Measuring the State of Online Disinformation in Europe on Very Large Online Platforms

First report of the SIMODS project

Structural Indicators to Monitor Online Disinformation Scientifically

Supported by

Executive Summary

The consortium led by Science Feedback and including Newtral, Demagog SK, Pravda, Check First, and the Universitat Oberta de Catalunya (UOC) completed the first large-scale, cross-platform, scientifically sound measurement of Structural Indicators of Disinformation. These indicators assess how permeable Very Large Online Platforms (VLOPs) are to misinformation and disinformation in Europe, how influential repeat misinformers are relative to credible sources, and the extent to which such content is monetised.

Amid debate over the Code of Conduct on Disinformation, now becoming a co-regulatory instrument under the DSA, and partial disengagement by some VLOPs from their commitments, this report offers comparable, evidence-based measurement to inform policy and enforcement.

WHAT WE MEASURED

Across six VLOPs (Facebook, Instagram, LinkedIn, TikTok, X/Twitter, YouTube) and four EU Member States (France, Poland, Slovakia, Spain), we report four Structural Indicators: **Prevalence** of mis/disinformation; **Sources** (relative influence of repeat misinformers vs. credible actors); **Cross-platform** presence; and **Monetisation**.

Datasets studied were either platform-provided (LinkedIn) or the result of large-scale keyword searches on high-salience topics (Ukraine/Russia, climate, health, migration, national politics). The corpus covers ~2.6 million posts totalling ~24 billion views. A view-weighted random sample (500 posts per platform and per country) approximates widely seen content; professional fact-checkers annotated posts to assess misinformation.

> **Data access note.** Despite DSA Article 40.12 requests, only LinkedIn supplied the requested random sample; TikTok granted API access too late for inclusion in this wave. This underscores persistent access gaps affecting independent audits.

KEY FINDINGS

1) **Prevalence**. TikTok shows the highest prevalence of mis/disinformation (~20% of exposure-weighted posts). Facebook (~13%) and X/Twitter (~11%) follow; YouTube and Instagram are at ~8%; LinkedIn is at ~2%.
   When including abusive (e.g., hate speech) and borderline content (content supporting a disinformation narrative without making a verifiably false claim), both of which

contribute to a less-informed public debate, prevalence rises to ~34% on TikTok, ~32% on X/Twitter, ~27% on Facebook, ~22% on YouTube, ~19% on Instagram, and ~8% on LinkedIn.

2) **Sources – "misinformation premium".** Accounts that repeatedly share misinformation (low-credibility) attract more engagement per post per 1 000 followers than credible (high-credibility) accounts on all platforms except LinkedIn.
   The ratio (low/high) is most pronounced on YouTube (~8×) and Facebook (~7×); it is ~5× on Instagram and X/Twitter, and ~2× on TikTok. This indicates systematic amplification advantages for recurrent misinformers. Only on LinkedIn is this premium absent, meaning that sharers of misinformation are not rewarded with extra visibility there.

3) **Cross-platform footprint.** Low-credibility actors are more likely than high-credibility actors to maintain accounts on X/Twitter (+34%), Facebook (+23%), and TikTok (+17%); the inverse holds for LinkedIn (-80%) and Instagram (-33%).

4) **Monetisation.** None of the assessed services fully prevents monetisation by recurrent misinformers. On YouTube, ~76% of eligible low-credibility channels are monetised (vs. ~79% for high-credibility). On Facebook, ~20% of eligible low-credibility Pages appear monetised (vs. ~60% for high-credibility). Google Display Ads appear on 27% of low credibility websites (vs. ~70% for high-credibility). Transparency limits prevented equivalent auditing on X/Twitter, TikTok, LinkedIn, and Instagram.

WHY THIS MATTERS

Under the DSA's systemic-risks framework, platforms must reduce the spread and impact of misleading content and avoid incentivising it financially. Our results show that misleading content is prevalent across platforms, recurrent misinformers benefit from a persistent engagement premium, and demonetisation is not fully operational, especially on YouTube. These patterns are inconsistent with an online environment that reliably privileges trustworthy information.

LIMITATIONS AND NEXT STEPS

This wave reflects a first collection period (first half of 2025) and constrained data access on some services. A second measurement will be published in early 2026 to track changes over time. Future iterations should benefit from improved platform data access and expanded monetisation coverage as Article 40 of the DSA requires platform cooperation.

1. Introduction

Many platforms have recently stepped back from earlier commitments to counter disinformation, for instance by reducing fact-checking programmes, staffing in relevant teams, or specific pledges. These policy shifts have been widely reported since early 2025 and are often framed as responses to political pressure in the United States.

Some claim that platforms are saturated with misinformation, while others argue that misleading content constitutes only a small fraction of users' exposure[1-2]. This ambiguity underscores the need for robust, comparable measurement so debates, and any resulting regulation spearheaded in Europe, are anchored in evidence rather than assertion.

1.1 THE CODE OF CONDUCT ON DISINFORMATION

The Code of Conduct on Disinformation (formerly the Code of Practice) is a co-regulatory instrument co-developed by the European Commission with online platforms, search engines, the advertising industry, fact-checkers and civil society. Signatories commit to a set of measures including (among others) promoting trustworthy sources, reducing the amplification of misleading content, demonetising disinformation, increasing transparency of political advertising, partnering with fact-checkers, and enabling researcher access to data[3].

On 13 February 2025, the Commission and the European Board for Digital Services formally integrated the 2022 Code into the Digital Services Act (DSA) framework, turning it into the Code of Conduct on Disinformation. As of 1 July 2025, the Code is operational under the DSA, with auditing and compliance mechanisms, meaning that the Code has become a "*significant and meaningful benchmark for determining compliance with the [DSA]*"[4].

1.2 STRUCTURAL INDICATORS

The concept of Structural Indicators was first introduced in the Commission's Guidance on Strengthening the Code of Practice on Disinformation, which called for Key Performance Indicators (KPIs) to track both implementation and effectiveness of the Code. These KPIs are structured into two complementary sets: Service-level Indicators, which assess the results and impact of specific policies; and Structural Indicators, which evaluate the broader systemic impact of the Code.

In response to this guidance, the European Digital Media Observatory (EDMO), specifically through the work of the Centre for Media Pluralism and Media Freedom (CMPF), developed

an initial set of Structural Indicators aimed at capturing the evolution and characteristics of online disinformation over time[5].

EDMO proposed indicators comprising a core set including: Prevalence of disinformation, Sources of disinformation, Audience of disinformation, and Collaboration and investments in fact-checking, and an extended set including Users' resilience, Demonetisation, Cross-platform disinformation, and Algorithmic amplification[6].

To compare how permeable each platform is to misleading content and how welcoming it is for actors spreading it, indicators must be defined in a way that is comparable across platforms and stable over time so that progress, or deterioration, can be quantified.

It is with this objective that the SIMODS (Structural Indicators to Monitor Online Disinformation Scientifically) project was designed: to provide independent, external measurement of key Structural Indicators and assess whether platforms respect users' rights to be informed truthfully and not manipulated and comply with the EU framework.

1.3 SIMODS

SIMODS (Structural Indicators to Monitor Online Disinformation Scientifically) is a project led by Science Feedback, in partnership with the Universitat Oberta de Catalunya (UOC), Check First, and fact-checking organisations Newtral, Demagog SK, and Pravda. The project measures four Structural Indicators:

1) Prevalence of Disinformation;
2) Sources of Disinformation;
3) Monetization of Disinformation;
4) Cross-platform Aspects of Disinformation.

Measurements cover six Very Large Online Platforms (VLOPs) and four countries: France, Poland, Slovakia and Spain. Under the DSA, a VLOP is designated at ≥45 million average monthly active recipients in the EU (≈10% of the EU population).

This European Media and Information Fund (EMIF)-funded project spans 18 months. This report presents the first measurement period; a second measurement and follow-up report will be published in early 2026.

While previous attempts have been made to measure Structural Indicators, most notably TrustLab's pilot implementation[7,8], they did not deliver a full prevalence metric, in part due to limited data collection scale. Indeed, it is not an easy feat to collect data at the scale required to produce meaningful and statistically robust results.

Despite the DSA's data-access provisions for researchers (Article 40), most platforms did not provide datasets in time for our analysis following our requests. Only LinkedIn provided the random sample that we requested, and TikTok granted API access too late for inclusion in this data-collection period.

SIMODS succeeds in delivering these measurements on Structural Indicators through an approach that:

- relies on large-scale datasets, which allows our results to be representative of contents that are highly viewed on each platform;
- rely on professional fact-checkers to assess whether each piece of content contains mis/disinformation, as they possess the most relevant expertise for this task given their experience through their daily work identifying and debunking false claims;
- applies rigorous protocols and statistical analysis, reviewed by UOC researchers within the consortium.

NOTE: WHY THE TERM MIS/DISINFORMATION?

The Code uses the term "disinformation" to cover "verifiably false or misleading information that is created, presented, and disseminated for economic gain or to intentionally deceive the public, and that may cause public harm".

This functional definition encompasses what is traditionally termed misinformation (false or misleading information spread unintentionally) and disinformation (misinformation spread intentionally). In this report, we apply the Code's definition when assessing content but use the shorthand "mis/disinformation" to avoid ambiguity and make explicit that both intentional and unintentional false or misleading content are included.

2. Findings

2.1 PREVALENCE of MIS/DISINFORMATION

The first and most direct indicator of the scale of the disinformation issue on a platform is its prevalence, that is, the proportion of the content that users are exposed to on the platform that contains mis/disinformation.

As EDMO explains, prevalence "*aims to measure how widespread disinformation is across platforms. As such, the share of content identified as disinformation in a selected sample of random content should be measured*"[6]. In reaction to EDMO's 2nd report, a group of experts who provided feedback on structural indicators further explained that prevalence should be measured "*by comparing it to content on similar topics rather than all non-disinformation content*"[9].

With this background information, we set out to measure prevalence consistently across the six very large online platforms. To do so, we collected hundreds of thousands of pieces of content on topics that are central to the public debate in Europe and at high risk of carrying mis/disinformation and asked professional fact-checkers to determine which posts contained mis/disinformation.

It is important to note that previous attempts to measure prevalence, such as TrustLab's 2023 pilot, were unable to construct a reliable measure of prevalence due to the limited scale of their data collection. Instead, TrustLab's study produced a metric of "discoverability" (or "findability"), i.e. the share of mis/disinformation among search results for disinformation-related keywords[7,8]. While valuable, this metric reflects what users find when they explicitly search for problematic content, rather than what they are incidentally exposed to in their everyday browsing. Our approach represents a significant methodological advance: it allows us to construct large, representative samples of content that reflect actual user exposure, thereby producing the first robust cross-platform, cross-country measure of prevalence.

2.1.1 Data Collection & Processing

A. KEYWORDS-BASED SEARCH

To approximate the information environment that users encounter, we built our corpus through keyword searches on topics of high public interest in Europe and high risk of mis/disinformation: the Russia-Ukraine war, climate change, health, migration, and national politics.

To minimize bias and allow meaningful comparison across countries, most keywords were translated identically across the four languages of the study. This was notably the case for Ukraine, climate, health, and migration topics. In contrast, local politics keywords were adapted to the national context in each country to ensure relevance.

In order not to bias our sampling towards mis/disinformation only, and to properly capture the diversity of content users are exposed to on platforms, the keyword lists, tested and designed by professional fact-checkers, included keywords in these three categories:

- **Neutral** terms (e.g. Zelensky, migrants, Covid-19): widely used across all information sources, ensuring that our dataset included mainstream reporting and discussion.
- **Ambiguous** terms (e.g. vaccine side effects, geoengineering, laboratories in Ukraine): terms often encountered in misleading narratives but that are not specific to it and also legitimately used in scientific or journalistic contexts.
- **Misinformation-related** terms (e.g. climate scam, Ukrainian Nazi, remigration): that are predominantly used in false or misleading claims and are unlikely to be used by credible sources when speaking about the topic.

The final list of search terms included around 100 keywords per language (French, Spanish, Polish, Slovak) and was balanced, with equal numbers of *neutral* and *ambiguous* + *misinformation-related* terms in each country. More details about the keywords can be found in Appendix 5.1.1. The first data collection period spanned 17 March to 13 April, and we collected posts that are published between these dates (inclusive).

To collect data from the selected platforms, we employed two different methods. First, given that this project investigates a systemic risk (under DSA Article 34) to civic discourse, we invoked Article 40.12 of the DSA and contacted all six VLOPs on 19 December 2024 to request a random sample of 200 000 posts per language. As only LinkedIn provided the requested dataset, we relied on a second method for the other platforms, using their search functions and associated tools to retrieve large numbers of posts containing any of our keywords of interest. More details on the tools, filters, and procedures used can be found in Appendix 5.1.1.

As a result, for the entire data collection period, we assembled a dataset comprising approximately **2.6 million posts** (with metadata) across four languages and six platforms, totaling around **24 billion views**. Given the varied contexts in which keywords can appear, the dataset still contained irrelevant content such as celebrity gossip, entertainment, or sports news. To address this, we used a Large Language Model (LLM), GPT 4o-mini, to filter the corpus, retaining only posts relevant to our study, that is, content contributing to

public discourse on the state of the world, such as health, science, politics, climate change, or other societal issues with a direct impact on people's lives or understanding of society. More details are provided in Appendix 5.1.2.

B. RANDOM SAMPLE

From this corpus, we then drew a random sample of 500 posts per platform and country for annotation by fact-checkers. A crucial methodological aspect is that **sampling was weighted by the number of views**. For instance, a video with 1 million views was 100 times more likely to appear in our sample than one with 10 000 views. This weighting ensures that the annotated sample reflects what users are actually seeing, not just what platforms return in search. It also mitigates potential distortions: if a platform's search algorithm systematically downranks low-credibility content, a highly viewed misleading post would still have a high probability of being sampled. More details on sampling are provided in Appendix 5.1.2.

C. ANNOTATION

With the random samples prepared, professional fact-checkers annotated each post to determine whether it contained mis/disinformation.

While our primary focus was distinguishing mis/disinformation (as defined in the Code of Conduct on Disinformation) from credible information, real-world content doesn't always fit neatly into a binary classification. Pilot tests conducted before the annotation period led us to define a broader set of categories to capture nuance:

- **Mis/disinformation**: Content stating or clearly implying a verifiably false or misleading claim that may cause public harm.
- **Credible and informative**: Content conveying true or credible information on important matters about the state of the world (excluding trivia, gossip, or anecdotes).
- **Borderline**: Content feeding a misleading narrative without necessarily containing outright falsehoods, but potentially reinforcing false beliefs.
- **Abusive**: Content not containing mis/disinformation but involving harmful material such as hate speech, insults, spam, or incitement to harmful behaviour.
- **Unverifiable**: Content that cannot be assessed as either credible or mis/disinformation (e.g. opinion-based).
- **Irrelevant**: Content not about public affairs or scientific/political issues (e.g. entertainment, sports, religious content, cooking recipes without health claims, geographically irrelevant to Europe).

- **Other language**: Content not in one of the languages spoken in the targeted country or English.
- **Deleted**: Content unavailable at the time of annotation (e.g. removed from the platform).
- **Don’t know**: Content not fitting any other category.

annotated as belonging to each of the main categories.

For the analysis, items labelled *Irrelevant*, *Other language*, *Deleted*, and *Don't know* were excluded. See Figure 2.1 for an illustration of the type of posts that were labeled in each of the main categories.

To ensure the robustness of our findings, each country had one fact-checker annotate the full dataset (500 posts per platform), while a second fact-checker independently reviewed a random subset of 100 posts per platform. Where content was labelled *Don't know*, the second fact-checker systematically reviewed it, and their judgment was retained. Once the data sample was fully labeled, the two fact-checkers discussed cases where discrepancies between the labels occurred and agreed on a final label for each piece of content.

This cross-verification was critical as it allowed us to account for the uncertainty and inevitable degree of subjectivity inherent in any annotation task. All results presented below include confidence intervals that quantify the uncertainties coming from both the sample size and the inter-annotator disagreement. Further details on annotation and confidence intervals can be found in Appendix 5.1.4.

2.1.2 Results

Once the data was processed and annotated by fact-checkers as outlined above, we were able to quantify the prevalence of posts belonging to each category.

A. PREVALENCE ACROSS CATEGORIES

Figure 2.2 shows an overview of the content breakdown across the six platforms, using the merged datasets from the four countries. The first observation is that the combined *Credible* and *Unverifiable* categories represent the majority of content on all platforms. We argue that these categories represent content that is legitimate to find on platforms. *Credible* content is intended to inform users on important matters regarding politics, health, science, etc., while *Unverifiable* content typically reflects people's opinions, commentaries, and thoughts about news and world events.

The distribution of *Credible* content is not uniform across platforms, with LinkedIn having the highest proportion at 59%, and X/Twitter having the lowest at only 27%. However, content that is generally harmful to users or society (the combination of *Abusive*, *Borderline*, and *Mis/disinformation*, which we collectively refer to as "*Problematic*" content in the rest of this analysis) can be found across all platforms. The share of *Problematic* content varies by platform, with TikTok and X/Twitter showing the highest levels at 34% and 31.5%, respectively.

Figure 2.2 – Percentage of posts belonging to each category for the six very large online platforms.

When comparing only *Credible* to *Problematic* content, we note that X/Twitter contains more *Problematic* content (31.5%) than *Credible* content (27%), while this is not the case for the other platforms (see Figure 5.2).

B. PREVALENCE OF MIS/DISINFORMATION

To assess the prevalence of mis/disinformation as required by the Code of Conduct on Disinformation, we calculated the ratio of content containing mis/disinformation compared to legitimate content on similar topics.

We define the prevalence metric $P_{misinfo}$ as:

$$P_{misinfo} = \frac{N_{misinfo}}{N_{misinfo}+N_{cred}+N_{unverif}} \times 100$$

where $N_{misinfo}$, N_{cred} and $N_{unverif}$ are the numbers of posts labeled as *Mis/disinformation*, *Credible*, and *Unverifiable*, respectively.

Figure 2.3 presents the values for the prevalence of mis/disinformation across platforms, using the merged dataset combining the four countries.

The results show significant differences between platforms:

- TikTok exhibits the highest prevalence of mis/disinformation at approximately 20% [17.7%, 22.6%], indicating that roughly one in five posts on the platform regarding the topics we investigated contains misleading or false information.
- Facebook and X/Twitter follow with elevated prevalence at 13% [11.5%, 15.4%] and 11% [9.0%, 12.2%], respectively.
- YouTube and Instagram have a prevalence of about 8%.
- LinkedIn has the lowest prevalence of mis/disinformation at 2% [1.3%, 3.2%], suggesting that exposure to misinformation on this platform is limited.

The confidence intervals displayed on the figure and mentioned in the text measure the uncertainty of our estimates; they measure both the uncertainty due to the size of our random samples and the uncertainty due to the labeling of content and potential disagreements between fact-checkers (see Appendix 5.1.4. for details). For those interested in measuring the proportion of all potentially misleading content (including both Mis/disinformation and Borderline content), refer to Appendix 5.1.5.B.

Figure 2.3 – Prevalence of mis/disinformation across the six very large platforms, aggregated across all languages. The error bars represent the 95% confidence intervals measuring the uncertainty around each estimate, calculated using a bootstrapping method (See Appendix 5.1.4 for details).

Platform	Prevalence [CI 95%]
LinkedIn	**2.3**% [1.3, 3.2]
Instagram	**8.0**% [6.4, 9.4]
Facebook	**13.4**% [11.5, 15.4]
YouTube	**8.5**% [6.9, 10.2]
TikTok	**20.2**% [17.7, 22.6]
X/Twitter	**10.6**% [9, 12.2]

Table 2.1 – Prevalence of mis/disinformation across the six very large platforms, aggregated across all languages (same values as on Figure 2.3). The confidence intervals (CIs) indicate the lower and upper bounds within which 95% of the estimates from the bootstrap calculation lie (See Appendix 5.1.4).

When considering all posts labelled Mis/disinformation, the topic with the highest share is health, representing 43.4% (Figure 2.4). The Russia-Ukraine war is the second most represented topic, with about one quarter of Mis/disinformation posts, followed by national politics (15.5%), which typically includes election-related claims in the country of interest or controversies surrounding new legislation, for instance. Climate and migration each account for 6.6% of Mis/disinformation posts.

our study.

C. PREVALENCE OF PROBLEMATIC CONTENT

Beyond content containing mis/disinformation, we have explained above that *Borderline* and *Abusive* content should also be considered to contribute to a less-informed public debate and not be confused with informative content. We propose adding their prevalence to the one of *Mis/disinformation* to create an indicator of the prevalence of harmful, or "problematic", content.

We calculate the prevalence of *Problematic* content P_{prob} as:

$$P_{prob} = \frac{N_{misinfo}+N_{bord}+N_{abus}}{N_{misinfo}+N_{bord}+N_{abus}+N_{cred}+N_{unverif}} \times 100$$

where $N_{misinfo}$, N_{bord}, N_{abus}, N_{cred} and $N_{unverif}$ are the numbers of posts labeled as *Mis/disinformation*, *Borderline*, *Abusive*, *Credible*, and *Unverifiable*, respectively.

Figure 2.5 shows that the prevalence of *Problematic* content is significantly higher than the prevalence of *Mis/disinformation* on all platforms.

With this metric, both TikTok and X/Twitter appear as the platforms with the highest prevalence of *Problematic* content, with prevalence values of 34% [31.7%, 36.5%] and 32% [29.9%, 34.1%], respectively (the overlapping confidence intervals indicate that the values for these two platforms are not statistically different). This means that roughly one in three posts on these platforms either contains false or misleading information, includes abusive language, reinforces misleading narratives, or promotes other forms of harmful content. Facebook ranks third with 27% [24.3%, 29.1%], followed by YouTube at 22% [20.0%, 24.5%], and Instagram with 18% [16.8%, 20.9%]. LinkedIn, as in previous analyses, reports the lowest prevalence at 7.5% [6.0%, 9.2%], indicating a comparatively safer information environment.

2.2 SOURCES of MIS/DISINFORMATION

A pervasive issue that has been frequently identified by researchers and civil society working on disinformation is that the most influential content is often produced or amplified by a limited set of actors who recurrently share misleading information. A small number of highly influential accounts can largely shape the spread of misleading narratives on a platform[10,11], and accounts that share mis/disinformation at a given point in time tend to continue doing so in the future[12].

To effectively measure the health of the information ecosystem on online platforms, it is crucial to examine the ability of these recurrent mis/disinformation sources to reach and influence large audiences. Furthermore, comparing the reach of these sources to that of

credible accounts provides insight into the platform's role in amplifying harmful content. The Code of Conduct on Disinformation encourages platforms to prioritize content from trustworthy sources while reducing the prominence of misleading or harmful content.

Figure 2.5 – Prevalence of *Problematic* content (defined by the grouping of *Mis/disinformation*, *Borderline* and *Abusive* content) across the six very large platforms, aggregated across all languages. The error bars represent the 95% confidence intervals measuring the uncertainty around each estimate, calculated using a bootstrapping method (See Appendix 5.1.4 for details).

Platform	Prevalence [CI 95%]
LinkedIn	**7.5**% [6.0, 9.2]
Instagram	**18.8**% [16.8, 20.9]
Facebook	**26.8**% [24.3, 29.1]
YouTube	**22.4**% [20.0, 24.5]
TikTok	**34.1**% [31.7, 36.5]
X/Twitter	**32.0**% [29.9, 34.1]

Table 2.2 - Prevalence of *Problematic* content across the six very large platforms, aggregated across all languages (same values as on Figure 2.5). The confidence intervals (CIs) indicate the lower and upper bounds within which 95% of the estimates from the bootstrap calculation lie (See Appendix 5.1.4).

To assess the effectiveness of the Code, the second Structural Indicator recommended by EDMO consists of measuring the characteristics and behaviors of accounts that repeatedly share mis/disinformation and comparing them to those of credible sources[6]. In response, our consortium developed metrics to compare the size of accounts' followings, their activity,

and the engagement their content receives across platforms. We propose using the average number of interactions per post per follower as a core structural indicator to estimate the relative influence of different sets of actors. This metric measures how much each platform helps amplify content from sources of misleading information compared to credible sources, while accounting for differences in their follower counts. We provide more details below.

2.2.1 Methodology

To contrast the engagement of misinformation spreaders with that of credible sources, we used two approaches to identify a list of accounts belonging to the two categories.

A. The Top 50 List Approach

One approach involved identifying the 50 most influential accounts on each platform and language based on the sample collected for the Prevalence section (Section 2.1). Accounts were ranked in descending order based on the cumulative number of views their content received in the dataset collected during the data collection period (March 17 – April 13). After excluding accounts that mostly shared content deemed irrelevant according to the project's definition (see Appendix 5.2.2 for details), we retrieved all posts published by these accounts during the same period, along with metadata such as the number of likes, comments, shares, and followers.

From this, we identified accounts that repeatedly shared mis/disinformation and those that were credible sources. Recognizing that not all accounts fit neatly into these two categories, we introduced a third category for accounts that do not belong to either group, such as those primarily sharing opinion-based content.

The categories used were:

- **Low-credibility:** Accounts that shared at least two posts containing false or misleading information.
- **High-credibility:** Accounts that almost exclusively shared credible and informative news, such as content from professional media outlets or scientific institutions.
- **Neither:** Accounts that did not fit into the two categories above, often sharing opinion-based content.

Fact-checkers in our consortium determined which category each account belonged to. To assist in the classification process, we used a Large Language Model (LLM), GPT 4o-mini, to classify posts based on their likelihood of belonging to the *Mis/disinformation*, *Credible*, or *Unverifiable* categories (as defined in the Prevalence section). Fact-checkers had access to this initial classification, which helped speed up the process, but the final decision

remained with the fact-checkers (see Section 5.2.2 for details on how we used an LLM for this task).

B. The Fact-Checkers' List Approach

Another approach involved asking fact-checkers to provide a list of accounts they know are frequent sources of misinformation and those they consider trustworthy, based on their day-to-day fact-checking activities. This list was developed independently of the Top 50 list. The fact-checkers' list typically included social media accounts frequently flagged for spreading false or misleading claims in their routine work. We also relied on the Consensus Credibility Scores, which aggregate multiple open-source credibility ratings for over 20 000 domains, to identify influential social media accounts associated with high or low credibility sources[13].

C. Comparison and Merging of the Two Lists

Upon comparing the two lists, we noted that the low-credibility and high-credibility sources from both datasets partially overlapped, giving a first indication of the robustness of the lists created. More importantly, we found consistent results on the average number of interactions per post per 1 000 followers for the low-credibility and high-credibility accounts using both the fact-checkers' and Top 50's lists. This consistency is a very important indication that the results discussed in this section are robust and do not depend on the specific methodology employed to construct the lists of low-credibility and high-credibility accounts. Consequently, we merged the two lists into one consolidated dataset for the results presented below. For a comparison of the results derived from the two lists, please refer to Appendix 5.2.3.

D. Note on handling of Political accounts

We treated accounts of politicians or political parties separately, categorizing them as 'Political'. These accounts were excluded from the primary analysis presented in the Results below. For results including political accounts, see Appendix 5.2.5.

2.2.2 Results

A. Accounts' Followership

The first observation is that high-credibility accounts consistently have larger audiences than low-credibility accounts. As shown in Figure 2.6, the average number of followers for accounts in the High-credibility lists is significantly higher than for those in the

Low-credibility lists across all platforms. Accounts classified as Neither generally have follower counts that are statistically similar to high-credibility accounts.

The confidence intervals for the High-credibility and Neither lists are relatively wide, reflecting the presence of accounts that have millions of followers more than other accounts in the dataset. These include well-known media organizations in the bootstrapping method (see Appendix 5.1.4 for details).

B. Accounts' Engagement Rates: The 'Misinformation Premium'

While followership provides insight into audience size, the Code of Conduct encourages platforms to increase the visibility of trustworthy content while reducing the amplification of misleading content, particularly from sources that repeatedly share mis/disinformation[3]. To capture this, we compared the average number of interactions per post per 1 000 followers across the different account groups. Normalizing by follower count allows us to fairly compare accounts of different sizes: given a similar audience, an account would be expected to receive comparable engagement for its posts.

Platform	High-credibility	Low-credibility	Neither
Facebook	**1 749** [1 191 - 2 398 104]	**318** [192 - 465]	**1 166** [551 - 1 906]
Instagram	**814** [606 - 1 051]	**155** [89 - 243]	**700** [467 - 985]
LinkedIn	**41** [18 - 69]	**6** [3.6 - 9]	**268** [127 - 450]
TikTok	**1 184** [799 - 1 664]	**271** [163 - 411]	**1 999** [663 - 3 745]
X/Twitter	**1 393** [641 - 2 479]	**304** [186 - 442]	**674** [165 - 1 494]
YouTube	**1 344** [1 029 - 1 688]	**428** [278 - 637]	**985** [701 - 1 325]

Table 2.3 – Average number of followers (in thousands) for accounts in the High-credibility, Low-credibility, and Neither lists on each platform, as displayed in Figure 2.6. Values in square brackets correspond to 95% confidence intervals.

Figure 2.7 shows that, across platforms, low-credibility accounts receive significantly higher interactions per post than high-credibility accounts. LinkedIn is the only exception, where differences are not statistically significant. The magnitude of engagement varies considerably across platforms: low-credibility accounts average approximately 5 interactions per post per 1 000 followers on Facebook, while on Instagram this figure is about ten times higher at about 46.

Considering the ratio of engagement for low-credibility versus high-credibility accounts, which can be seen as a "**misinformation premium**", the differences are striking (see Figure 2.8). On YouTube, low-credibility accounts receive more than 8 times the engagement of high-credibility accounts. Facebook shows a similar pattern (7.2×), while Instagram and X/Twitter exhibit ratios of approximately 5×. The lowest ratios are observed on TikTok (2×) and LinkedIn, where low-credibility accounts do not outperform high-credibility accounts in interactions per post per follower.

Figure 2.7 – Average number of interactions per post per 1 000 followers for accounts classified as High-credibility, Low-credibility, and Neither on each platform. Error bars represent 95% confidence intervals, calculated using a bootstrapping method (see Appendix 5.1.4 for details).

Platform	High-credibility	Low-credibility	Neither
Instagram	**9.0** [8.7 - 9.4]	**45.4** [34.3 - 60.0]	**30** [24.3 - 38.1]
Facebook	**0.76** [0.71 - 0.82]	**5.6** [5.2 - 6.0]	**1.9** [1.8 - 2.2]
X/Twitter	**2.6** [2.4 - 2.8]	**9.9** [9.2 - 10.7]	**7.2** [6.7 - 7.6]
Youtube	**0.67** [0.62 - 0.70]	**6.6** [6.2 - 7.1]	**4.2** [3.7 - 4.7]
TikTok	**19.5** [16.7 - 22.2]	**39.2** [31.8 - 47.8]	**36.4** [23.0 - 55.0]
LinkedIn	**29.2** [22.9 - 36.2]	**25.5** [5.81, 58.1]	**34.2** [26.4 - 44.1]

Table 2.4 – Average number of interactions per post per 1 000 followers for accounts in the High-credibility, Low-credibility, and Neither lists on each platform, as displayed in Figure 2.7. Values in square brackets correspond to 95% confidence intervals.

Note that this section compares platforms using the metric of interactions per post per follower. Although view-based metrics were available for some platforms, we did not manage to obtain views on posts we collected from sources of high and low credibility across all six platforms. To ensure comparability, we therefore report interaction-based metrics here. For reference, results based on views per post per follower are presented in Appendix 5.2.4. The total number of interactions were calculated by aggregating the number of comments, shares and likes of each post.

To ensure our findings are not an artefact of follower-count differences (i.e., low-credibility accounts typically having fewer followers, which can be correlated with interactions per post) we conducted a robustness test restricting the comparison to high- and low-credibility accounts with similar follower counts (see Appendix 5.2.6). The results are generally unchanged when stratifying by account size: low-credibility accounts still exhibit a significant interaction premium, confirming the robustness of the effect reported here.

C. Proportion of High/Low credibility accounts in the Top 50

Another indicator of how relatively influential repeat mis/disinformation accounts are compared to credible sources on a platform can be obtained by looking at the proportion within the Top 50 of accounts that are low-credibility versus high-credibility.

Figure 2.9 shows that the platform with the highest proportion of low-credibility accounts in the Top 50 is X/Twitter with about 34%, followed by TikTok with 29%. On Facebook and

YouTube, about 20% in the Top 50 are low-credibility accounts. The platform with the

2.3 CROSS-PLATFORM ASPECTS of MIS/DISINFORMATION

As demonstrated in the Prevalence and Sources sections, mis/disinformation and accounts that consistently publish such content can be found across all platforms. Research has shown that misleading narratives often travel between platforms[14]. Although a comprehensive study of how specific mis/disinformation narratives propagate across services is beyond the scope of this report, we aim to explore how welcoming different platforms are to sources of mis/disinformation. Specifically, this section investigates the existence of mis/disinformation sources across platforms, as recommended by EDMO's second report, which urges the investigation of "the existence of disinformation sources on other platforms based on users/accounts identified as sources of disinformation in the sample".

In alignment with this recommendation, we present indicators that assess the extent to which low and high credibility actors operate across different platforms and the audience sizes they manage to cultivate on each.

2.3.1 Methodology

Starting with the consolidated dataset of high- and low-credibility accounts developed in Section 2.2, which combines the Top 50 and fact-checker-identified accounts, we examined

whether these accounts maintained a presence on other very large online platforms. For each account, we manually searched other platforms using their name and username. Accounts were then labeled as 'active' if they had posted at least once in 2025 or 'inactive' if they had not posted during this period. Only active accounts were included in the study.

This process resulted in the creation of a cross-platform actors dataset, where an actor is defined as a collection of accounts operated by the same source across different platforms. In total, the dataset includes 341 high-credibility actors and 315 low-credibility actors, which amounts to 656 unique actors that have an active account on at least one of the six platforms analyzed.

2.3.2 Results

A. Platforms Favored by Low and High credibility Actors

Figure 2.10 illustrates the proportion of actors with active accounts on each platform. Among low-credibility actors in our dataset, Facebook is the most popular platform, with 50% maintaining an active account, followed by X/Twitter (43%) and YouTube (38%). High-credibility actors in this dataset, on the other hand, favor Instagram (45% of active accounts), followed by Facebook (41%) and YouTube (39%).

To understand how much more likely a low-credibility actor is to have an account on a given platform compared to a high-credibility actor, we calculate the ratio of the percentage of low-credibility actors with an account on the platform to the percentage of high-credibility actors with an account.

Our analysis shows that X/Twitter is the most comparatively attractive platform for low-credibility actors in this dataset, with a 34% higher likelihood of having an account than high-credibility actors (see Table 2.5). Similarly, Facebook (+23%) and TikTok (+17%) appear to be relatively more attractive to low-credibility actors. YouTube shows similar proportions for both groups, with both low- and high-credibility actors maintaining active accounts at approximately the same rate (38% versus 39%). In contrast, LinkedIn is the platform where low-credibility actors are least likely to maintain an account, with an 80% lower likelihood compared to high-credibility actors, followed by Instagram (33% lower likelihood).

Note that only a small portion of actors maintained a presence across all six platforms: 7 Low-credibility actors and 39 High-credibility actors. However, when excluding LinkedIn (the platform with the fewest accounts among low-credibility actors), the number of actors with accounts on the remaining five platforms increases to 28 for low-credibility actors and 55 for high-credibility actors. This suggests that credible sources have a greater ability to

maintain a significant cross-platform presence.

Figure 2.10 – Proportion of High- and Low-credibility actors with an active account on each platform.

Platform	High credibility proportion	Low credibility proportion	Ratio Low/High credibility proportions
Facebook	41%	50%	1.23
X/Twitter	32%	43%	1.34
YouTube	39%	38%	0.96
Instagram	45%	30%	0.67
TikTok	25%	29%	1.17
LinkedIn	27%	5.4%	0.20

Table 2.5 – Proportion of high- and low-credibility actors with an active account on each platform. The last column displays the ratios of low- to high-credibility proportions.

B. Ratio of Low-credibility to High-credibility Followership Size

To further assess how welcoming each platform is to accounts that repeatedly share mis/disinformation, we analyzed the number of followers these accounts manage to gather compared to high-credibility accounts on each platform.

By this measure, YouTube emerges as the platform most favorable to low-credibility actors, with the ratio of their average number of followers to that of high-credibility accounts

around 0.5 (see Figure 2.11). This indicates that low-credibility actors attract an audience size that is approximately half that of high-credibility sources, which typically represent professional teams dedicated to providing accurate information.

On Facebook, low-credibility actors have a followership size about one-third that of high-credibility actors, followed by TikTok (about one-quarter) and X/Twitter (about one-fifth). At the opposite end, LinkedIn and Instagram are platforms where low-credibility

In summary, these findings highlight the differential reception of low- and high-credibility actors across platforms. Platforms like YouTube and Facebook appear particularly favorable to low-credibility actors, both in terms of the number of accounts active on these platforms and their relative audience size. In contrast, LinkedIn and Instagram tend to be more favorable to high-credibility actors, reflecting a more professional user base.

These results should be considered in conjunction with the findings in Section 2.2.2, which showed that low-credibility sources consistently generate higher engagement, both in terms of interactions and, where available, views (see Appendix 5.2.4), compared to high-credibility accounts. Taken together, these findings suggest that while low-credibility actors attract fewer followers, often by a significant margin, their content achieves disproportionately high levels of engagement once posted.

This discrepancy highlights a structural imbalance in how content from low-credibility actors is surfaced and amplified by platforms, as well as how users engage with it. While High-credibility actors maintain a stronger follower base, low-credibility actors are able to leverage platform dynamics and user behavior to achieve visibility far beyond what their follower count alone would predict.

Platform	Ratio of Low- and High-credibility actors (# of followers)
Facebook	**0.31** [0.30, 0.32]
TikTok	**0.26** [0.24, 0.27]
YouTube	**0.47** [0.46, 0.48]
Instagram	**0.09** [0.08, 0.09]
X/Twitter	**0.21** [0.20, 0.22]
LinkedIn	**0.14** [0.13, 0.15]

Table 2.6 – Ratio of the average number of followers for low-credibility over high-credibility actors with an active account on each platform.

2.4 MONETIZATION of MIS/DISINFORMATION

Commitment 1 of the Code of Conduct sets out five Measures that online platforms and search engines should take to reduce the financial incentives for the production and dissemination of disinformation, at both the content and the account level. These Measures include platforms adopting policies to avoid placing ads next to mis/disinformation content, having systems in place to ensure that systematically-violative accounts cannot benefit from monetisation programs, and providing more transparency and third-party scrutiny over those Measures' effectiveness.

While platform Signatories to the Code (with the exception of LinkedIn) have unsubscribed from some of these Measures in early 2025, measuring the extent to which signatories are funding the production of mis/disinformation remains as relevant as it was when the Code was first negotiated. As such, a robust Code monitoring framework should include a Structural Indicator on monetization.

2.4.1 Methodology

A. DATA ACCESS

As pointed out in EDMO's second report, a methodologically-sound Structural Indicator on demonetization requires data that is, so far, inaccessible to outside researchers using

publicly available data. Two critical data points are currently not available for any of the platforms studied, specifically:

- the amount of revenue that a given account is generating from the platform, across monetization methods (e.g., ad-revenue sharing, tipping, creator marketing partnerships with brands, on-platform shops, etc),
- whether a given piece of content is contributing to the creator's revenue stream from that platform.

Because they are directly tied to financial payouts, it is highly likely that these data points are readily available. Future iterations of the Structural Indicators should request this data from the platforms under DSA Article 40.4, as the system is expected to become operational by the end of 2025.

B. A PRELIMINARY LOOK AT ACCOUNT-LEVEL AD REVENUE SHARING

In the absence of such relevant, comparable-across-platforms data, a "best-effort" approach was adopted to highlight the gap between the current publicly-available data offering and the Structural Indicators' ambitions. Our study focused on ad-revenue sharing or other platform payouts related to content popularity, as all platforms of interest offered such programs. Other monetization features, such as brand partnerships (disclosed or undisclosed), tipping or subscriptions were left out, but will be requested in future iterations.

In most cases, platforms impose two types of criteria to benefit from ad-revenue payouts: meeting some activity and audience thresholds (e.g. at least 5 videos posted in the last 90 days, garnering 100 000 views) as well as being in good standing with regards to the platform's community guidelines.

To isolate to the extent possible the latter factor, as this is where we would expect the effect (if any) of disinformation-relevant policies to materialize, we adopted a comparative approach, differentiating between high-credibility and low-credibility accounts. We started from the lists of accounts used in the cross-platform Structural Indicator (section 2.3.1), keeping only accounts that had posted at least one piece of content of a monetization-eligible format during the April-June 2025 period. We filtered out the accounts that did not meet the publicly-observable activity and audience criteria, leaving only accounts potentially eligible for monetization (the full platform-specific methodology is available in Appendix 5.3). We hypothesized that, under properly functioning demonetization systems, eligible high-credibility accounts would be monetized to a large extent, while low-credibility ones would not be monetized.

2.4.2 Results

TikTok, LinkedIn, X/Twitter and Instagram did not offer usable data and/or made it impossible to reasonably infer (even indirectly) a given account's monetization status and were consequently left out (see Appendix 5.3 for a discussion). Google Display Ads was included, as the service is covered by the Code (under Google Ads) and, unlike the other Structural Indicators, it could be audited using the same methodology as the other platforms.

For the services covered, each actor's likely monetization status was inferred, either by checking if it was present in the official list of monetization partners (in the case of Facebook) or by seeing how frequently ads appeared on their content (YouTube, Google Display Ads). Results were then aggregated for each credibility category (see Appendix 5.3 for full details).

	High-credibility assets			Low-credibility assets		
	Number of Assets	Assets that meet eligibility criteria	Monetized Assets (% of eligible)	Number of Assets	Assets that meet eligibility criteria	Monetized Assets (% of eligible)
Facebook	136	131	79 (**60.3%**)	141	70	14 (**20.0%**)
YouTube	110	107	84 (**78.5%**)	72	63	48 (**76.2%**)
Google Display Ads	81	81	57 (**70.4%**)	113	113	30 (**26.5%**)

Table 2.7 – Number and proportion of high- and low-credibility assets likely benefitting from ad-revenue sharing with different services. Assets refer to Pages (Facebook), channels (YouTube), or web domains (Google Display Ads).

Table 2.7 summarizes our results, showing how many channels or domains there are in our dataset, how many of them are eligible for monetization and the proportion of eligible accounts for which we have been able to confirm their monetized status. In absolute terms, we observe that none of the services is fully successful in ensuring that low-credibility accounts do not receive a share of ad revenue. However, the level to which this was the case varied widely across platforms: while only a minority of low-credibility actors were likely monetizing using Facebook and Google Display Ads (20% and 26.5%, respectively), a vast majority of low-credibility YouTube channels are, with more than three-quarter of eligible accounts monetized (76.2%).

Comparing monetization levels of high- and low-credibility actors confirms this observation: both groups are monetized at the same levels on YouTube (78.5% vs 76.2%), while a gap exists between them on Facebook and Google Display Ads (60.3% vs 20% on Facebook and 70.4% vs 26.5% on Google Display Ads).

3. Recommendations

FOR POLICYMAKERS AND REGULATORS

1) Make Structural Indicators part of routine supervision

Structural Indicators should be embedded in the ordinary supervisory cycle for VLOPs/VLOSEs. Regulators should require platforms to assist independent third-parties so they can audit a harmonised set of indicators on a regular schedule (e.g., twice a year).

2) Operationalise and enforce researcher access (DSA Art. 40.12)

Art. 40.12 access should move from ad-hoc, platform-specific negotiations to a predictable, enforceable regime, in line with researcher access to non-public data under Article 40.4. The Board and DSCs should publish a common EU data schema, a minimum technical standard, and service-level agreements for response times.

Access decisions should be logged, reasoned, and appealable within fixed deadlines that are compatible with research timelines. We fail to see a compelling need for well-established, reputable organizations studying topics evidently linked to DSA systemic risks to have to routinely wait months before receiving access to publicly-available data. Where a platform denies, unduly delays, or offers degraded/obsolete endpoints, researchers should have a clear, time-effective pathway to escalate 40.12 complaints to regulatory authorities.

3) Require random content samples

To estimate prevalence credibly, each platform should provide random samples of public content per Member State large enough to reach agreed precision (e.g., ±2-3 percentage points at 95% confidence). Each delivery should include a signed manifest describing: population covered, inclusion/exclusion rules (e.g., account types, surfaces), sampling method and seed, and any known coverage gaps.

4) Make samples auditable and reproducible

Regulators should require platforms to furnish reproducible sampling artefacts (e.g., seed values, hashing logic for selection), stable identifiers for content and accounts, and minimal metadata necessary to verify inclusion. Independent auditors designated by authorities should be able to re-draw the same sample ex post and to verify that no classes of content were silently excluded.

FOR PLATFORMS

To enable reproducible, cross-platform Structural Indicators while protecting users' privacy, platforms should implement two complementary access pathways.

1) Research APIs / bulk exports (ongoing access)

Platforms should maintain stable, well-documented endpoints (or periodic bulk files) exposing public content and public account data, partitioned by Member State and language. The minimum fields required are:

- Content-level exposure & interactions:
  Unique content ID; creator account ID; post timestamp; language; surface of exposure (feed, search, recommendations, ads adjacency); impressions; interactions (reactions/likes, comments, reshares/reposts, saves/bookmarks).
- Account-level metadata:
  Account ID; followers number; audience geography; participation in creator/partner programmes (eligibility, enrollment dates).

2) One-off "regulatory samples" (time-bounded audits)

Alongside ongoing APIs, platforms should deliver time-bounded datasets for specific audit windows (e.g., a defined month around an election). These samples must be platform-generated and accompanied by a signed sampling manifest (inclusion/exclusion criteria, randomisation method, seed…). They should include records for content later edited or removed within the window to avoid survivorship bias. Where payouts or ad adjacency apply, include content-level monetisation eligibility flags and account-level payout aggregates for the window.

Together, the ongoing APIs (for longitudinal research) and the regulatory samples (for verifiable point-in-time audits) provide the minimum infrastructure needed to compute Structural Indicators that are comparable across platforms and reproducible over time.

4. Can Large Language Models help measure Structural Indicators?

We piloted the use of large language models (LLMs) to assist with content filtering and pre-labelling, with two goals: test where automation can safely reduce manual workload, and quantify the reliability and cost of such automation. Human annotation remained the source of truth for all indicators in this wave.

4.1 POST-LEVEL ASSISTANCE FOR THE PREVALENCE INDICATOR

Filtering out Irrelevant content

For the Prevalence Indicator, we used GPT-4o-mini to help filter posts retrieved via keyword searches that fell outside the scope of our study (e.g., celebrity gossip, recipes, personal anecdotes, or uses of “COVID-19” purely as a time marker). The model also flagged Geographically Irrelevant items (e.g., Spanish or French content about Latin America or francophone Africa).

Because processing the full corpus (~3 million posts) would have been cost-prohibitive, we applied this automated filtering to a random sample of up to 20 000 posts per platform when available. Early performance was promising but imperfect; several prompt iterations and spot-checks by fact-checkers were needed to curb both over- and under-filtering. Further prompt refinement is warranted.

During the process, we encountered several challenges.

First, applying LLMs to video-first platforms introduced additional constraints. For YouTube and TikTok, effective text-based classification requires access to transcripts; acquiring them is costly, transcripts are not always available, and many TikTok videos contain only music with on-screen text, which is time-consuming and expensive to extract reliably.

Second, even on text-centric platforms, attached media (images/videos) can alter meaning or conceal misinformation. Because of storage and processing costs, we did not systematically download and parse attached media, which may have reduced the LLM’s ability to capture the full semantics of some posts.

Categorising content

After fact-checkers completed annotations on the view-weighted random sample (labeling posts as *Mis/disinformation*, *Credible*, *Borderline*, *Abusive*, *Unverifiable*, *Irrelevant*), we

tested whether LLMs could replicate these labels. The goal was to assess whether LLMs could accurately replicate the work done by fact-checkers and potentially scale this process for larger datasets.

To do so, we used the dataset annotated by the fact-checkers as a testing ground to evaluate the performance of various LLMs. We tested three models: Mistral Medium 3, Magistral Medium 1 and GPT-4o-mini, primarily on text-heavy platforms (Facebook, LinkedIn).

The categorization process involved several key steps:

1. **Relevancy Rating:** The LLMs first rated posts to determine whether they are Irrelevant. Posts deemed irrelevant were filtered out at this stage.
2. **Initial Labeling:** The LLMs analyzed relevant posts, comparing it to factual information in its knowledge base.
3. **Enrichment (optional):** If the LLM identified gaps in context or knowledge, it conducted a web search to gather additional information and enrich its understanding of the post's content.
4. **Final Labeling:** With the enriched context, the LLM re-evaluated the post, now assigning it one of the labels.

After several tests, the results showed room for improvement. The accuracy of the LLMs to properly identify posts containing mis/disinformation varied significantly by language, with the Mistral models offering the highest accuracy, but still only achieving between 50% and 70% accuracy, depending on the language.

Crucially, enrichment via web search, often required for correct interpretation, significantly increased cost. Given accuracy and cost, scaling this pipeline for core labelling was not deemed sustainable for this wave; we retained it as an R&D track.

4.2 ACCOUNT-LEVEL ASSISTANCE FOR THE SOURCES INDICATOR

For the Sources Indicator, LLMs supported fact-checkers in assessing the credibility of influential accounts (the whole process is described in Section 5.2.2). We collected posts by candidate accounts during 17 March–13 April 2025, including associated media where feasible. On YouTube and TikTok, we downloaded available video transcripts and fed the first ~250 words into the model; we applied the same approach to short-form video (e.g., Instagram reels) when transcripts existed. When images were attached to posts, they were included in the data examined by the model to capture additional cues.

Here, LLMs were helpful as triage tools: while they are of limited reliability for determining whether a single post is definitively mis/disinformation, they usefully surfaced a shortlist of

posts from a given account that likely warranted closer human review. Fact-checkers then made the final credibility determination more quickly.

4. Acknowledgements

We gratefully acknowledge the financial support of the European Media and Information Fund (EMIF), the main funder of this work*.

We also thank the Bright Initiative (powered by Bright Data) for providing access to selected services that facilitated data collection.

We are grateful to LinkedIn and TikTok for their cooperation and assistance, which helped enable portions of our analysis through access to data and technical interfaces.

Finally, we thank Jacopo Amidei, Ishari Amarasinghe and Andreas Kaltenbrunner, who are researchers at the Universitat Oberta de Catalunya, for their scientific review and constructive feedback on our methodology as well as EDMO for drafting detailed recommendations on how to measure Structural Indicators.

* The sole responsibility for any content supported by the European Media and Information Fund lies with the author(s) and it may not necessarily reflect the positions of the EMIF and the Fund Partners, the Calouste Gulbenkian Foundation and the European University Institute.

HOW TO CITE THIS REPORT

Vincent EM, Crisan D, Carniel B (2025) Measuring the State of Online Disinformation in Europe on Very Large Online Platforms. First report of the SIMODS project (Structural Indicators to Monitor Online Disinformation Scientifically).

5. Appendices

5.1 METHODOLOGY FOR THE PREVALENCE INDICATOR

5.1.1 Data collection

To collect data that reflects the diverse type of content users are exposed to on platforms, we selected approximately 100 keywords per language. These words were chosen for their relevance to the public conversation within the European space and their connection to topics that are often found in misleading claims or local issues in our target countries. Additionally, given the various spellings and declensions of certain keywords depending on the language, we included plural forms and grammatical variations to ensure broader coverage and capture more relevant posts, as well as accounting for compound words, using an exact match search when possible. We focused on five major topics: the Russo-Ukrainian conflict, climate change, general health (including Covid-19), migration, and local politics. The full list of keywords is available to scientists upon request for any legitimate research project.

The keyword lists were developed by the fact-checkers in our consortium, with the goal of striking a balance between topic-relevant terms across a spectrum of proximity to misleading claims. These include “neutral” keywords typically used in news reporting or general discussions (e.g., *Zelensky*, *migrants*, *Covid-19*), “ambiguous” terms associated with certain narratives (e.g., *vaccine side effects*, *geoengineering*, *laboratories in Ukraine*), and “misinformation-related” terms that are more commonly linked to misinformation (e.g., *Ukrainian Nazi*, *climate scam*, *remigration*).

The lists of keywords were tested by fact-checkers to ensure they yielded relevant and pertinent results. For each keyword in the local language, fact-checkers conducted searches on at least two platforms from the six VLOPs included in this study. The criteria for determining whether a keyword should be included in the analysis were as follows:

- The results should contain viral posts, defined as those with more than 50k views or over 1k interactions. These posts should be relatively recent (within the last six months) and appear on both platforms. If no viral and recent posts are found on at least one platform, the keyword was excluded.
- More than 50% of the content in the search results should be directly related to the topic being searched (e.g., climate change, health, Ukraine war, or immigration) and relevant (excluding entertainment or opinion posts). If most of the content was off-topic or irrelevant, the keyword was excluded.

- For keywords in the ambiguous or misinformation-related categories, fact-checkers were instructed to exclude the keyword if no viral *misinformation* posts were found within the first 20 results.

This process ensured that only keywords with a substantial presence of relevant and viral content were included for the analysis.

The first data collection period spanned 17 March to 13 April, and we collected posts that are published between these dates (inclusive). To collect data from the selected platforms, we employed two different methods. First, given that this project investigates the systemic risk of disinformation, as defined in the Digital Services Act (DSA), Article X, we invoked Article 12 of the DSA and contacted all six VLOPs on 19 December 2024 to request a random sample of 200 000 posts per language. Of the six platforms we contacted, only LinkedIn provided a random data sample. This dataset consisted of public posts from LinkedIn members and companies whose location is set to our countries of interest, with a maximum of 200 000 posts per language. The exact number varied depending on data availability in each respective country. TikTok granted us access to its researcher API on 31 March 2025; however, this access came too late in the data collection process to be used. Following our initial outreach, we sent follow-up reminders to Meta and YouTube but did not receive a response. As for X/Twitter, our application was denied on 9 January 2025, with the platform stating that the project does not meet the requirements under Article 34 of the Digital Services Act. We submitted an appeal on 17 January 2025, but as of September 2025, we had not received a reply.

The second method involved identifying alternative tools that enabled access to each platform's native search functionality and allowed us to perform keyword-based searches. Since access to platform data varies depending on the technical restrictions imposed by each platform, we adopted a tailored approach for data collection. Specific methods were selected and implemented based on the technical and policy constraints of each platform. A detailed breakdown of the data collection approach used for each platform is provided below.

A. META

Data collection for the META platforms (Facebook and Instagram) was carried out manually on a biweekly basis using the Meta Content Library. Each week, data which was posted in the timeframe Monday to Wednesday was collected on Thursday, and data for Thursday to Sunday was collected the following Monday. This was made possible by using the date filter available in the platform's user interface, which allowed us to target only those specific days and retrieve all of the content that was made available to us on that specific week.

For Facebook, Meta offers the possibility to download a subset of the public content dataset, which includes posts from *Pages* with 15 000 or more likes or followers, and from *Profiles* with a verified badge and at least 25 000 followers. We conducted a manual search using the boolean search function on the 100 keywords per language, targeting *Profiles*, *Pages*, and *Groups*. To narrow the scope, we applied the *Post Surface Country* and *Language* filters corresponding to each country of interest.

For Instagram, Meta also allows access to a subset of public content, including posts from *Business*, *Creator*, and *Personal* accounts with at least 25 000 followers or a verified badge. Just as for Facebook, we conducted a manual search using the boolean search function on approximately 100 keywords per language, targeting *Business*, *Creator*, and *Personal accounts*, and applied the *Language* filter for each respective country. Additionally, we used the image-text search feature to identify relevant content that included keywords within images. The resulting dataset included various types of content, such as posts, reels, and images, along with associated metadata. This metadata comprised elements such as the content description, number of likes, comments, interactions, and views for each post.

To ensure an accurate reflection of the content's reach and engagement, all posts were collected within a maximum of four days from their publication. This short time frame allowed us to capture content while it was still actively circulating. To account for the potential increase in user engagement over time, we revisited the same posts at the end of the data collection period to update their metadata, such as the interactions metrics and view counts.

B. X/TWITTER

Data collection for X/Twitter was conducted daily using Apify, a licensed third-party tool, searching for the language related keywords. Apify relies on a scraper that leverages X/Twitter's native search functionality, specifically through the 'searchTerms' field, allowing for keyword-based content retrieval. Similar to X/Twitter's search interface, Apify enables users to select the type of content to display, including Latest posts, Trending posts, Photos, or Videos. For the purposes of this study, we selected the "Latest" filter to capture the most recent posts published at the time of each search.

Apify also offers filters by date, which we used to retrieve content posted on the exact day of collection, as well as language filters to target content specific to the countries of interest. It is important to note that X/Twitter does not provide a dedicated filter for the Slovak language. As an alternative, we used the Czech language filter, based on guidance from Demagog SK's fact-checking team. This decision was informed by their confirmation that the Czech language filter returns content in Slovak and that Slovak audiences

frequently consume Czech-language content, given the linguistic similarities between the two. To ensure the relevance of the dataset for the Slovak context, fact-checkers were instructed to label content that was not relevant to the Slovak population as *Irrelevant*, following the guidelines outlined in Appendix 5.1.3.

To account for the potential increase in user engagement over time, we revisited the same posts at the end of the data collection period to update their metadata, such as the interactions metrics and view counts.

C. YOUTUBE

Data collection for YouTube was conducted on a daily basis using Check First's monitoring system called CrossOver. This tool simulates user behaviour on the platform by replicating native search functionality, returning results as they would appear to an actual user based on their actual geographical location. Each day and for each language, search queries were conducted, to capture a wide range of relevant content.

Due to platform search limitations and resource constraints, no date filters were applied during the initial data collection phase. Instead, we filtered content by publication date during post-processing to ensure that only posts from 2025 were included in the analysis. While content from other platforms was restricted to posts published between March 13 and April 17, we chose to include all YouTube videos posted at any point in 2025 (the same applies to TikTok as detailed below). This decision reflects the longer content lifespan and engagement cycle of YouTube videos compared to other platforms.

In addition to the primary search results, the system also captured the recommended videos associated with each result. This approach allowed us to collect not only direct search results, but also the broader content ecosystem that users are exposed to when interacting with YouTube on our topics of interest.

D. TIKTOK

Data collection for TikTok was conducted daily using Check First's monitoring system, CrossOver. This tool replicates TikTok's native search functionality, simulating real user behaviour and returning results as they would appear to an actual user based on their geographical location.

For each language, search queries were performed each day to capture a broad range of relevant content. Since TikTok does not provide a native date filter, all videos retrieved through keyword searches were later filtered during the post-processing phase, and only content published in 2025 was included in the final dataset, matching our approach with YouTube.

To enhance data robustness and ensure broader coverage, we complemented this approach with a data collection using Bright Data, a third-party technology provider offering web data collection and proxy services. Bright Data was used to collect additional data based on the same predefined keyword searches. We merged both datasets collected from CrossOver and Bright Data to constitute our final dataset.

5.1.2 Data Processing & Sampling

A. DATA CLEANING & PROCESSING

The resulting dataset from the data collection period, spanning 17 March to 13 April, comprised a total of 2.6M posts across all platforms and target languages, after duplicates were removed. For TikTok and YouTube, only posts published in 2025 were retained, while for the remaining platforms, we included exclusively the content published within the defined data collection period.

Additionally, content not published in one of the targeted languages, i.e., Spanish, French, Slovak/Czech, English, or Polish was excluded.

As shown in Table 5.1, TikTok leads in total views, accumulating approximately 8.3 billion, whereas X/Twitter records around 500 million, highlighting a substantial disparity between platforms. Across the full dataset of 2.6 million posts, spanning six platforms and four languages, the combined total reached 24 billion views, reflecting the overall reach of the content analyzed.

Platform	Views (entire dataset collected)	Posts (entire dataset collected)
Facebook	2 842 746 331	221 011
Instagram	4 075 375 809	77 629
YouTube	2 165 804 082	15 007
X/Twitter	470 224 104	1 030 272
TikTok	8 2313 62 104	15 270
LinkedIn	6 391 619 710	1 227 569

Table 5.1 – Total number of posts and views within the keyword search dataset

Given the nature of the selected keywords and their potential use in varied contexts, some retrieved posts were unrelated to our topics of interest. For example, terms like “Covid-19”

were sometimes used as temporal markers rather than referring to the pandemic or the disease itself, leading to the inclusion of irrelevant content.

To remove such content, we employed a Large Language Model (LLM), specifically a GPT 4o-mini, to filter out contextually irrelevant posts. Multiple versions of the classification prompt were tested on a random sample of a total of 250 posts, with the results validated by the fact-checkers, until we identified a prompt that satisfactorily allowed us to discriminate relevant from irrelevant contents.

Due to the size of the dataset and the computational costs associated with LLM processing, we selected a random sample of 20 000 posts per platform and per language on which to apply the filtering. For text-based platforms (X/Twitter, Facebook, Instagram, LinkedIn), the LLM was applied to the post descriptions. For video-based platforms (YouTube and TikTok), we downloaded the video transcripts and used these as input for the filtering process.

The LLM classified content into three categories:

- Relevant: content contributing to public discourse on the state of the world, such as health, science, politics, climate change, or other societal issues with a direct impact on people's lives or understanding of society
- Irrelevant: content on topics unrelated to our study or that do not match the definition of Relevant above; typically including celebrity gossip, sports, cooking recipes without health claims, beauty routines, and personal religious opinions.
- Geographically Irrelevant Content:
    - posts that fall outside the geographical scope of the analysis such as posts in French discussing African politics, for instance, or posts in Spanish addressing political developments in South America.
    - content not written in one of the targeted languages: French, Spanish, Slovak/Czech, Polish, or English

Content labeled as *Irrelevant* or *Geographically Irrelevant* were removed from the dataset.

B. RANDOM SAMPLING

To obtain a reliable proxy of the state of online discussions on high-sensitivity topics across platforms, we drew a random sample of 500 posts per platform and language, weighted by the number of views of each post.

Weighting by views was a critical step for three main reasons. First, it allows us to ensure that widely viewed posts are more likely to appear in the sample, thus reflecting what users are actually seeing on the platform. As shown in Table 5.2, the resulting weighted sample accounts for a total of 3.8 billion views, with 1.9 billion views coming from TikTok

alone. Second, it helps mitigate potential biases introduced by platform-specific search algorithms, which may be influenced by personalization or ranking mechanisms. Third, it captures variations in topic salience, meaning that if, for example, posts about the war in Ukraine receive significantly more engagement than those on climate change, they will be proportionally more represented in our annotated sample.

Platform	Total number of views in the random sample
Facebook	325 392 155
Instagram	592 692 176
YouTube	684 116 016
X/Twitter	14 024 616
TikTok	1 945 557 092
LinkedIn	207 183 002

Table 5.2 – Total number of views per platform in the random sample (2 000 posts per platform)

5.1.3 Annotation

To ensure high-quality annotated data, each sample of 500 posts per platform and per language was individually reviewed by fact-checkers with relevant language and topic expertise. These reviewers assessed whether the content contained misinformation, using a classification framework developed collaboratively by the fact-checking team. The framework was designed to reflect the wide variety of content typically encountered on social media and to enable consistent application across platforms and linguistic contexts.

To ensure the robustness of our findings, assurance, a cross-verification process was implemented. For each platform and language, a first fact-checker was responsible for annotating the full sample of 500 posts, while a second fact-checker independently reviewed a randomly selected 20% subset of the same sample. Both fact-checkers worked blindly and independently, without access to each other's annotations, ensuring an unbiased second layer of review.

In cases where discrepancies arose between the two reviewers, a resolution phase followed the initial annotation. After the full dataset had been labeled, the fact-checkers reviewed the cases with conflicting classifications, discussed their assessments, and agreed on a final label for each disputed item. This process not only ensured consistency and accuracy in content classification but also enabled us to measure inter-annotator agreement, an important indicator of reliability, and to incorporate this information into the calculation of confidence intervals for the final prevalence estimates, which is explained in detail in Annex 5.1.4.

Fact-checkers were tasked to label each piece of content with one of the options below:

- **Mis/disinformation**: Content stating or clearly implying a verifiably false or misleading claim that may cause public harm.
  This definition is a simplified version of the one from the Code of Conduct.
- **Credible and informative**: Content conveying true or credible information on important matters about the state of the world (excluding trivia, gossip, or anecdotes).
  The *Credible* label was only applied to content that presents factual information on topics with direct relevance to people's lives or public understanding of society, such as health, science, politics, or social issues. These posts had to be accurate and informative, i.e. contribute constructively to public discourse.
- **Borderline**: Content feeding a misleading narrative without necessarily containing outright falsehoods, but potentially reinforcing false beliefs.
  This category captures content that does not meet the criteria for being labeled as mis/disinformation but does nonetheless contribute to the spread or normalization of misleading narratives. Research has shown that "factually accurate but deceptive content" about vaccines, for instance, can be "more consequential for driving vaccine hesitancy than flagged misinformation" as it is more prevalent than strictly false or misleading information[15]. This is the phenomenon we are intending to capture with the *Borderline* category.
- **Abusive**: Content not containing mis/disinformation but involving harmful material such as hate speech, insults, spam, or incitement to harmful behaviour.
  Hateful or discriminatory content was only classified as mis/disinformation if it also included false or misleading claims. Content that included hate speech or offensive language alone, without a misinformation component, was labeled under the *Abusive* category.
- **Unverifiable**: Content that cannot be assessed as either credible or mis/disinformation (e.g. opinion-based).
  For contents that address important societal topics but cannot be classified as either credible or mis/disinformation, typically because they involve personal or political opinions, or subjective commentary that fall outside the scope of factual verification, we used the *Unverifiable* category.
- **Irrelevant**: Content not about public affairs or scientific/political issues (e.g. entertainment, sports, religious content, cooking recipes without health claims, geographically irrelevant to Europe).
  Contents unrelated to public-interest information or verifiable factual claims were labeled *Irrelevant*. This includes song lyrics, sports updates, cooking recipes without health claims, celebrity gossip, expressions of religious belief without factual

assertions, and purely personal anecdotes. Although we used an LLM to filter content outside the study's geographical scope (Appendix 5.1.2) , this automated step was not foolproof; residual off-scope items could remain. To address this, fact-checkers were instructed to flag any posts as Irrelevant when centered on events in Latin America or Francophone Africa for instance, so as to maintain our focus on Europe.

- **Other language**: Content is not in one of the languages spoken in the targeted country or English.

Posts written in languages other than the targeted ones, French, Spanish, Slovak, Polish, or English, were labeled as *Other Language*.

- **Deleted**: Content unavailable at the time of annotation (e.g. removed from the platform).

Contents that had been deleted from platforms at the time of review were labeled as *Deleted*.

- **Don't know**: Content not fitting any other category.

Contents that could not be reliably classified under any of the defined categories due to ambiguity, lack of context, or incomplete information, were assigned the label *Don't Know*. This ensured that all reviewed posts were accounted for, even when a definitive classification was not possible.

5.1.4 Inter-annotator Agreement and related Confidence Intervals

Each content in the random sample was first annotated by one fact-checker, who assigned it to one of the categories described above. A second fact-checker independently reviewed a randomly selected 20% subset of the sample.

Figure 5.1 displays the number of cases when both fact-checkers agreed or disagreed on the categorization of content across the five categories we study (*Misinformation*, *Credible*, *Borderline*, *Abusive*, and *Unverifiable*) across all platforms and languages. The rows represent the classifications made by the second fact-checker, while the columns represent the classifications by the first fact-checker. The diagonal of the matrix shows the cases where both fact-checkers agreed on the classification; we observe a high level of agreement between the fact-checkers given that the numbers on the diagonal are always higher than the numbers outside of it on any given row or column. The sum of the numbers on the diagonal shows that the agreement rate is 87.5%. A source of disagreement occurred between the categories *Credible* and *Unverifiable*, for instance; as we can see on the confusion matrix, the first fact-checker rated content as *Credible* 36 times while the second rated it as *Unverifiable* and the second fact-checker rated content as *Credible* 34 times while the first rated it as *Unverifiable*.

Figure 5.1 – Confusion Matrix Showing Inter-Annotator Agreement and Disagreement in Independent Content Labeling

Once the data sample was fully labeled, the two fact-checkers discussed cases where discrepancies between the labels occurred and agreed on a “final label” for each piece of content. When a final label was available for a given content, this label was used in the prevalence calculation. However, when only the first fact-checker label was available, we took into account the probability that the label proposed by the first fact-checker could be wrong as described below.

To quantify the uncertainty around our estimates of prevalence, we applied a bootstrapping technique with 1 000 iterations. Bootstrapping is a resampling method that involves repeatedly drawing samples with replacement from the original dataset and recalculating the prevalence metric in each iteration. This process generates an empirical distribution of the prevalence estimate, from which confidence intervals can be derived without relying on parametric assumptions. The bootstrapping thus allows us to quantify the confidence intervals around our estimate related to the size of our sample.

In addition to measuring the uncertainty related to sample size, we also accounted for uncertainty arising from potential disagreements between annotators. Specifically, we incorporated the frequencies with which the initial labels assigned by the first fact-checker differed from the final label (when available), which we take as our best estimate of ground truth.

To give a concrete example, in Slovakia the first fact-checker labeled 66 pieces of content as *Abusive*. The final label agreed in 57 cases (86%), 4 were relabeled *Unverifiable* (6%), 4 *Borderline* (6%), and 1 *Mis/disinformation* (1.5%). In the bootstrapping process, at each iteration, we therefore randomly swapped the *Abusive* label with *Borderline* with probability 6%, with *Unverifiable* with probability 6%, with *Mis/disinformation* with probability 1.5% and left it unchanged with probability 86% (percentages may not sum to exactly 100% in this example due to rounding). Note that this procedure was applied separately for each country, to reflect potential team-specific biases.

5.1.5 Results

A. PREVALENCE ACROSS CATEGORIES

As outlined in Section 2.1.2-A, the content breakdown across all six platforms reveals that the combined categories of *Credible* and *Unverifiable* account for the majority of content on all platforms. While the *Unverifiable* category provides valuable context for the type of content commonly found on social media, as it typically encompasses personal opinions, commentaries, and individual perspectives on global events and news, we sought to highlight the distribution of *Credible* versus *Problematic* contents in isolation. As defined in Section 2.1.2-A, *Problematic* content refers to the combination of *Abusive*, *Borderline*, and *Mis/disinformation*.

Figure 5.2 reproduces Figure 2.2 excluding the *Unverifiable* category. This allows to highlight that LinkedIn has the lowest proportion of problematic content, followed by Instagram and then YouTube and Facebook. The highest levels of *Problematic* content as compared to *Credible* content is found on TikTok and X/Twitter, where it represents a bit less than half (47%) and a bit more than half (53%) respectively.

B. PREVALENCE OF MIS/DISINFORMATION + BORDERLINE

As we have explained above, content that is factually accurate can still lead users to misleading conclusions, which we captured in the *Borderline* category. If one wants to measure the proportion of all potentially misleading content, one needs to assess the prevalence of Mis/disinformation and Borderline content together.

Figure 5.3 displays a measure of the proportion of posts labeled as *Misinformation* or *Borderline*, relative to the total number of posts labeled as *Misinformation*, *Borderline*, *Credible*, or *Unverifiable*. In line with the patterns observed in the prevalence of Mis/disinformation content, TikTok has the highest combined prevalence of *Misinformation* and *Borderline* content, with 32% of posts falling into these categories. It is followed by

Facebook at 25% and X/Twitter at 24%. LinkedIn displays the lowest prevalence of such content at 7%, indicating a significantly more limited presence of misleading content on this platform.

Figure 5.2 – Percentage of posts belonging to each category for each of the six very large online platforms, same as Figure 2.2 but excluding *Unverifiable*.

Figure 5.3 – Prevalence of *Mis/disinformation + Borderline* content across the six very large platforms, aggregated across all languages. The error bars represent the 95% confidence intervals measuring the uncertainty around each estimate, calculated using a bootstrapping method (See Appendix 5.1.4 for details).

C. MISINFORMATION PREVALENCE BY COUNTRY

Figure 5.4 illustrates the prevalence of misinformation across the four countries in our study. Results differ by country: in Poland and Spain, prevalence is not statistically different across platforms (approximately 10% and 5%, respectively) on all platforms except LinkedIn. By contrast, in France TikTok shows the highest prevalence (~40%), followed by X/Twitter and Facebook (~20%). In Slovakia, TikTok and Facebook exhibit the highest prevalence (~18%), followed by YouTube (~10%).

Figure 5.4 – Misinformation prevalence for each country across the six very large platforms per language. The error bars represent the 95% confidence intervals measuring the uncertainty around each estimate, calculated using a bootstrapping method (See Appendix 5.1.4 for details).

5.2 METHODOLOGY FOR THE SOURCES INDICATOR

5.2.1 Data collection

The process of building a dataset of accounts that repeatedly share mis/disinformation involved several stages and drew on multiple sources, which allowed us to test the sensitivity of our results to the chosen methodology.

A. THE FACT-CHECKERS' LIST APPROACH

In one approach, fact-checkers from the consortium compiled preliminary lists of trustworthy sources and social media channels known for sharing mis/disinformation for each country. These lists were developed based on the fact-checkers' expertise, internal databases of accounts whose posts have been fact-checked, and Science Feedback's Consensus Credibility Scores, which aggregate multiple open-source credibility ratings for over 20 000 sources, providing a reliable basis for identifying recurrent misinformation sources[13].

B. THE TOP 50 LIST APPROACH

In another approach, we leveraged the keyword-based dataset used in Indicator 2.1 (Prevalence) to identify the most influential accounts in our dataset. For each platform and language, we selected the top 200 accounts based on the cumulative number of views their posts received during the data collection period (17 March to 13 April 2025). These accounts were manually reviewed by fact-checkers to determine their relevance to the study. Accounts were considered relevant if they regularly discussed topics of interest for our study, disseminated news, or shared content related to public affairs and misinformation. Conversely, accounts focused exclusively on entertainment, sports, or celebrity news were excluded. From the pool of relevant accounts, we retained the top 50 per platform-language pair.

For both of the lists from above, the top 50 and fact checkers recommendation, we collected all of their posts during the data collection window using third-party tools such as BrightData and Apify, capturing not only post content but also associated metadata (engagement metrics, number of followers, images, and videos). Due to platform limitations, view counts per post were only available for YouTube, TikTok, and X/Twitter. For LinkedIn, where fewer accounts were active during the target period, we extended the data collection window by three weeks before and after the data collection period i.e., 24 February to 4 May.

It should be noted that, during dataset construction, we were unable to retrieve every post from every account, leading to missing entries in both the Top 50 and the fact-checkers' recommended lists. These gaps stem from platform and tool constraints, as well as accounts that did not post during the collection period or had deleted content. Consequently, we cannot guarantee that the retrieved posts represent the complete set of posts published by these accounts during the period. While this is not necessarily problematic for metrics such as interactions per post, it prevents us from drawing definitive conclusions about the total interactions or views generated by all posts from the accounts

in our dataset.

5.2.2 Annotation of Sources

Fact checkers labeled the accounts in the Top 50 list into three different categories, following these guidelines:

- **Low-credibility:** Accounts that shared at least two posts containing false or misleading information.
- **High-credibility:** Accounts that almost exclusively shared credible and informative news, such as content from professional media outlets or scientific institutions.
- **Neither:** Accounts that did not fit into the two categories above, often sharing opinion-based.

High-credibility sources typically include reputable news organisations and digital-native publishers known for producing accurate and informative content. These actors are characterised by adherence to organisational editorial standards and by operating under legal and regulatory frameworks that hold them accountable for the reliability of the information they disseminate.

To assist the fact-checkers in this task, we used a Large Language Model (GPT 4o-mini) to give a first estimation of whether each post shared by one of the top 50 accounts was likely to contain mis/disinformation.

We developed a prompt that asked the LLM to rate each post for its likelihood of containing mis/disinformation on a scale ranging from 0 to 10. Based on our testing conducted on a sample of labeled posts, we noted that posts scoring:

- between 0 and 3, typically corresponded to posts labeled as Credible by fact-checkers,
- between 7 and 10, typically corresponded to posts labeled as Mis/disinformation by fact-checkers,
- between 4 and 6, tended to correspond to posts labeled as neither Credible nor Mis/disinformation.

We thus used these thresholds to label each post as either *Credible*, *Misinformation* or *Neither*. To enhance the reliability of the LLM's classification, we validated the prompt on a random sample of posts annotated by a fact-checkers, calculating the agreement rate to assess model performance. We initially used GPT 4o-mini to label all posts, taking into account not only the textual description but also any associated media, such as images and videos. For posts that received a score equal to or above 7, indicating a higher likelihood of

misinformation, a second round of evaluation was performed using GPT 4o-mini with web-browsing capabilities. This additional step was crucial for weeding out false positives, typically associated with recent claims being wrongly labeled as misinformation by the LLM's out-of-date information. Posts initially labeled as misinformation with GPT 4o-mini were re-evaluated using the same prompt in GPT with browsing capabilities. If the post's score dropped below 7 during this second round, its classification was revised accordingly.

Once each post had been individually labeled as either *Misinformation*, *Credible* or *Neither*, we proceeded to aggregate these labels at the account level. This step allowed us to propose a first, LLM-based, categorisation of accounts based on the overall nature of the content they shared. An account was classified as:

- Low-credibility according to the LLM if it had two or more posts with scores above 7.
- High-credibility according to the LLM if at least 95% of its posts had scores below 3.
- Neither according to the LLM if none of the above conditions were met.

These suggestions by the LLM were then made available to fact-checkers who provided the final classification. These pre-classifications by the LLM helped save time for fact-checkers, as it allowed them to quickly identify the posts from each account that were more likely to contain mis/disinformation.

For readers interested in the accuracy of the LLM, Figure 5.5 presents a confusion matrix comparing its account-level pre-classifications to the final classifications made by the fact-checkers. The matrix shows that the LLM correctly identified 213 low-credibility accounts but misclassified 64 of them, for instance. Overall, GPT reached a precision of about 66% and recall of 75% for high-credibility accounts, and a precision of 62% and recall of 77% for low-credibility accounts.

5.2.3 Sensitivity of results to the two lists

To ensure our indicators on Sources weren't biased by the methodology used to constitute the lists of high-credibility and low-credibility accounts, we tested the sensitivity of the results by comparing the results obtained with the fact-checkers' list approach and with the Top 50 list approach. The main metric we proposed as an indicator of whether platforms welcome repeat sources of mis/disinformation is the number of interactions per post per follower that low-credibility accounts get as compared to the same metric for high-credibility accounts.

Figure 5.5 – Confusion Matrix Showing Agreement and Disagreement between GPT pre-classifications and fact-checkers final classifications of accounts in the Top 50 as High-credibility, Low-credibility or Neither.

Figure 5.6 shows that, regardless of the specific approach used to construct the lists of low-credibility and high-credibility accounts, the results are broadly consistent. Low-credibility accounts systematically outperform high-credibility accounts in terms of interactions per post per 1 000 followers, and they do so by similar multiplying factors in both datasets. Based on these results, we decided to merge the high- and low-credibility lists originating from the fact-checkers and Top 50 approaches. Note that for LinkedIn, fact-checkers weren't able to identify lists of high- or low-credibility based on their fact-checking activities, so only the dataset of the top 50 most influential accounts was used.

bootstrapping method (see Appendix 5.1.4 for details).

5.2.4 View-based 'Misinformation Premium'

In Section 2.2.2-B, we show that posts from low-credibility accounts consistently receive higher engagement than posts from high-credibility accounts on all platforms except LinkedIn. Here we test whether this observation still holds when considering the number of views, instead of the number of interactions. Due to the platforms' inherent features and capabilities of the third-party tools we used, we were able to collect the number of views for all the posts published by accounts in our lists of high- and low-credibility accounts only on TikTok, YouTube and X/Twitter.

Figure 5.7 shows that low-credibility accounts receive more views per post per 1 000 followers than high-credibility accounts. On YouTube, low-credibility accounts receive about three times as many views per post per 1 000 followers as high-credibility ones, while on TikTok and X/Twitter low-credibility accounts receive about twice as many views as high-credibility accounts.

This result is consistent with the interactions-based metric discussed in section 2.2.2-B, although the magnitude of the difference between high- and low-credibility accounts is greater when comparing the numbers of interactions: the interactions-based misinformation

Figure 5.7 – Average number of views per post per 1 000 followers for accounts classified as High-credibility (green), Low-credibility (red), and Neither (grey) on the three platforms where the number of views were available. Error bars represent 95% confidence intervals, calculated using a bootstrapping method (see Appendix 5.1.4 for details).

5.2.5 Results for 'Political' Accounts

Section 2.2.2 presented results that excluded accounts labeled as 'political' (accounts of politicians or political parties). Given that the speech of politicians is usually treated differently in public discourse, notably by journalists, we did not want our results to be potentially driven by the level of engagement they receive, which we expected could be higher than that of other accounts. However, in the labeling phase, posts from political accounts were treated in the same way as posts from other accounts, so we were able to label political accounts as either low-credibility if they shared two or more posts containing mis/disinformation, or Neither if they didn't. No political accounts were labeled as high-credibility given the typically partisan nature of the content they share.

Figure 5.8 illustrates the mean number of interactions per post per 1 000 followers, similar to Figure 2.6 but also including the two categories of political accounts. The figure shows that, across most platforms, low-credibility accounts that are also political receive more interactions per post per 1 000 followers than both high-credibility and low-credibility accounts. The exceptions are LinkedIn, where low-credibility political accounts receive significantly less interactions than other accounts, and TikTok, where our sample contains few posts from low-credibility political accounts making the confidence interval overlap the values of low-credibility non-political accounts.

5.2.6 Robustness test of the 'Misinformation Premium'

Our findings for the Sources Indicator show that low-credibility accounts outperform high-credibility accounts in interactions per post and views per 1 000 followers. To ensure this result is not driven by the typically smaller follower counts of low-credibility accounts, we replicated the analysis by stratifying on account followership.

For each platform, we divided low-credibility accounts into four groups (quartiles) based on follower count. Using the same follower-count ranges, we then grouped high-credibility accounts into four corresponding groups. For example, on Instagram the quartiles were: Q1: 0-33 000 followers; Q2: 33 000-70 000; Q3: 70 000-140 000; and Q4: 140 000-3 500 000, applied to both low- and high-credibility accounts; any high-credibility accounts above 3.5 million followers were excluded from the comparison. LinkedIn was the exception: follower counts lacked sufficient dispersion, so only two groups were created.

Figure 5.8 - Average number of views per post per 1 000 followers for accounts classified as High-credibility, Low-credibility, and Neither, Low-credibility Political and Neither Political on all six platforms where Political accounts were present. Error bars represent 95% confidence intervals, calculated using a bootstrapping method.

Platform	High-credibility	Low-credibility	Neither	Low-credibility Political	Neither Political
Instagram	**9.0** [8.7 - 9.4]	**45.4** [34.3 - 60.0]	**30** [24.3 - 38.1]	**82.7** [73.1 - 90.2]	**30.5** [25.3 - 37.0]
Facebook	**0.76** [0.71 - 0.82]	**5.6** [5.2 - 6.0]	**1.9** [1.8 - 2.2]	**26.1** [23.6 - 28.9]	**37.7** [33.7 - 41.9]
X/Twitter	**2.6** [2.4 - 2.8]	**9.9** [9.2 - 10.7]	**7.2** [6.8 - 7.6]	**17.3** [15.2 - 19.4]	**12.2** [11.3 -13.1]
YouTube	**0.67** [0.62 - 0.70]	**6.6** [6.2 - 7.1]	**4.2** [3.7 - 4.7]	**35.9** [33.1 - 38.5]	-
TikTok	**19.5** [16.7 - 22.2]	**39.2** [31.8 - 47.8]	**36.4** [23.0 - 55.0]	**70.8** [32.4 - 113.2]	-
LinkedIn	**29.2** [22.9 - 36.2]	**25.5** [5.81 - 58.1]	**34.2** [26.4 - 44.1]	**4.8** [3.3 - 6.5]	**22.2** [10.8 - 37.1]

Table 5.3 - Average number of interactions per post per 1 000 followers for accounts in the High-credibility, Low-credibility, Neither, Low-credibility Political and Neither Political lists on each platform, as displayed in Figure 5.8. Values in square brackets correspond to 95% confidence intervals.

This approach allowed us to compare high- and low-credibility accounts at similar audience sizes, minimizing bias that could arise when comparing all accounts jointly, given their different follower-count distributions. Within each quartile, we calculated the mean interactions per post per 1 000 followers, as in Section 2.2.2.B.

Table 5.4 presents the results, confirming those in Figure 2.7: across nearly all platforms and quartiles, the mean interactions per post per 1 000 followers is significantly higher for low-credibility accounts than for high-credibility accounts.

There are two exceptions:
(i) LinkedIn, where high-credibility accounts do not differ statistically from low-credibility accounts in interactions, consistent with Figure 2.7; and
(ii) TikTok, where high-credibility accounts surpass low-credibility accounts in the 4th quartile; both are statistically tied in the 1st and 2nd quartiles; and low-credibility accounts exceed high-credibility accounts only in the 3rd quartile (120 000-730 000 followers). These contrasts explain why, overall, the misinformation premium on TikTok is "only" ~2×.

5.3 METHODOLOGY FOR THE MONETIZATION INDICATOR

5.3.1 Facebook

As part of its advertiser brand safety offerings, Facebook publishes "partner-publisher lists", which "*show publishers that have signed up for monetization and follow our Partner Monetization Policies*". Those lists refer specifically to accounts whose video content can be used for monetization (ads playing during videos or Reels).

As they do not capture all types of ads (e.g. ads on users' feeds), nor all types of monetization (subscription, Meta Stars, Facebook Content Monetization program, branded content), these lists are not exhaustive and can only be considered indicative of the broader phenomenon of the monetization of disinformation on Facebook.

Starting with the Facebook accounts collected in Section 2.3 for which a fact-checker had assigned a credibility label, the following criteria were applied:

- Only Pages were kept, filtering out personal profiles as most are not eligible for monetization,

- Because of the partner-publisher lists' focus on videos, Pages that had not recently published videos or Reels, or that had reached minimal audiences on such content, were marked as ineligible[1].

Table 5.5 shows the results for Facebook, same as Table 2.7 but stratified by country.

Platform		Q1	Q2	Q3	Q4
Instagram	**High-credibility:**	**66** [58.9, 73.5]			
	Low-credibility:	**117** [79.8, 163.8]			
Facebook	**High-credibility:**	No data			
	Low-credibility:	**18** [16.2, 20.4]			
X/Twitter	**High-credibility:**	**6.5** [5.1, 8.1]			
	Low-credibility:	**23** [21.0, 26.0]			
YouTube	**High-credibility:**	**2.2** [2.1, 2.4]			
	Low-credibility:	**16.7** [15.6, 18.0]			
TikTok	**High-credibility:**	**70** [41.7, 105.0]			
	Low-credibility:	**62** [41.6, 87.2]			
LinkedIn	**High-credibility**	**60** [53.0, 66.0]			
	Low-credibility	**44** [25.3, 65.0]			

Table 5.4 - Stratification analysis of average interactions per post per 1 000 followers for high-credibility and low-credibility accounts across platforms. Accounts are divided into quartiles based on follower count. Values in square brackets indicating 95% confidence intervals.

In July 2025, Facebook was in the process of rolling out its Facebook Content Monetization Program, which aims to combine many of the platform's monetization features. Starting in September 2025, this should result in the end of the specific in-stream Ads and ads on Reels programs, which formed the basis of the partner-publisher lists.

Given the absence of publicly-available data on accounts already onboarded onto the Facebook Content Monetization Programme, it is doubtful whether the transition will result

[1] Specifically, a Page must have had a minimum of 5 videos or Reels published in the last 30 days AND at least 60 000 minutes watched on videos from the last 60 days (calculated as number of views times video duration, divided by 2 to account for mid-play drops). As Facebook does not publish official activity and audience criteria anymore, these thresholds were derived from earlier guidance and might have changed since.

	10.7 [9.9, 11.6]	**9.4** [8.9, 10.0]	**6.7** [6.5, 6.8]
	34.6 [30.8, 39.0]	**17** [15.4, 19.1]	**27** [24.0, 31.3]
	1.5 [1.4, 1.6]	**0.9** [0.79, 1.06]	**0.26** [0.24, 0.28]
	8.5 [7.7, 9.4]	**6.4** [5.9, 6.9]	**1.8** [1.6, 2.0]
	6.0 [5.4, 6.6]	**1.1** [0.96, 1.3]	**1.2** [1.1, 1.3]
	11.6 [10.7, 12.5]	**4.6** [4.3, 4.9]	**2.4** [2.2, 2.5]
	0.5 [0.48, 0.58]	**0.4** [0.37, 0.41]	**0.16** [0.15, 0.17]
	4.2 [4.0, 4.5]	**0.4** [0.39, 0.46]	**0.9** [0.8, 1.0]
	55 [30.4, 86.1]	**21** [18.2, 24.9]	**14** [12.7, 15.1]
	25 [15.1, 40.2]	**62** [47.6, 79.3]	**7** [5.5, 9.2]
	19 [17.4, 19.8]	No data	No data
	8 [6.0, 11.6]	No data	No data

in sufficient data to conduct further analysis.

	High-credibility accounts			Low-credibility accounts		
	Nr. Pages or Accounts	Pages that meet eligibility criteria	Monetized Pages (% of eligible)	Nr. Facebook Pages or Accounts	Pages that meet eligibility criteria	Monetized Pages (% of eligible)
Slovakia	25	22	1 (**4.5**%)	35	13	0 (**0.0**%)
Poland	31	30	21 (**70.0**%)	56	37	10 (**27.0**%)
France	40	39	33 (**84.6**%)	37	16	4 (**25.0**%)
Spain	40	40	24 (**60**%)	13	4	0 (**0.0**%)

Table 5.5 – Number and proportion of high- and low-credibility accounts appearing in Facebook's partner-publisher lists, indicating likely monetization.

5.3.2 Instagram

Instagram does not offer meaningful data to track account-level monetization. Transparency has taken a step back as Instagram stopped publishing its "partner-publisher lists" that detailed the accounts that were eligible for monetization in H1 2025.

5.3.3 X/Twitter

X/Twitter does not offer publicly-available data as to the accounts its flagship revenue sharing mechanism (the "Creator Revenue Program") supports. Consequently, we were not able to study monetization on X/Twitter. Future iterations will have to rely on access to platform data under DSA Article 40.4.

5.3.4 LinkedIn

In 2024, LinkedIn launched its first revenue-sharing program (Wire, rebranded as BrandLink in May 2025). No consolidated data is publicly available as to which creators are taking part in the pilot program, although the few names cited in communications material belong to broadly credible actors, such as Der Spiegel or the Washington Post, alongside individual influencers. Consequently, we were not able to study monetization on LinkedIn. Future iterations will have to rely on access to platform data under DSA Article 40.4.

5.3.5 TikTok

TikTok offers two flagship programs to reward creators for the views their content garners:

- The TikTok Creator Rewards Programme, which, broadly, pays out users on the basis of how well their videos perform. The list of accounts eligible to partake in the program is not public.
- TikTok Pulse, which shares with creators the revenue from ads appearing next to the most trending videos (videos with a "Pulse Score" in the top 4 percent of all videos on TikTok - the Pulse Score being an internal metric blending "user engagement, video views and recent growth"). With no data on which videos are in Pulse Score top 4% nor on which accounts are eligible, a systematic study of the TikTok Pulse funding was not possible.

Consequently, we were not able to study monetization on TikTok. Future iterations will have to rely on access to platform data under DSA Article 40.4.

5.3.6 YouTube

Similar to Facebook, we screened YouTube channels of high- and low-credibility actors (see section 2.3) to see which were eligible for monetization on the basis of publicly-observable criteria (i.e., criteria not related to the channel's or content's standing vis-a-vis YouTube's community guidelines).

These criteria are:

- Whether the channel has more than 1 000 subscribers,
- Whether the channel has more than 4 000 hours of watch time over the last 12 months on its videos (excluding Shorts). As this is not directly observable, we made the same assumption as for Facebook (sum of number of views on videos posted in the last 12 months multiplied by the video length, divided by 2 to account for view drops).

We then aimed to check the effective monetization status of each. However, the monetization status of the channel is not available from official databases. The last ten videos published by a channel were collected and we observed how many of these videos had ads (either before or during the video playing, or in the top-right-hand corner of the video). Any channel with three or more videos displaying ads in the last ten was considered monetized.

5.3.7 Google Display Ads

While the Google Display Ads network operates a very different service from that of social media platforms, the monetization of websites repeatedly serving mis/disinformation content plays an important role in the for-profit disinformation ecosystem. In addition, Google as a whole is a signatory to the Code, reporting for some Commitments as Google Ads, which includes its Display Ads Network. Accordingly, we decided to include it in our

analysis. Google Display Ads do not set a traffic threshold for web domains to be eligible to have ads served by Google displayed on their website. No eligibility filtering step was therefore necessary.

	High credibility accounts			Low credibility accounts		
	Nr. of channels	Channels that meet eligibility criteria	Monetized channels (% of eligible)	Nr. of channels	Channels that meet eligibility criteria	Monetized channels (% of eligible)
Slovakia	20	20	16 (**80.0**%)	14	14	12 (**85.7**%)
Poland	21	21	17 (**81.0**%)	19	17	12 (**70.6**%)
France	25	25	15 (**60.0**%)	20	17	10 (**58.8**%)
Spain	44	41	36 (**87.8**%)	19	15	14 (**93.3**%)

Table 5.6 – Number and proportion of high- and low-credibility YouTube channels frequently displaying ads, indicating likely monetization.

Web domains were selected on the basis of the list of high- and low-credibility actors detailed in section 2.3. For each social media account, a manual search was performed to establish whether the entity or individual operating the account also had an official web domain. The social media account's credibility rating was extended to the web domain, under the assumption that their editorial standards were equivalent.

For each domain, up to five pages (randomly accessed from the homepage) were visited by the analyst. If the analyst saw ads served by one of Google's main ad-serving services (doubleclick.net or googlesyndication.com), the website was marked as being monetized by Google. If no such ads were found on the 5 pages, the website was marked as not being monetized by Google.

	High credibility websites		Low credibility websites	
	Number of websites	Monetized websites (%)	Number of websites	Monetized websites (%)
Slovakia	14	10 (**71.4**%)	24	8 (**33.3**%)
Poland	15	9 (**60.0**%)	30	5 (**16.7**%)
France	16	12 (**75.0**%)	29	8 (**27.6**%)
Spain	36	26 (**72.2**%)	30	9 (**30.0**%)

Table 5.7 – Number and proportion of High- and Low-credibility web domains displaying ads served by Google services, indicating monetization.

6. References

[1] Budak C, Nyhan B, Rothschild DM et al. (2024) Misunderstanding the harms of online misinformation. Nature https://doi.org/10.1038/s41586-024-07417-w

[2] Ecker U, Roozenbeek J, Van Der Linden S, Tay LQ, Cook J, Oreskes N, Lewandowsky S (2024) Misinformation poses a bigger threat to democracy than you might think. Nature https://doi.org/10.1038/d41586-024-01587-3

[3] The Code of Conduct on Disinformation https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation

[4] COMMISSION OPINION of 13.2.2025 on the assessment of the Code of Practice on Disinformation within the meaning of Article 45 of Regulation 2022/2065 https://ec.europa.eu/newsroom/dae/redirection/document/112679

[5] Nenadic I, Brogi E, Bleyer-Simon K (2024) Structural indicators to assess effectiveness of the EU's Code of Practice on Disinformation, EUI, Centre for Media Pluralism and Media Freedom https://hdl.handle.net/1814/75558

[6] European Digital Media Observatory (2024) Structural Indicators of the Code of Practice on Disinformation: The 2nd EDMO report https://edmo.eu/wp-content/uploads/2024/03/SIs_-2nd-EDMO-report.pdf

[7] Trustlab (2023) A Comparative Analysis of the Prevalence and Sources of Disinformation across Major Social Media Platforms in Poland, Slovakia, and Spain https://test2.disinfocode.eu/wp-content/uploads/2023/09/code-of-practice-on-disinformation-september-22-2023.pdf

[8] Trustlab (2024) A Comparative Analysis of the Prevalence of Misinformation and Sources of Disinformation across Major Social Media Platforms in Poland, Slovakia, Spain, and France https://test2.disinfocode.eu/wp-content/uploads/2024/09/code-of-practice-2-supplementary-report-designed-2024.pdf

[9] Chystoforova K & Reviglio U (2024) EDMO experts' feedback on structural indicators for the EU code of practice on disinformation. European University Institute (EUI) https://cadmus.eui.eu/server/api/core/bitstreams/4172771f-27e9-51fb-84b2-60df46c7f1a4/content

[10] Center for Countering Digital Hate (2021) The Toxic Ten How 10 fringe publishers fuel 69% of digital climate change denial. https://counterhate.com/wp-content/uploads/2021/11/211101-Toxic-Ten-Report-FINAL-V2.5.pdf

[11] DeVerna MR, Aiyappa R, Pacheco D, Bryden J, Menczer F (2024) Identifying and characterizing superspreaders of low-credibility content on Twitter. PLoS ONE https://doi.org/10.1371/journal.pone.0302201

[12] DeVerna MR, Aiyappa R, Pacheco D, Bryden J, Menczer F (2024) Identifying and characterizing superspreaders of low-credibility content on Twitter. PLoS One https://doi.org/10.1371/journal.pone.0302201

[13] Carniel B (2023) Consensus Credibility Scores: a comprehensive dataset of Web domains' credibility. Science Feedback https://science.feedback.org/consensus-credibility-scores-comprehensive-dataset-web-domains-credibility/

[14] Denniss E, Lindberg R (2025) Social media and the spread of misinformation: infectious and a threat to public health. Health Promotion International https://doi.org/10.1093/heapro/daaf023

[15] Allen J, Watts DJ, Rand DG (2024) Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. Science https://www.science.org/doi/10.1126/science.adk3451