POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Auditing Exposure to Harmful Content on TikTok using Multimodal Language Models: A Cross-National, Age-Stratified Study

Laurea Magistrale in Computer Science and Engineering – Ingegneria Informatica

Author: Hamidreza Saffari

Advisor: Prof. Francesco Pierri

Academic year: 2025-2026

1. Introduction

TikTok is one of the most influential gateways through which young users encounter video content online, and its For-You feed adapts rapidly to micro-interactions such as watch time. Independent audits have documented self-harm content reaching newly created teen accounts within hours and age-gating that does not meaningfully shield underage personas relative to adults, while the EU Digital Services Act has made reproducible audits of short-video platforms increasingly urgent. The two age-stratified TikTok audits closest to this work rely on manual annotation of a limited subsample [1, 5], which bounds the scale at which per-country exposure can be measured.

The obstacle is a measurement bottleneck. Deciding whether a short video is harmful requires watching it, reading its caption, and often parsing its audio, consistently and across several languages; human annotation does not scale to tens of thousands of videos, and the judgment is itself subjective and culturally dependent. Aligning the harm taxonomy with TikTok's own Community Guidelines lets the audit measure the platform against the standard it claims to enforce, rather than against an external definition of harm, but it does not remove the labelling cost.

One way past that bound is to enlist multimodal large language models (MLLMs) as automated annotators of harmful video. Recent work shows MLLM judgments that align with humans and uses sampled frames plus text metadata to label large video corpora [3]. At the same time, different MLLM moderators disagree substantially on the same items [2], so a single frontier model cannot be trusted without validation. This thesis asks how harm exposure on TikTok varies across *age*, *country*, and *input modality*, and whether a validated MLLM can measure it at full corpus scale.

We run a two-stage audit on three EU countries (France, Italy, Sweden) and four age personas (13, 16, 19, 40), via a within-account `scroll-pre` → `SEARCH` → `scroll-post` cycle that captures both baseline For-You-page (FYP) exposure and the platform's response to a harm-keyword probe. Stage-1 validates an MLLM auditor against native-speaker labels on a 300-video subset; Stage-2 runs that auditor on a phase-stratified 10% sample of the full 36,971-video corpus. The audit yields three contributions: empirical findings on harm exposure across countries and ages; a validated, low-cost MLLM annotation pipeline (approximately $50

of API spend); and released artifacts (a 13-category harm taxonomy, prompts, per-country keyword lists, and corpus metadata).

2. Methodology

Data collection. We audit TikTok through sockpuppet persona accounts organized on two axes, country and age. Each of France (FR), Italy (IT), and Sweden (SW) is paired with personas aged 13, 16, 19, and 40, for twelve (country, age) combinations, with three independent accounts each. Accounts report their age at registration, use the native system locale, and route traffic through a VPN endpoint in the target country, since TikTok keys recommendation and search to the IP-linked country. The three countries span the high, medium, and low end of TikTok's per-language EU moderator allocation in the DSA transparency data analyzed by Tonneau et al. [4] (French ∼620, Italian ∼396, Swedish ∼98 moderators).
The four ages bracket the policy-relevant thresholds: 13 is TikTok's stated minimum, 16 is mid-adolescence, 19 is the youngest age in the platform's adult tier, and 40 is an adult control, so the 13-vs-40 contrast bounds the effect of the age signal. Collection has two phases. The *passive* phase records what the algorithm surfaces under pure FYP scrolling (14,093 unique videos). The *active* phase probes keyword-level intent: each run follows a `scroll-pre` → `SEARCH` → `scroll-post` cycle, issuing three native-language keywords per harm category over a five-day window (22,878 unique videos), with the same account serving as its own control for whether the search contaminates the subsequent feed. The two phases together cover 36,971 unique videos.

Harm taxonomy and annotation. Videos are judged under a 13-category harm taxonomy aligned with TikTok's Community Guidelines, spanning disordered eating, self-harm, dangerous challenges, nudity, sexually suggestive and shocking content, hate speech, sexual and physical abuse, trafficking, gambling, regulated substances, integrity, and harassment. Each label is a three-way verdict (harmful, not harmful, or unavailable when the embed fails or the content is region-locked), with a required primary and optional secondary subcategory when harmful and a low/medium/high confidence rating. The same taxonomy is injected into the auditor's system prompt, so that human and model judgments are made against identical definitions; aligning the taxonomy with the platform's own guidelines means a flagged video is, by construction, a candidate violation of the rules the platform claims to enforce.

Two-stage MLLM audit. Stage-1 selects the auditor. Four MLLMs (Gemini 2.5 Flash, Qwen3-VL-32B, GPT-4o-mini, Mistral Large 3) are scored against a 300-video reference set, randomly sampled and stratified by age, independently labeled by two native-speaker annotators per country and reconciled by joint resolution. Each model is evaluated under three input conditions: E1 (text only: caption plus audio transcript), E2 (native video upload), and E3 (eight frames sampled uniformly, plus caption and transcript, following the frame-plus-metadata protocol of Jo et al. [3]). All conditions share a system prompt that injects a 13-category harm taxonomy aligned with TikTok's Community Guidelines.
Gemini 2.5 Flash under E3 is the strongest configuration at aggregate Cohen's $\kappa = 0.42$, improving monotonically with visual input (E1 $0.17 \rightarrow$ E2 $0.38 \rightarrow$ E3 0.42; Figure 1) at roughly $2\times$ lower per-call cost than native-video upload. Stage-2 applies this auditor to a phase-stratified 10% sample of the corpus under both E3 (primary) and E2, for approximately $49 of API spend; text-only E1 is dropped because no model clears fair agreement on text alone.

Validation. The task is genuinely hard: even the winning configuration does not clear moderate agreement in every country, and per-country agreement for Gemini E3 is $\kappa_{\mathrm{FR}} = 0.290$, $\kappa_{\mathrm{IT}} = 0.463$, $\kappa_{\mathrm{SW}} = 0.356$, tracking the per-country class balance and the pre-resolution two-annotator agreement (aggregate 0.54) rather than a country-specific model deficit. The monotonic gain from E1 to E3 (Figure 1) is the evidence that the auditor uses the visual input rather than answering from the caption alone, which is the failure mode the text-only E1 baseline is designed to surface. Against the reconciled reference the model is conservative, under-flagging more than it over-flags

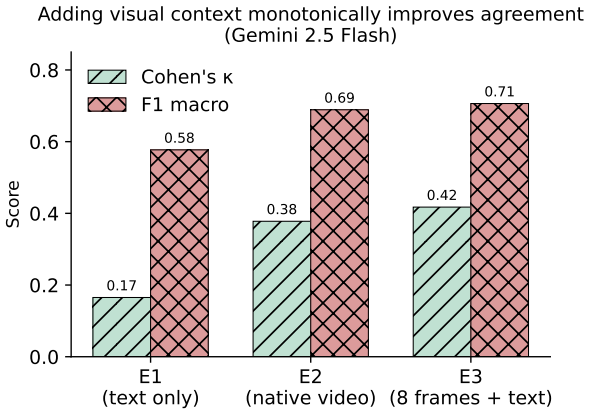

Figure 1: Cohen's κ and macro-F_1 across the three input conditions for Gemini 2.5 Flash on the 300-video Stage-1 subset. Agreement rises from text-only (E1) to eight frames (E3); E3 is strongest at ∼2× lower per-call cost than native video (E2).

(48 false negatives vs. 24 false positives), so its Stage-2 estimates are best read as a validated lower bound on exposure.

3. Results

Cross-country, cross-age prevalence. Pooled over the Stage-2 sample under Gemini E3, Italy carries the highest harm rate at all four age personas, with IT-16 the most-exposed case in the audit at 40.4%; France ranges from 23.6% to 32.4% and Sweden from 24.2% to 31.0%. Within France and Sweden the rate is relatively flat across ages (a 7–9 percentage-point spread), whereas Italy inverts the youth-safety-first prior, with its 13-year-old persona more exposed than its adult. Sexually suggestive content is the dominant harm category in every (country, age) combination, from 21.7% of harmful items (SW-13) to 50.0% (IT-16). The For-You feed and the search endpoint also differ sharply in reach: FYP-surfaced videos carry play counts roughly an order of magnitude higher than search-surfaced ones, so the harmful content a determined user reaches by querying is, on average, far less viral than what the feed recommends.

Harm-keyword search is the main driver of exposure. The within-account cycle (Figure 2) shows that the `SEARCH` endpoint returns 35–56% harmful content, a 1.5–7.5× lift over the immediate `scroll-pre` baseline in ten of twelve combinations, with no visible search-time block or warning. The lift is short-lived: `scroll-post` reverts to within a few percentage points of `scroll-pre` within the same session. The probe also collapses the FR/SW age gradient, since the youngest French and Swedish personas reach 37–42% harm under search, close to the adult combinations in the same countries. This is search returning what the keyword asks for in the absence of search-time moderation, not recommender-side amplification, and it means a passive-only audit understates reachable harm by an order of magnitude in many combinations. The largest lifts fall on the older French and Italian personas, whose `scroll-pre` rate sits below 25% and whose `SEARCH` endpoint returns 50–56% harmful items before reverting; Sweden's adult combination shows the same shape at smaller magnitude. The two exceptions are IT-13 and IT-16, already at 34–44% harm at `scroll-pre`, where the probe has no headroom to lift further; the contrast with the same-age French and Swedish combinations (at 9–19% pre-probe) indicates that country, not persona age, sets the headroom on this audit window. Across all twelve combinations the `scroll-pre` and `scroll-post` confidence intervals overlap and the absolute change is bounded by a few percentage points, so the search elevates exposure during the search session itself rather than as a persistent drift in the recommendation feed; a day-by-day breakdown confirms both the lift and the Italian ceiling effect hold on every day of the five-day window.

Passive exposure is highest in Italy at every age. Under purely passive FYP scrolling (Figure 3), Italy carries the highest harm rate at every age persona. The cross-country gap widens from a 4-point spread at age 13 (FR 24.5%, IT 27.9%, SW 23.7%) to over 25 points at age 19 (FR 22.5%, IT 48.6%, SW 38.4%); Italian age-19 is the most-exposed case in the audit at 48.6%. France stays flat near 23% across all four ages, while Sweden rises from 24% to 38%. Italy's two youngest combinations are already at 34–44% harm at `scroll-pre`, leaving the search probe little headroom. Several factors could explain the Italian pattern (content supply, the Italian-language moderator workforce,

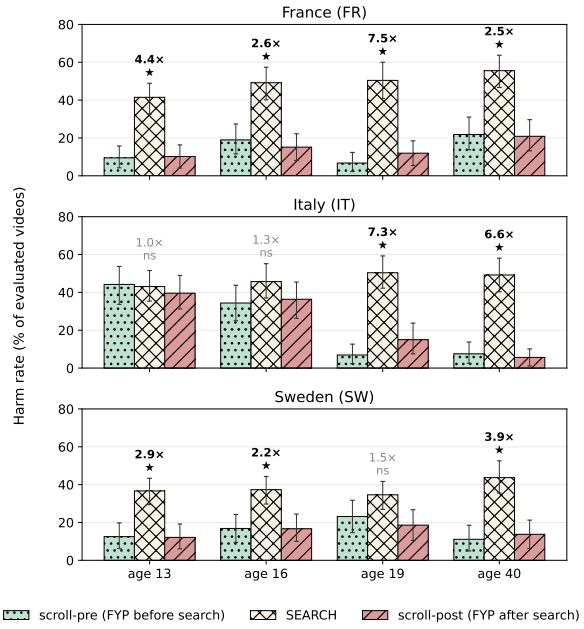

Figure 2: Per-combination harm rate at the three steps of the active cycle (`scroll-pre` → `SEARCH` → `scroll-post`) under Gemini E3, with 95% CIs. Annotated: the `SEARCH`/`scroll-pre` ratio; stars mark non-overlapping CIs.

persona behavior, or annotator thresholds); the cross-country ranking is a measurement, not an account of cause.

Modality and provider blocks. At scale, native-video E2 reports harm rates 3–12 percentage points higher than E3 while preserving the cross-combination ordering (binary-verdict agreement $\kappa = 0.614$ over 5,108 paired items). The Gemini safety layer refuses about 1.1% of inputs, and these refusals are not noise-distributed: they concentrate on the most explicit categories (Nudity and Sexually Suggestive content), so reported prevalences are underestimates for exactly the categories the audit targets. A worst-case sensitivity bound shifts headline rates by at most 1–3 percentage points and preserves both the IT > SW ≥ FR ordering and the Italian ceiling effect.

4. Conclusions

This work audits TikTok harm exposure across three EU countries and four age personas with a validated, two-stage MLLM pipeline. Two findings stand out. First, harm-keyword search rather than recommender-side amplification is

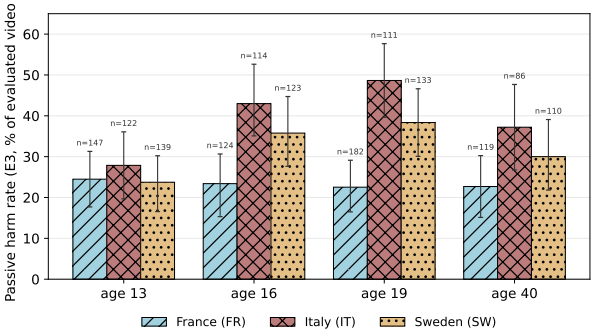

Figure 3: Stage-2 harm rate per (country, age) on the passive phase (FYP scrolling only) under Gemini E3, with 95% video-level bootstrap CIs. Italy leads at every age and the cross-country gap widens with age.

the largest lever in the exposure budget within our audit window: the search endpoint returns 35–56% harmful content with no visible moderation, and the elevated exposure does not persist into the post-search feed. Second, baseline algorithmic exposure is sharply uneven across countries, with Italian-speaking users the most exposed at every age and Italian age-19 reaching 48.6%.

Methodologically, MLLM-based auditing scales cross-national youth-safety audits at API costs of order $50, with annotation throughput rather than compute as the remaining bottleneck, subject to two caveats: the unresolved E2-vs-E3 modality gap, and provider safety filters that bias the most explicit categories downward. Both caveats point the same way, toward a Stage-2 human annotation pass that would convert the current validated lower bound into a calibrated point estimate. For the platform, the two levers that follow most directly are search-time moderation, scoping the returned pool by the same taxonomy already enforced on the recommendation feed, and per-language moderation infrastructure, noting that exposure does not line up monotonically with per-language moderator allocation, since the most exposed country (Italy) sits in the middle of the allocation range rather than at the bottom; the audit measures this co-occurrence but cannot establish the moderator workforce as its cause. The audit pipeline reported here is reusable as a regression test against either intervention, and the same Stage-1 plus Stage-2 design transfers to other short-video platforms under the same per-

sonas.

5. Acknowledgements

I thank my advisor, Prof. Francesco Pierri, for his guidance, and the native-speaker annotators who labeled the Stage-1 reference subset across French, Italian, and Swedish content.

References

[1] Fatmaelzahraa Eltaher, Rahul Krishna Gajula, Luis Miralles-Pechán, Patrick Crotty, Juan Martínez-Otero, Christina Thorpe, and Susan McKeever. Protecting young users on social media: Evaluating the effectiveness of content moderation and legal safeguards on video-sharing platforms, 2025. URL https://arxiv.org/abs/2505.11160.

[2] Neil Fasching and Yphtach Lelkes. Model-dependent moderation: Inconsistencies in hate speech detection across LLM-based systems. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-acl.1144/.

[3] Claire Wonjeong Jo, Miki Wesołowska, and Magdalena Wojcieszak. Harmful YouTube video detection: A taxonomy of online harm and MLLMs as alternative annotators, 2024. URL https://arxiv.org/abs/2411.05854.

[4] Manuel Tonneau, Diyi Liu, Ryan McGrady, Kevin Zheng, Ralph Schroeder, Ethan Zuckerman, and Scott A. Hale. Language disparities in moderation workforce allocation by social media platforms. SocArXiv preprint, August 2025. URL https://osf.io/preprints/socarxiv/amfws.

[5] Linda Xue, Francesco Corso, Nicolo Fontana, Geng Liu, Stefano Ceri, and Francesco Pierri. Towards an automated framework to audit youth safety on TikTok. In *Proceedings of the Fourth Workshop on Bridging Human–Computer Interaction and Natural Language Processing (HCI+NLP)*, Suzhou, China, November 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.hcinlp-1.9/.