Supplementary Material

From Detection to Counterspeech: Auditing AI Moderation and Fact-Checking Practices in Ethiopia's Multilingual Online Sphere

This document provides supplementary methodological detail for the computational audit and the document analysis. Interview materials are not reproduced here: participant consent and safety preclude their release. The full annotated corpus, the classifier-audit code, and a script that regenerates the article's Table 1 are openly available at https://github.com/Endalk-Chala/ethiopia-ai-moderation-audit, archived with a persistent DOI. Code and data are not reproduced here.

1. Corpus and Annotation (Computational Audit)

1.1 Corpus composition and event coverage

The audit corpus comprises 951 in-scope Amharic and Afan Oromo social-media posts, purposively sampled from political-activist pages, influencers, and ordinary users on public Facebook, X, and TikTok (2020–2025). The corpus spans major Ethiopian political events of the period: the Tigray conflict, the 2020 assassination of Oromo singer Hachalu Hundessa, COVID-19, and the Amhara and Oromia conflicts. After two-coder annotation, reconciliation, and exclusion of cases not mapping onto the binary decision, 838 items were audit-eligible (Amharic n = 470; Afan Oromo n = 368). Gold labels were binarized (Hate = 1; Counter and Neutral = 0), yielding 597 Hate and 241 not-Hate items.

The deposited corpus file holds 1,007 physical rows: the 951 in-scope items plus 56 English-language rows retained for provenance and excluded from analysis. Each ineligible row carries an ExclusionReason value, so the reduction from 1,007 rows to the 838 audited items reconciles row by row.

1.2 Coding scheme and collapse rules

Each item was independently labeled by two coders under a three-class scheme, then collapsed to a binary gold standard for the audit:

Class	Definition (as applied in this study)	Binary label
Hate	Discourse attacking or demeaning a group on the basis of protected characteristics, following Sponholz (2021); boundaries treated as contested in public deliberation rather than as a fixed legal category.	1
Counter	Counterspeech: speech responding to, challenging, or correcting hate speech (Benesch, 2014; Buerger, 2021).	0
Neutral	Content that is neither hateful nor counterspeech (e.g., reporting, commentary, or unrelated discourse).	0

Collapse rule: Hate → 1; Counter and Neutral → 0. Items that did not map onto this binary decision were excluded from the audit-eligible set.

1.3 Inter-annotator agreement

Agreement on the underlying three-class scheme was strong, Cohen's $\kappa = .971$ (n = 783; Amharic .955, Afan Oromo .993). The three-class figure excludes rows where either coder assigned Misinformation or

Ambiguous. On the full six-class scheme, $\kappa = .786$ ($n = 951$). The protocol target was $\kappa \geq .70$, met on both schemes and in both languages.

2. Classifier Audit Protocol (Computational Audit)

Three classifiers were evaluated across tiers of language specificity: Perspective API (generic); a fine-tuned AfroXLMR model trained on the AfriHate dataset (Africa-centric; Muhammad et al., 2025); and an Amharic mBERT model (language-specific). Perspective API, which supports neither language, was evaluated in native mode and on human English translations. The AfroXLMR model was fine-tuned for this study because no deployable checkpoint is released; Amharic mBERT was applied to the Amharic subset only. Performance was assessed using precision, recall, and F1 against the binarized gold standard, disaggregated by language. Over-enforcement was operationalized as false positives on Counter or Neutral items, and under-enforcement as false negatives on Hate items. The audit follows external algorithmic auditing for contexts without direct access to proprietary systems (Bandy, 2021; Sandvig et al., 2014).

2.1 Model provenance, fine-tuning, and evaluation settings

Setting	Value
Perspective API	commentanalyzer v1alpha1, TOXICITY attribute. Native mode: languages=['am'] for Amharic, auto-detect for Afan Oromo. Translation mode: languages=['en'] on the human-produced English gloss. Hate = score ≥ 0.50. Scores collected June 2026; the hosted model is versioned by Google and cannot be pinned.
AfroXLMR (AfriHate)	Davlan/afro-xlmr-base-hate-v1, fine-tuned on the AfriHate Amharic training split. Seed 42, 3 epochs, max_length 128, batch size 16, learning rate 2e-5, macro-F1 model selection. Validation macro-F1 = 0.714.
Amharic mBERT	amengemeda/amharic-hate-speech-detection-mBERT, applied to the Amharic subset only.
Inference	Batch size 32, max_length 256, inputs truncated to 512 characters, evaluated under torch.no_grad().
Metric convention	Coverage-conditional: precision, recall, and F1 computed over rows a classifier actually scored, with coverage reported separately. Perspective covers 82% of eligible rows and Amharic mBERT 56%, so the convention is consequential and is stated alongside every figure.

The full audit protocol and analysis code are in the project repository; reproduce.py verifies the corpus against every figure reported in the article and regenerates Table 1.

3. Document Analysis: Source Inventory

Sources comprise platform self-descriptions and policies, transparency reports, Meta Oversight Board decisions on AI and automation, and triangulating civil-society, academic, and journalistic materials. Analysis used a thematic coding scheme focused on platform accountability and the visibility of Ethiopian-language moderation.

Category	Sources (as drawn on in the main text)
Platform policies / self-descriptions	Meta/Facebook Community Standards; TikTok Community Guidelines (prohibited-behavior taxonomies; intervention typologies beyond removal); X rules.
Transparency reports	Meta transparency reporting; Twitter-era Transparency Reports (through H2 2021); X Global Transparency Report (X Corp.); TikTok transparency reporting.
Oversight Board	Meta Oversight Board decisions on AI and automation, including Ethiopia-relevant matters tied to the 2021 election.
Triangulating materials	Amnesty International (2023) — disclosures of Ethiopian-language moderation capacity surfaced through Kenyan litigation; Azoulay (2024); civil-society, academic, and journalistic reporting.

The three themes organizing the document analysis in the main text are: visible rules and less visible operations; crisis-visible Ethiopia; and infrastructural opacity and the limits of metrics.