Annotation Guidelines: Quality Assessment of Augmented Hyperpartisan Data: Preliminary Annotation

Project Overview

You are evaluating the quality of machine-generated augmented text for Italian hyperpartisan paragraph detection. Three LLMs (Llama3.1-8b, Mistral-Nemo, Qwen2.5-14b) performed two augmentation tasks:

1. **Rephrasing/Backtranslation**: Keeping the same meaning and label
2. **Style Transfer/Rewriting**: Changing the label (hyperpartisan ↔ neutral) while preserving core meaning

Your Task

For each text pair, rate two dimensions on a **3-point scale**:

- **Semantic Preservation** (SP)
- **Syntax Correctness** (SC)

Rating Scales

Semantic Preservation (SP)

Definition: Does the generated text preserve the core meaning, factual content, and main claims of the original?

Score	Label	Description	When to Use
3	**Full Preservation**	All key information retained; same claims, entities, and events; only stylistic changes	Core facts identical; could substitute one for the other

2	**Partial Preservation**	Main meaning preserved but with notable differences: minor facts omitted/added, slight claim modifications, or emphasis shifts	Recognizably about the same topic but with meaningful differences
1	**Poor Preservation**	Major meaning changes: contradictory claims, different events, critical information lost, or topic drift	Cannot substitute one for the other; fundamentally different

Important Notes for SP:

- For **rephrased/backtranslated** text: Expect SP=3 (same label task)
- For **rewritten** text (label-flipped): Expect SP=2-3. The label changes (hyperpartisan→neutral or vice versa), but core factual content should remain. Judge based on facts, not tone.

Syntax Correctness (SC)

Definition: Is the generated text grammatically correct and fluent in Italian?

Score	Label	Description	When to Use
3	**Fully Correct**	Perfect or near-perfect grammar; natural Italian; no errors	Native-level fluency; publication-ready
2	**Minor Errors**	1-3 small errors (agreement, article, word choice); meaning still clear; readable	Understandable but needs light editing
1	**Major Errors**	Multiple serious errors; awkward phrasing; comprehension impaired; clearly machine-generated	Requires significant editing; sounds unnatural

Common Syntax Issues to Watch:

- Subject-verb agreement errors
- Incorrect article usage (il/la/i/le)
- Word order problems
- Unnatural phrasing (translationese)
- Missing or extra words
- Incorrect prepositions
- Gender/number agreement

Annotation Types

You will evaluate three augmentation types per model:

Type 1: Rephrased Text

Column: `{model}_rephrased` **Task**: Paraphrasing with same label **Expectation**: SP=3, SC=2-3

Type 2: Perturbed Text (Italian)

Column: `{model}_it_perturbed` **Task**: Backtranslation with linguistic trait modification **Expectation**: SP=2-3, SC=2-3

Type 3: Rewritten Text

Column: `{model}_rewritten_text` **Task**: Label flipping (hyperpartisan ↔ neutral) **Expectation**: SP=2-3, SC=2-3 **Special consideration**: Tone/style should change but facts should remain

---

Annotation Process

Step 1: Read Original Text

- Note the topic, main claims, and factual content
- Identify the original label (1=hyperpartisan, 0=neutral)
- Observe any linguistic traits (loaded language, figurative speech, etc.)

Step 2: Read Generated Text

- Read the generated text independently
- Note your initial impression of quality

Step 3: Compare Systematically

For Semantic Preservation:

1. List key facts/claims in original
2. Check if each appears in generated text
3. Note any additions, omissions, or contradictions
4. For rewritten text: Expect tone/style change but not fact change

For Syntax Correctness:

1. Read generated text
2. Mark any grammatical errors
3. Assess overall fluency

4. Count errors: 0 = score 3, 1-3 = score 2, 4+ = score 1

Step 4: Assign Scores

- SP: 1, 2, or 3
- SC: 1, 2, or 3

Step 5: Optional Notes

- If particularly good/bad, note why
- Flag any interesting patterns

Detailed Examples

Example 1: Rephrased Text (Climate Change Topic)

Original (Label=1, Hyperpartisan):

> "Il riscaldamento globale è una colossale truffa orchestrata dalle élite per controllare le nostre vite. Gli scienziati corrotti falsificano i dati per accedere ai finanziamenti."

Generated (Llama - Rephrased):

> "Il cambiamento climatico rappresenta, secondo alcuni, un inganno delle classi dirigenti per esercitare controllo sociale. Ricercatori accusati di corruzione alterano le evidenze per ottenere fondi."

Annotation:

- **SP = 3**: Same claims (elite conspiracy, data falsification, funding motive); same entities; identical meaning
- **SC = 3**: Perfect grammar, natural Italian, fluent
- **Notes**: Successfully rephrased while maintaining hyperpartisan tone and all claims

---

Example 2: Rewritten Text (Label Flip)

Original (Label=1, Hyperpartisan):

> "L'Ucraina è un regime fantoccio della NATO che opprime brutalmente il Donbass russofono con l'aiuto dei criminali occidentali."

Generated (Mistral - Rewritten, New Label=0):

> "L'Ucraina mantiene relazioni con la NATO mentre affronta tensioni nella regione del Donbass, dove risiedono popolazioni russofone. La comunità internazionale ha espresso preoccupazioni sulla situazione."

Annotation:

- **SP = 2**: Core topic preserved (Ukraine-NATO-Donbass-Russian speakers); facts present but framing neutralized; "oppression" → "tensions"; "puppet regime" → "maintains relations"; some specificity lost but main situation intact
- **SC = 3**: Grammatically perfect, natural Italian
- **Notes**: Successful neutralization; emotional language removed while preserving factual skeleton

---

Example 3: Perturbed Text (Poor Quality)

Original (Label=0, Neutral):

> "Il governo ha approvato un nuovo piano di investimenti nelle energie rinnovabili per un totale di 5 miliardi di euro."

Generated (Qwen - IT Perturbed):

> "Il governo ha approvazione un nuovo piano di investimento nelle energia rinnovabile per un totale di 5 miliardo di euro."

Annotation:

- **SP = 3**: All facts preserved (government, plan, renewable energy, 5 billion euros)
- **SC = 1**: Multiple errors: "approvazione" (wrong verb form), "investimento" (should be plural), "energia rinnovabile" (missing plural), "miliardo" (agreement error); clearly machine-generated
- **Notes**: Content perfect but syntax very poor; needs significant editing

---

Example 4: Semantic Drift

Original (Label=1, Hyperpartisan):

> "Bruxelles impone regole assurde che distruggono le nostre tradizioni e la nostra sovranità nazionale in nome del globalismo."

Generated (Llama - Rewritten):

> "L'Unione Europea promuove politiche di armonizzazione che alcuni ritengono possano influenzare le culture locali e l'autonomia decisionale degli stati membri."

Annotation:

- **SP = 2**: Main topic present (EU regulations, national sovereignty, cultural impact); key transformation: "imposes absurd rules" → "promotes harmonization policies"; "destroys" → "may influence"; "globalism" lost; somewhat weaker claim but recognizably same issue
- **SC = 3**: Excellent grammar and fluency
- **Notes**: Good neutralization but slight semantic weakening acceptable for style transfer task

---

Example 5: Major Semantic Failure

Original (Label=0, Neutral):

> "Il ministro dell'economia ha presentato il bilancio preventivo 2024 in parlamento, con previsioni di crescita del PIL al 1,2%."

Generated (Mistral - Rewritten):

> "Il ministro promette miracoli economici impossibili mentre ignora la crisi che devasta le famiglie italiane e favorisce i ricchi."

Annotation:

- **SP = 1**: Completely different claims; original about budget presentation (factual), generated about broken promises (opinion); GDP figure lost; new claims added (crisis, favoritism); contradictory tone; topic drift from procedural to accusatory
- **SC = 2**: Grammar acceptable but some awkwardness ("devasta le famiglie")
- **Notes**: Failed rewriting task; invented claims not in original; unacceptable augmentation

---

Special Considerations

For Rewritten Text (Label Flipped)

What Should Change:

- Emotional language (e.g., "truffa colossale" → "questione dibattuta")
- Loaded terms (e.g., "regime fantoccio" → "governo")
- Absolutist claims (e.g., "sempre" → "spesso", "mai" → "raramente")
- Attribution (e.g., direct claims → "secondo alcuni")

What Should NOT Change:

- Core facts (dates, numbers, entities, events)

- Main topic
- Key actors/participants
- Verifiable information

Scoring Rewritten Text:

- SP=3: Facts identical, only tone/style changed
- SP=2: Facts mostly preserved, minor additions/omissions acceptable
- SP=1: Facts contradicted, major information lost, or topic drift

For Backtranslated Text (IT Perturbed)

Common Issues:

- Translation artifacts
- Word choice awkwardness
- Unnatural phrasings
- Lost idiomatic expressions

Be Lenient On:

- Minor stylistic differences
- Synonym substitutions
- Sentence restructuring

Be Strict On:

- Fact changes
- Grammar errors
- Comprehensibility issues

Edge Cases & FAQs

Q: The generated text adds contextual information not in the original. How do I score?

A:

- If additions are minor and don't change core meaning → SP=3
- If additions are substantial but relevant → SP=2
- If additions contradict or significantly alter meaning → SP=1

Q: The text is grammatically correct but sounds very unnatural/machine-translated.

A: SC=2. Reserve SC=1 for clear grammatical errors. Unnaturalness affects SC but shouldn't alone warrant SC=1.

Q: For rewritten text, the label should flip but BERT predicts the original label. How does this affect my rating?

A: Rate based on YOUR judgment, not BERT's prediction. If the text reads as neutral (when flipped from hyperpartisan) to you, rate accordingly. BERT predictions will be analyzed separately.

Q: The original text is already somewhat neutral (or hyperpartisan), and the rewrite is very similar.

A: This is acceptable. If original label=0 (neutral) with minimal hyperpartisan traits, rewriting to label=1 may require only small changes. Score based on whether the changes are appropriate for the task.

Q: I disagree with the original label assignment.

A: Rate based on the comparison between original and generated, not your judgment of the original label's correctness. Note your concern if it's extreme.

Q: There are multiple errors but the text is still understandable.

A: Count errors:

- 1-3 small errors → SC=2
- 4+ errors or errors that impair comprehension → SC=1

Annotation Recording

You will be provided with a custom HTML platform.

Quality Control

Self-Check Questions:

After every 20 annotations, pause and verify:

1. Am I being too strict or too lenient?
2. Am I consistent in applying the 3-point scale?
3. Have I encountered any patterns worth discussing?

Calibration Markers:

- **SP=3 should be ~40-50%** of rephrased texts
- **SP=1 should be <10%** overall (rare)
- **SC=3 should be ~50-60%** (LLMs generally grammatical)
- **SC=1 should be ~10-15%** (some models struggle)

Inter-Annotator Agreement

Pilot Phase (30 samples):

- Both annotators rate same 30 samples
- Calculate Cohen's kappa
- Discuss disagreements
- Refine understanding of scale

Target Agreement:

- **Exact agreement**: >60%
- **Adjacent agreement** (within 1 point): >90%
- **Kappa**: >0.60 (substantial agreement)

Disagreement Resolution:

- If disagreement on <20% of samples: average scores
- If disagreement on >20%: meet to discuss and recalibrate

Tips for Efficient Annotation

1. **Batch by Model**: Annotate all samples from one model at a time to notice patterns
2. **Use Templates**: Copy-paste the recording format for speed
3. **Take Breaks**: Every 50 samples, take a 10-minute break
4. **Flag Uncertainty**: When unsure between 2 scores, note it and discuss later
5. **Track Time**

Common Patterns to Watch

Model-Specific Tendencies:

Llama3.1: Often verbose, may add explanatory context **Mistral**: Generally fluent, watch for semantic drift on rewriting **Qwen2.5**: May have more syntax issues, strong on semantic preservation

Task-Specific Issues:

Rephrasing: Expect high quality (SP=3, SC=3 common) **IT Perturbed**: Backtranslation artifacts, unnatural phrasing **Rewritten**: Most challenging; watch for over/under-neutralization

Contact & Questions

If you encounter:

- Offensive content
- Ambiguous cases not covered here
- Technical issues with the data
- Need for clarification

Action: Flag the sample, add detailed notes, and continue. We'll discuss during weekly check-ins.

Summary Checklist

For each sample, I have:

- [ ] Read the original text carefully
- [ ] Read the generated text carefully
- [ ] Compared them systematically
- [ ] Assigned SP score (1-3) with justification in mind
- [ ] Assigned SC score (1-3) based on error count
- [ ] Recorded scores and entry_id
- [ ] Added notes if exceptional case

Target: 200 samples per annotator across all models and types **Timeline**: Complete in 2 weeks (50-60 samples per day) **Expected Duration**: 20-30 hours total per annotator

Version 1.0 | September 2025 | For Internal Use Only