.42	−0.01	**0.59****
Neutrality		—	0.41	−0.10
$\text{SBERT}_{\text{src-gen}}$			—	**0.66*****
$\text{SBERT}_{\text{gen-ref}}$				—
BERTScore	**0.70*****	−0.17	0.42	**0.94*****

Table 17: Pearson correlations between metrics. $^{**}p < 0.01$, $^{***}p < 0.001$.

Model	ΔBLEU	ΔNeutrality	$\Delta\text{SBERT}_{\text{src}}$
mT5-base	**+8.7**	+0.00	−0.13
mT5-large	+6.2	−0.09	−0.12
BART-large	+5.9	−0.39	−0.30
FLAN-T5-large	+4.8	−0.09	−0.22
FLAN-T5-base	+4.3	−0.12	−0.22
IT5-large	−9.4	+0.14	−0.28
Mean	+3.4	−0.09	−0.21

Table 18: Change in metrics when transitioning from SFT to DPO. Negative Δ indicates DPO underperformance.

Metric	Δ (CLASS−SFT)	p-value	Significant?
BLEU	−0.07	0.956	No
Neutrality	+0.02	0.660	No
$\text{SBERT}_{\text{src-gen}}$	−0.02	0.414	No
$\text{SBERT}_{\text{gen-ref}}$	−0.02	0.316	No
BERTScore	−0.05	0.272	No

Table 19: Paired comparison between SFT and SFT-CLASS configurations (t-test).

Model	Size	Config	Rationale
Zero-shot Prompting			
Gemma3	27B	0-shot	Too large for FT
Qwen3	30B	0-shot	Too large for FT
Qwen2.5	72B	0-shot	Too large for FT
Llama3.3	70B	0-shot	Too large for FT
Apertus	70B	0-shot	Too large for FT
LoRA Fine-tuning			
Mistral	7B	FT	Fits in memory
Llama3.1	8B	FT	Fits in memory
Llama3.2	3B	FT	Fits in memory
Gemma3	4B	FT	Fits in memory
Qwen2.5	3B	FT	Fits in memory
Phi3.5	Mini (3.8B)	FT	Fits in memory

Table 20: LLM configurations by model size. Large models ($\geq$27B) use zero-shot prompting due to GPU memory constraints; smaller models ($\leq$8B) use LoRA fine-tuning.

Italian political terminology and stylistic preferences for neutral reformulation.

Configuration Selection Rationale. The dichotomy between zero-shot and fine-tuned configurations reflects practical constraints rather than experimental design preference. Ideally, all models would undergo fine-tuning for fair comparison; however:

1. **Memory constraints**: A 70B model required approximately 140GB in FP16, far exceeding single-GPU capacity even with LoRA (which reduces but does not eliminate memory requirements for large models).

2. **Compute budget**: Fine-tuning 70B models with distributed training would require multi-node setups beyond our allocation.

3. **Practical relevance**: Zero-shot evaluation reflects real-world deployment scenarios where users leverage large models via API without fine-tuning access.

This configuration strategy enables comprehensive evaluation across the model size spectrum while remaining computationally feasible.

Model Selection We selected models representing diverse architectures and training approaches:

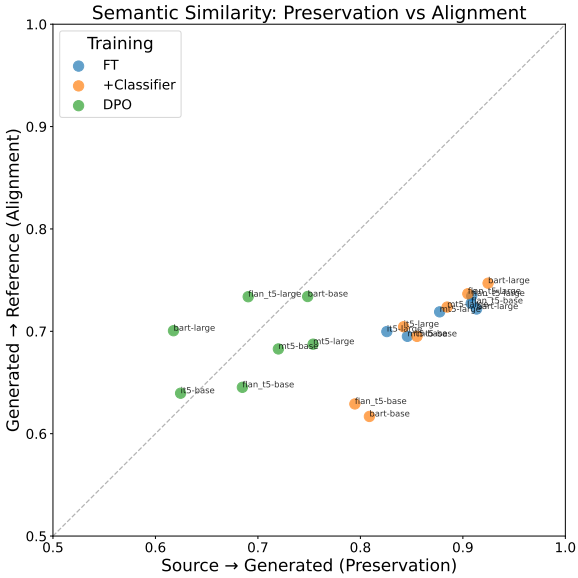

Figure 14: Scatter plot of semantic preservation ($SBERT_{src\text{-}gen}$) vs. reference alignment ($SBERT_{gen\text{-}ref}$). SFT models cluster in the upper-right (high on both), while DPO models spread toward lower preservation.

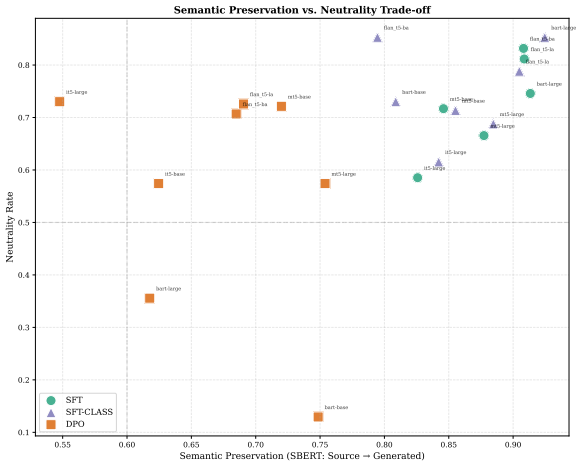

Figure 15: Preservation-alignment analysis showing the trade-off between source fidelity and reference matching across training configurations.

- **Qwen series**: Strong multilingual capabilities, including Italian
- **Llama series**: Widely-adopted open models with proven performance
- **Gemma series**: Google's efficient instruction-following models
- **Mistral**: European-developed model with strong multilingual support
- **Apertus**: Italian-focused variant based on Llama architecture
- **Phi3.5**: Microsoft's small-but-capable model (included for comparison)

F.1 GPT-4o-mini as a Judge Prompts

G Computational Infrastructures

All the experiments were run on two NVIDIA Tesla V100 32GB, one NVIDIA A100 40GB, one NVIDIA A100 80GB.

Component	Prompt Content
User Prompt	You are an expert Italian rewriter and back-translator for hyper-partisan text. - CONTEXT - 1. Original Italian paragraph: original 2. Masked Italian paragraph: masked 3. Span mapping: span_table The placeholders have this structure: [linguistic_trait_n], where the linguistic trait are: LOADED_LANGUAGE, EPITHET, IRONY/SARCASM, HYPERBOLIC_LANGUAGE, FIGURATIVE_SPEECH, NEOLOGISM and the n indicate the occurrence of the each linguistic traits. - TASK 1: REPHRASE - Rephrase the **original Italian paragraph** in natural, fluent Italian while **preserving the exact meaning**. Avoid changing tone or bias. Output should be different in wording but identical in semantics. - TASK 2: BACK-TRANSLATION (UNPERTURBED) - • Translate masked Italian to English (keep placeholders) • Translate that English back to Italian (keep placeholders) • The placeholders **MUST** remain in unperturbed outputs both in Italian and English outputs. Keep them in bold text as in the original text. This is an example of the output for TASK 2: BACK-TRANSLATION (UNPERTURBED): "Text Text Text [LOADED_LANGUAGE_n] Text Text [EPITHET_n] Text" - TASK 3: BACK-TRANSLATION (PERTURBED) - • Replace each placeholder with a **synonymous, bias-amplifying Italian expression** • Translate the full replaced Italian to English • Translate that English back to Italian • **NO placeholder may remain** in perturbed outputs This is an example of the output for TASK 3: BACK-TRANSLATION (PERTURBED): "Text Text Text Text Text Text Text Text" Return **ONLY** this JSON (no markdown, no extra text): "rephrased": "rephrased Italian", "it_unperturbed": "Italian with placeholders", "it_perturbed": "Italian with replaced expressions"

Table 21: Prompt used for data augmentation with LLMs. The prompt instructs the model to perform three parallel tasks on each hyperpartisan paragraph: surface-level rephrasing (preserving bias), unperturbed back-translation (preserving placeholders), and perturbed back-translation (amplifying bias by replacing placeholders with synonymous loaded expressions). The structured JSON output facilitates automatic extraction and subsequent human supervision.

Component	Prompt Content
Judge Prompt (Italian)	Sei un esperto di analisi del linguaggio giornalistico italiano. Esegui il seguente compito seguendo la definizione di hyperpartisan adottata nelle guidelines: 1-Etichetta il testo come hyperpartisan o neutrale. guidelines Testo da valutare: text Rispondi SOLO con un JSON in questo formato (nessun testo aggiuntivo): "label": <hyperpartisan, neutrale>,

Table 22: Prompt used for LLM-as-a-judge evaluation with GPT-4o-mini.

Component	Prompt Content
System Instruction	Sei un giornalista esperto. Riscrivi il seguente paragrafo in modo completamente neutrale, rimuovendo ogni tono hyperpartisan, sarcastico, caricaturale. Mantieni tutti i fatti e il significato originale, usando un linguaggio oggettivo e imparziale. Ti verranno forniti degli esempi di neutralizzazione di paragrafi hyperpartisan nel seguente formato: ESEMPIO: Hyperpartisan:"text", Neutrale: "text".
Few-Shot Examples Esempio 2: Hyperpartisan: "{text}" Esempio 3: Hyperpartisan: "{text}"	Esempio 1: Hyperpartisan: "{text}"
User Instruction	Questo è il paragrafo che devi neutralizzare: Paragrafo: {input_text} Rispondi esclusivamente con la versione neutralizzata, senza aggiunte, spiegazioni o commenti.

Table 23: Prompt used for zero-shot neutralization with big size LLMs (>70B parameters). The system instruction defines the role and task, three high-quality few-shot examples illustrate the expected neutralization style (preserving facts while removing hyperpartisan tone, sarcasm, and loaded language), and the user instruction provides the target paragraph with a strict requirement to output only the neutralized version.

Component	Prompt Content
Judge Prompt (Italian)	Sei un esperto di analisi del linguaggio giornalistico italiano. Esegui i seguenti giudizi: 1-Etichetta il testo come hyperpartisan o neutrale. 2-Assegna un valore (int) in una scala da 1 a 3 per la semantica del testo prodotto e per la correttezza sintattica seguendo le guidelines. guidelines Testo da valutare: text Rispondi SOLO con un JSON in questo formato (nessun testo aggiuntivo): "label": <hyperpartisan, neutrale>, "semantic_score": <1,2,3>, "syntax_score": <1,2,3>,

Table 24: Prompt used for LLM-as-a-judge evaluation with GPT-4o-mini. The prompt instructs the model to rate the neutrality of a given text on a 1–5 Likert scale, provide a confidence level, list any biased phrases identified, and give brief reasoning. The strict JSON-only output format ensures structured, parseable responses for automated analysis and comparison with human judgments.

Component	Prompt Content
Judge Prompt (Italian)	JUDGE_PROMPT = """Sei un esperto di analisi del linguaggio giornalistico italiano. Valuta quanto il seguente testo sia neutrale e imparziale. Scala di valutazione (1-5): 1 = Molto di parte, linguaggio fortemente emotivo o carico. Oppure generazione fallita: presenza di errori gravi, testo incomprensibile, output ripetuto. 2 = Parzialmente di parte, diverse espressioni cariche. Generazione con errori ed incoerenze semantiche come l'aggiunta di contesto inesistente o troncamento del contesto fornito. 3 = Moderatamente neutrale, ma alcune tracce di bias. Generazione generalmente corretta ma con alcune imprecisioni grammaticali e semantiche. Il testo può contenere lievi allucinazioni e del contesto aggiunto con correttezza referenziale. 4 = Principalmente neutrale, linguaggio equilibrato. Generazione sostanzailmente corretta con lievi imperfezioni grammaticale (concordanza, piccoli refusi). Non viene aggiunto alcun tipo di contesto, tutte le informazioni nel testo originale vengono preservate, ma neutralizzate. 5 = Completamente neutrale, oggettivo e imparziale. Linguaggio chiaro e professionale. Generazione priva di errori grammaticali o semantici. Testo originale NON da valutare: input Testo generato da un LLM: generated Rispondi SOLO con un JSON in questo formato (nessun testo aggiuntivo): "neutrality_score": <1-5>, """

Table 25: Prompt used for LLM-as-a-judge evaluation with GPT-4o-mini. The prompt instructs the model to rate the neutrality of a given text on a 1–5 Likert scale. The strict JSON-only output format ensures structured, parseable responses for automated analysis and comparison with human judgments.