Second Measurement of the State of Online Disinformation in Europe on Very Large Online Platforms

Second report of the **SIMODS** project

Structural Indicators to Monitor Online Disinformation Scientifically

Supported by
European MEDIA AND INFORMATION Fund
Managed by
Calouste Gulbenkian Foundation

Executive Summary

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

Against a backdrop of platforms walking back earlier commitments to counter disinformation, this second report brings something no single measurement can offer: a basis for comparison. The consistency of results across two independent measurement periods strengthens the credibility of the findings and confirms that what we are measuring is not noise, but structural features of the platforms themselves.

WHAT WE MEASURED

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

The second data collection period ran throughout October 2025, covering five topics (the Russia–Ukraine war, climate change, health, migration, and national politics) and yielding approximately 3.3 million posts. 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 of posts. TikTok and YouTube provided API access, which required additional effort to produce comparable results. This concerns publicly available data: the barrier platforms are erecting against independent researchers has no technical justification.

For non-public data, including monetisation records, there was no cooperation from any platform. This opacity makes it practically impossible to study the systemic risks these platforms impose on society, as the DSA requires.

KEY FINDINGS

1) **Prevalence.** TikTok shows the highest prevalence of mis/disinformation (~25% of exposure-weighted posts), up from ~20% in the first measurement period. YouTube also saw a notable increase, from ~8.5% to ~12%. Facebook (~15%), X/Twitter (~11%), and Instagram (~8%) remained broadly stable. LinkedIn continues to show the lowest prevalence at ~1%.

   When including abusive (e.g., hate speech) and borderline content (content that reinforces a disinformation narrative without making an outright false claim), levels are substantially higher: TikTok reaches ~43% problematic content, Facebook ~34%, X/Twitter ~32%, YouTube ~27%, Instagram ~16%, and LinkedIn ~4%. Notably, three platforms (TikTok, X/Twitter, and YouTube) now show more problematic content than credible content in our samples, compared to only one (X/Twitter) in the first measurement period.

   Health misinformation remains the dominant category across all platforms (~43% of all mis/disinformation posts).

2) **Sources.** Across almost all platforms, low-credibility accounts receive disproportionately high engagement relative to their audience size, a pattern we term the "**misinformation premium**". On most platforms, this premium persisted or worsened compared to the first measurement period: on X/Twitter it rose from ~4 to ~10, and on YouTube from ~8.5 to ~11. This means that on X/Twitter, an account posting false or misleading information repeatedly now receives around 10 times as much engagement per post as a credible source with a comparable following.

3) **Monetisation.** Monetisation data remain entirely inaccessible on four of the six platforms. On YouTube, 81% of eligible low-credibility channels appear to benefit from monetisation, compared to 90% of eligible high-credibility channels. On Facebook, the gap is wider (22% vs. 51%). In both cases, the fact that a high proportion of eligible low-credibility accounts appear to be monetised indicates that demonetisation policies are not functioning as intended. These results are consistent with those of the first measurement period: platforms are, to a meaningful extent, benefiting from and financially sustaining the very accounts that repeatedly spread misleading content.

4) **Consistency across measurement periods.** The overall coherence of results between the two waves is a key finding in itself. Prevalence estimates, the misinformation premium and monetisation patterns are consistent with those observed in the first wave. This reproducibility confirms that our methodology is sound and that the phenomena we measure are structural, and not incidental.

5) **AI-generated disinformation.** This wave introduces a new indicator tracking the share of mis/disinformation that is AI-generated. On video platforms, AI-generated content accounts for approximately one quarter of all identified mis/disinformation on TikTok (24%) and approximately one fifth on YouTube (19%). For a phenomenon that barely existed a few years ago, these figures indicate rapid growth and a significant and escalating risk to the quality of public information. Health misinformation accounts for the largest share of AI-generated mis/disinformation on both platforms.

   Critically, the overwhelming majority of this content carries no label: across all platforms, only 16.5% of AI-generated mis/disinformation was visibly marked as synthetic. This is a failure by platforms to inform their users of what they are watching, and to protect them from manipulation and deception. The prevalence of unlabelled AI-generated health misinformation, including fabricated videos featuring AI avatars posing as medical professionals, illustrates concretely the real-world harms this failure enables.

6) **Audience growth.** This wave also introduces a new indicator tracking the relative growth rate of audiences for high- and low-credibility accounts. On most platforms, no statistically significant difference in follower growth was observed between the two groups. One exception is X/Twitter, where low-credibility accounts are growing their audiences at ~3.5 times the rate of high-credibility accounts. X/Twitter thus appears to favour the expansion of accounts that repeatedly share misleading content.

WHY THIS MATTERS

Two waves of measurement, using a consistent methodology now point to the same conclusion: the structural permissiveness of major online platforms to misleading content appears to be a persistent feature of how these platforms are designed and operated.

The integration of the Code of Conduct on Disinformation into the DSA framework in 2025 creates, for the first time, a legal basis for enforcement. The indicators developed by the SIMODS project are designed to serve that purpose: they are comparable across platforms, reproducible over time, and grounded in independent, transparent methodology. What is now required is the political and regulatory will to use them.

1. Introduction

Six months after the publication of the first SIMODS report, the context in which disinformation on online platforms is discussed has continued to shift in important ways. Early 2025 saw several major platforms step back from earlier voluntary commitments to counter disinformation, reducing fact-checking programmes, staffing in relevant teams, or withdrawing support for disinformation research and efforts to counter it, moves widely reported and often framed as responses to political pressure in the United States. At the same time, the Code of Conduct on Disinformation was formally integrated into the Digital Services Act framework, becoming operational in July 2025 and for the first time giving independent measurement a direct role in regulatory enforcement.

The debate over whether platforms are saturated with misleading content, or whether such content represents only a marginal share of what users see, is no longer merely academic[1-2]: it now has legal and policy consequences. This report provides the evidence to anchor that debate.

1.1 THE CODE OF CONDUCT ON DISINFORMATION

The Code of Conduct on Disinformation (ex 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.

With this objective, the SIMODS 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 present report focuses on measuring four Structural Indicators:

1) Prevalence of Disinformation
2) Sources of Disinformation (engagement and growth);
3) Monetisation of Disinformation
4) AI-Generated Disinformation.

This European Media and Information Fund-funded project spans 18 months. The first report of SIMODS was published in September 2025[7], based on data collected in the spring of 2025. One indicator was replaced: the Cross-platform Aspects of Disinformation indicator included in the first report has been replaced by an indicator on AI-generated mis/disinformation, reflecting the rapid growth of this phenomenon and the need to track it systematically. In addition, the Sources indicator has been extended to include a measure of audience growth for high- and low-credibility accounts.

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).

While previous attempts have been made to measure Structural Indicators, most notably TrustLab's pilot implementation[8,9], 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), several platforms did not provide datasets in time for our analysis following our requests. Only LinkedIn provided the random sample that we requested, TikTok and YouTube granted API access that allowed us to collect data for this second 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 content that is 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.

Following the publication of the first SIMODS report, UOC researchers within the consortium conducted a study examining an additional source of uncertainty in our prevalence estimates: the sensitivity of results to the choice of keywords used for data collection while the present and previous reports quantify uncertainty arising from sampling and inter-annotator disagreement. It has been submitted for publication[10].

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 definition is typically contrasted with "misinformation", which refers to false or misleading information spread unintentionally, without deceptive intent. However, it is not possible to formally identify intentionality when assessing isolated posts, as is required when measuring prevalence at scale. In this report, we use the shorthand "mis/disinformation" to refer to all false or misleading information, making explicit that both intentional and unintentional 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, i.e., the proportion of content 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 response 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"[11].

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 central to the public debate in Europe and at high risk of containing 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[8,9]. 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 minimise 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, national 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, and Slovak) and was balanced, in each country, with equal numbers of *neutral* keywords on one hand and of *ambiguous* + *misinformation-related* terms on the other hand. More details about the keywords can be found in Appendix 5.1.1. The second data collection period (the one analysed in this second SIMODS report) spanned 1 October to 31 October 2025, and we collected posts that were 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 to request a random sample of 200 000 posts per language, efforts that started with the first iteration of the report on the 19 December 2024 and continued with a second request to the platforms on the 7 October 2025.

As only LinkedIn provided the requested dataset and TikTok and YouTube provided access to their APIs, we relied on a second method for the other platforms, using their search functions and third-party 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 second data collection period, we assembled a dataset comprising approximately **3.3 million posts** (with metadata) across four languages and six platforms, totalling around **18 billion views**. This is comparable to the data collected for the first measurement period comprising approximately 2.6 million posts, 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 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.

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 labelled 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 labelled, 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.

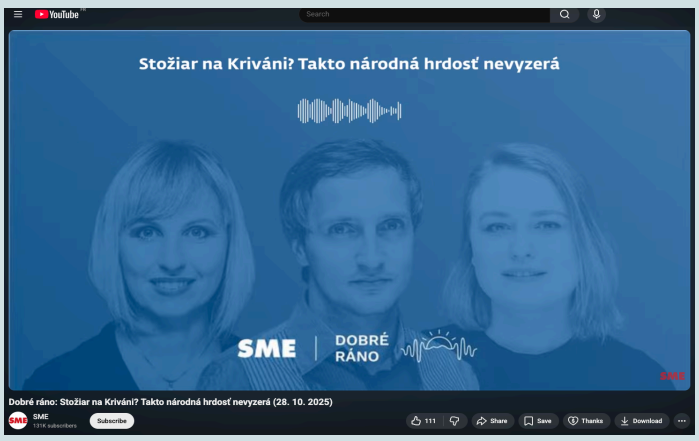
2.1.2 Results

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.

CREDIBLE POST (Instagram)

An Instagram post from a reputable source, updating users on the Ukrainian War and attacks conducted by Russia.

UNVERIFIABLE VIDEO (YouTube)

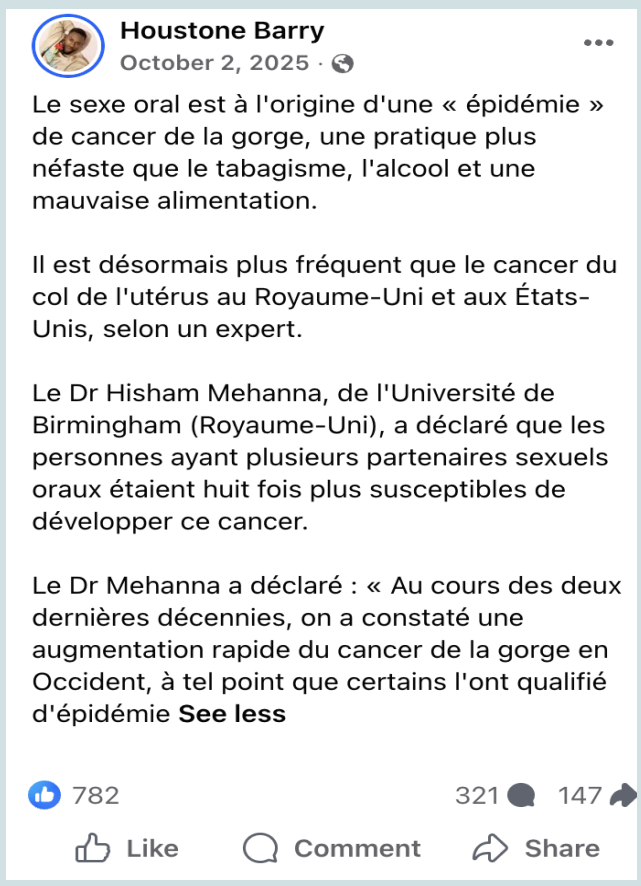


A YouTube video discussing national laws in Slovakia. The topic is relevant to our study, but the post presents only a subjective experience without verifiable information.

IRRELEVANT POST (LinkedIn)

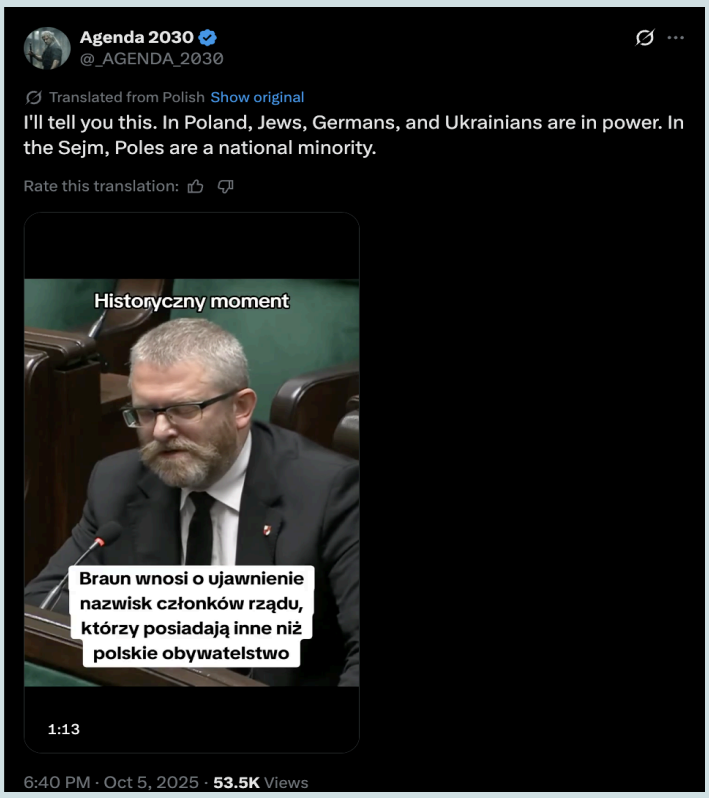
A LinkedIn post promoting insurance policies for people diagnosed with breast cancer. This content does not provide information that informs readers on societal issues.

MISINFORMATION POST (Facebook)



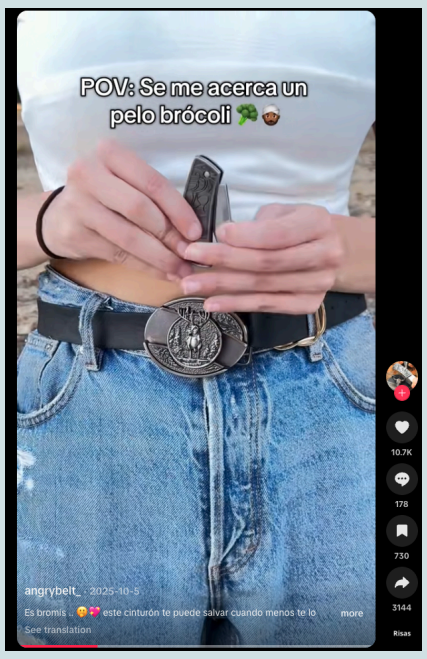
A Facebook post stating that oral sex causes throat cancer and is more detrimental than smoking, drinking alcohol and low-quality food.

BORDERLINE POST (X/Twitter)



An X/Twitter post targeting Jewish and Ukrainian people, insinuating that they exert conspiratorial control over the Polish government.

ABUSIVE VIDEO (TikTok)



A TikTok video targeting Muslim immigrants using a derogatory term to refer to them (pelo brocoli) and inciting to violence (handling of a knife “if one approaches”).

Figure 2.1 – Screenshots of examples of posts that were annotated as belonging to each of the main categories.

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 60% while YouTube, TikTok and X/Twitter have the lowest at around 21%, 24% and 26.7% respectively. 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, Facebook and X/Twitter showing the highest levels at 43%, 34% and 32%, respectively.

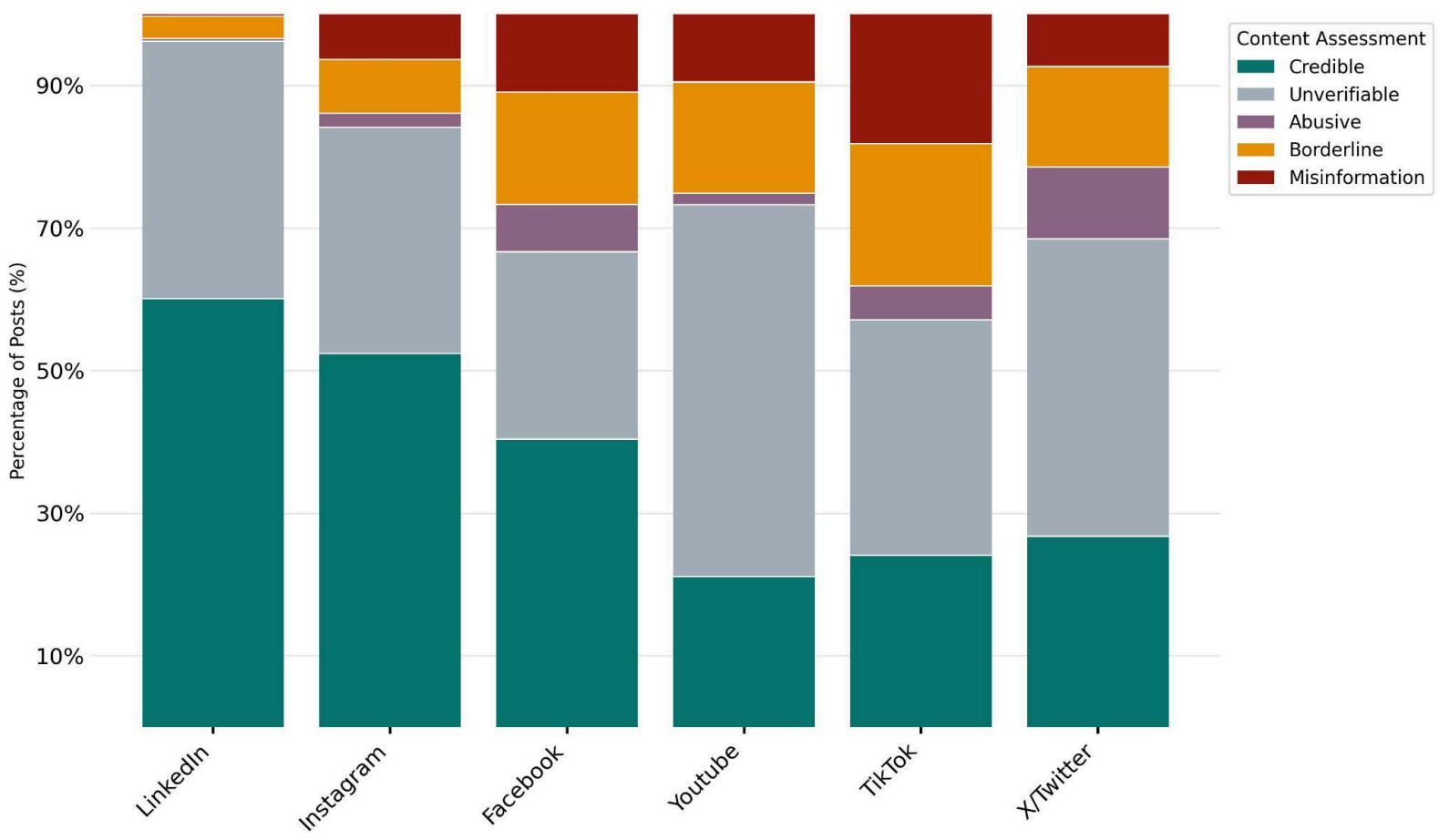


Figure 2.2 – Percentage of posts belonging to each category for the six very large online platforms.

When comparing only *Credible* to *Problematic* content, we note that half of the platforms covered (X/Twitter, TikTok, and YouTube) contain more *Problematic* content than *Credible* content, while this was the case only for X/Twitter in the first measurement period (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 labelled as *Mis/disinformation*, *Credible*, and *Unverifiable*, respectively.

This prevalence calculation is designed to approximate the proportion of misinformation that users are likely to encounter on each platform, as compared to credible and opinionated content on the same topics. In this calculation, posts that are *Irrelevant* (humorous, satirical, or otherwise unrelated to the informational dimension of the topic), are excluded from the calculation.

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 25% [22.6%, 27.5%], indicating that roughly one in four posts on the platform regarding the topics we investigated contains misleading or false information.
- **Facebook**, **YouTube** and **X/Twitter** follow with elevated prevalence at 15% [13.2%, 16.8%], 12% [10.6%, 13.9%] and 11% [9.0%, 12.2%], respectively.
- **Instagram** has a prevalence of about 8% [6.5%, 9.2%].
- **LinkedIn** has the lowest prevalence of mis/disinformation at around 1% [0.5%, 1.5%], 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 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.

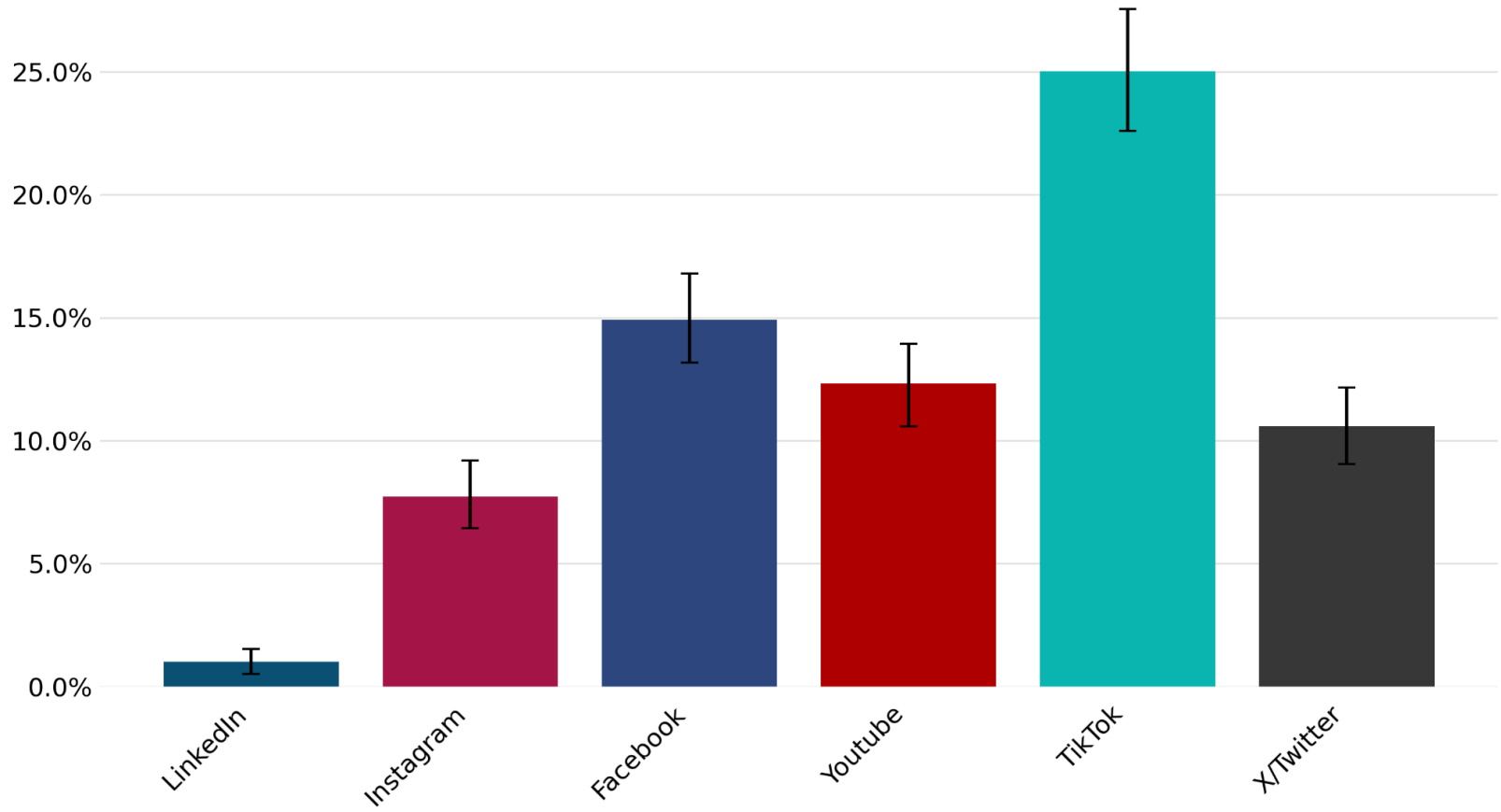


Figure 2.3 – Prevalence of mis/disinformation across the six very large platforms, aggregated across all languages. The error bars represent the 95% confidence intervals measuring the uncertainty around each estimate, calculated using a bootstrapping method (see Appendix 5.1.4).

Platform	Prevalence [CI 95%]
LinkedIn	**1.0%** [0.5%, 1.5%]
Instagram	**7.7%** [6.5%, 9.2%]
Facebook	**14.9%** [13.2%, 16.8%]
YouTube	**12.3%** [10.6%, 13.9%]
TikTok	**25.0%** [22.6%, 27.5%]
X/Twitter	**10.6%** [9.0%, 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 42.8% (Figure 2.4). The Russia-Ukraine war is the second most represented topic, with about 23% of mis/disinformation posts, followed by national politics (12%), which typically includes election-related claims in the country of interest or controversies surrounding new legislation. Migration and climate account for 7.7% and 6.4% of mis/disinformation posts, respectively. These results are broadly consistent with what we observed during the first period (Figure 5.9).

To better understand the nature and diversity of the claims made in these misleading posts, and provide a qualitative understanding of the most prevalent misleading claims, we performed an analysis of the narratives (or "broad messages") they convey. Using an LLM, we clustered the claims contained in the posts based on their semantic similarity and classified them into narratives building on the methodology of Huang and He[12].

Within the health category, the most prevalent misleading narrative in the entire dataset is that "Major diseases can be treated with diet, natural or folk remedies". Examples of claims that fall under this narrative include "Lemon detoxifies the body" or "Essential oils have been proven to cure cancer", for instance. The second most prevalent narrative is that "Conventional medicine is unnecessary / harmful", with examples of claims here being that "the majority of medical decisions are incorrect" or that "long-term use of medications leads to additional health problems rather than solutions". These two narratives complement each other in their attempt to cast doubt on scientifically supported health advice and replace them with unproven, alternative approaches.

The third narrative by number is that "Vaccines are broadly unsafe, untested, and/or promoted by a corrupt pharmaceutical complex". Claims under this narrative include: "vaccinated children have higher rates of neurodevelopmental disorders", "mRNA vaccines are gene therapies" and "COVID-19 vaccines are linked to increase in cancer rates".

As can be seen in Figure 2.5, the platform which holds the most health misinformation is Instagram, with about 80% of the misinformation posts on the platform being health-related, followed by TikTok and Facebook where it is about half.

For the Russia-Ukraine war topic, the most prevalent narrative identified is that "Ukraine is an inherently corrupt puppet state controlled by foreign powers, not a legitimate sovereign actor". Examples of claims under this narrative include: "Ukraine is financially, politically, and militarily controlled by foreign powers" and "Zelensky is transferring 50 million dollars every month to a bank in Saudi Arabia". The second-most represented narrative is "The West, NATO, and the EU are orchestrating or provoking conflicts (including Ukraine) as proxies to weaken other states"; an example of a claim under this cluster is "The West spent nearly five billion dollars to support a coup in Ukraine".

For the National Politics topic, no obvious cluster stands out due to the diversity of the claims that are made in each country.

Regarding migration-related misinformation, the most prevalent narrative in our dataset relates to the so-called 'Great Replacement': "Immigration is a coordinated plot to replace native populations and destroy national identity". Examples of claims include "'The native

French population is being systematically replaced by immigrants as part of a political agenda to alter France's cultural identity" and "The majority of migrants coming to Spain are criminals". The platforms where migration-related themes are most prevalent are X/Twitter and TikTok, where they represent 16% and 10% respectively of all misleading posts (Figure 2.5).

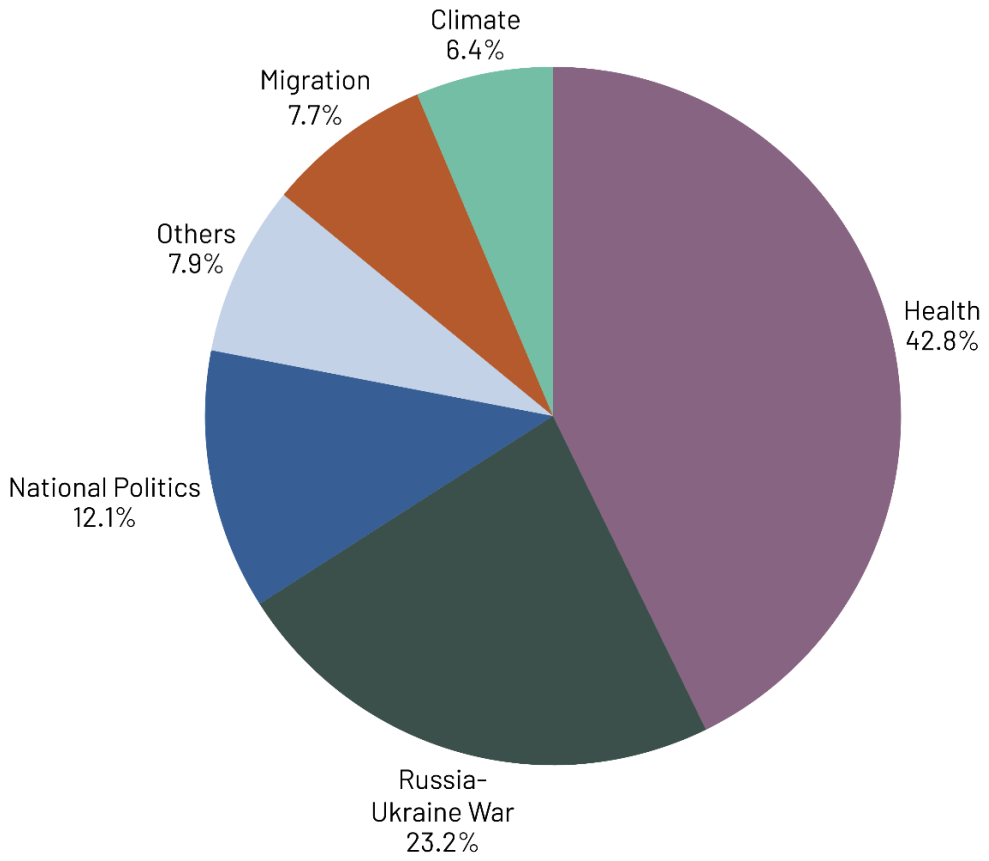


Figure 2.4 – Topic distribution of mis/disinformation posts across the studied data sample.

For the Climate topic, three clusters appear as roughly equally frequent: "Human-caused climate change and the role of CO_2 are massively overstated or irrelevant", "Geoengineering is a secret government program deliberately manipulating weather to harm populations", "Environmental policies (e.g. the Green Deal) are a deliberate scheme to seize property, impoverish ordinary citizens and impose social control". The fourth cluster identified by prevalence is the narrative that "Renewable energy in general is unreliable, not truly clean, or promoted as an economic scam".

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 that of *Mis/disinformation* to create an indicator of the prevalence of harmful, or "problematic", content.

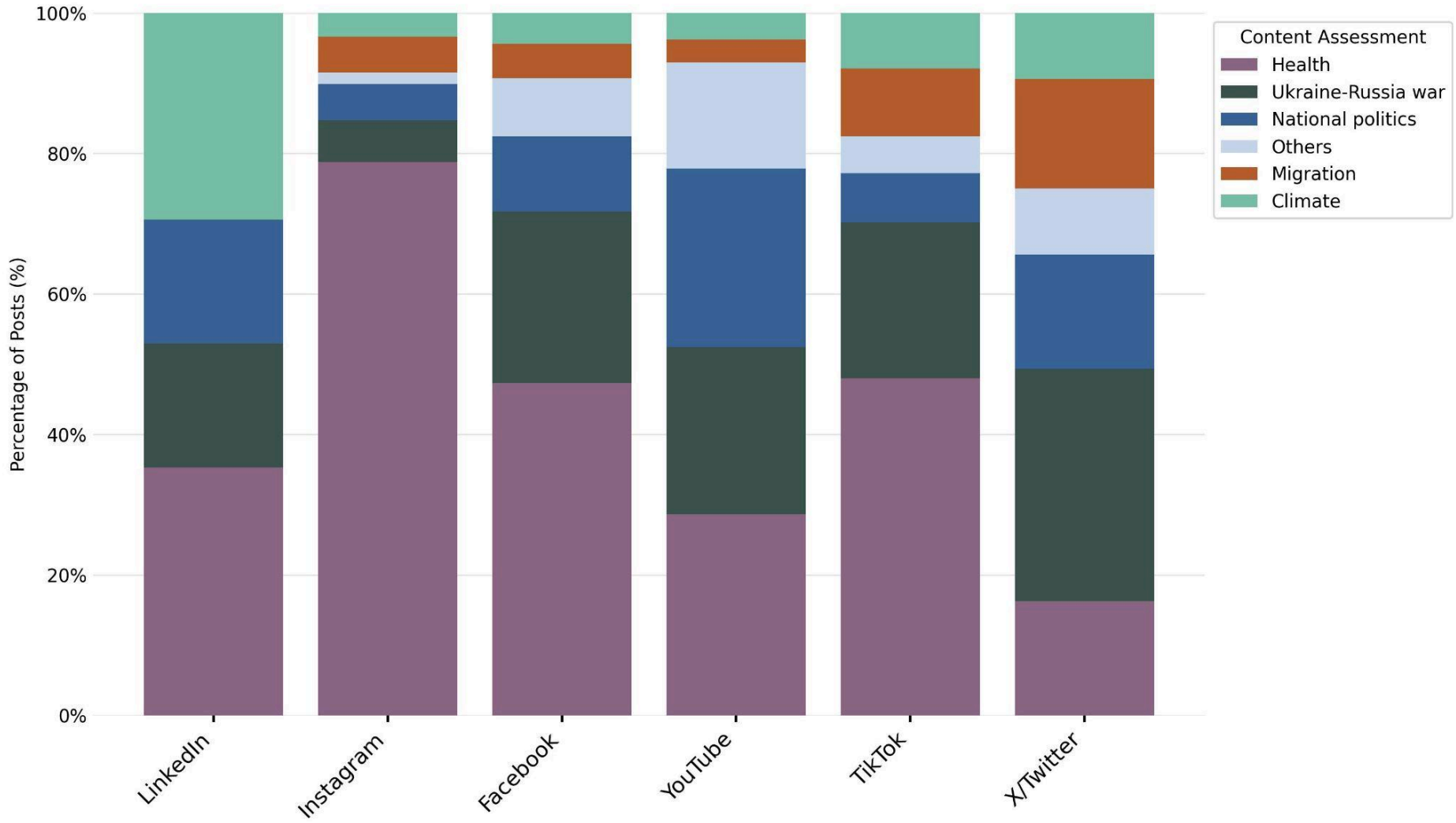


Figure 2.5 – Topic distribution of mis/disinformation posts across the platforms studied; same data as Figure 2.4 separated by platform

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 labelled as *Mis/disinformation*, *Borderline*, *Abusive*, *Credible*, and *Unverifiable*, respectively.

Figure 2.6 shows that the prevalence of *Problematic* content is significantly higher than the prevalence of *Mis/disinformation* on all platforms.

With this metric, TikTok appears as the platform with the highest prevalence of Problematic content, with prevalence values of 43% [40.6%, 45.9%], which means that roughly two out of five of the posts on TikTok either contain false or misleading information, include abusive language, reinforce misleading narratives, or promote other forms of harmful content. Facebook and X/Twitter take the next two spots, with a prevalence of about 34% [31.5%, 35.9%] and 32% [29.7%, 33.9%], the overlapping confidence intervals indicating that the values for these two platforms are not statistically different. YouTube ranks fourth at 27% [24.9%, 29.2%], followed by Instagram at 16% [14.4%, 17.9%]. Lastly, LinkedIn reports the

lowest prevalence at 4% [3.1%, 5.2%], indicating a comparatively safer information environment, a finding consistent with the results from the first SIMODS Report.

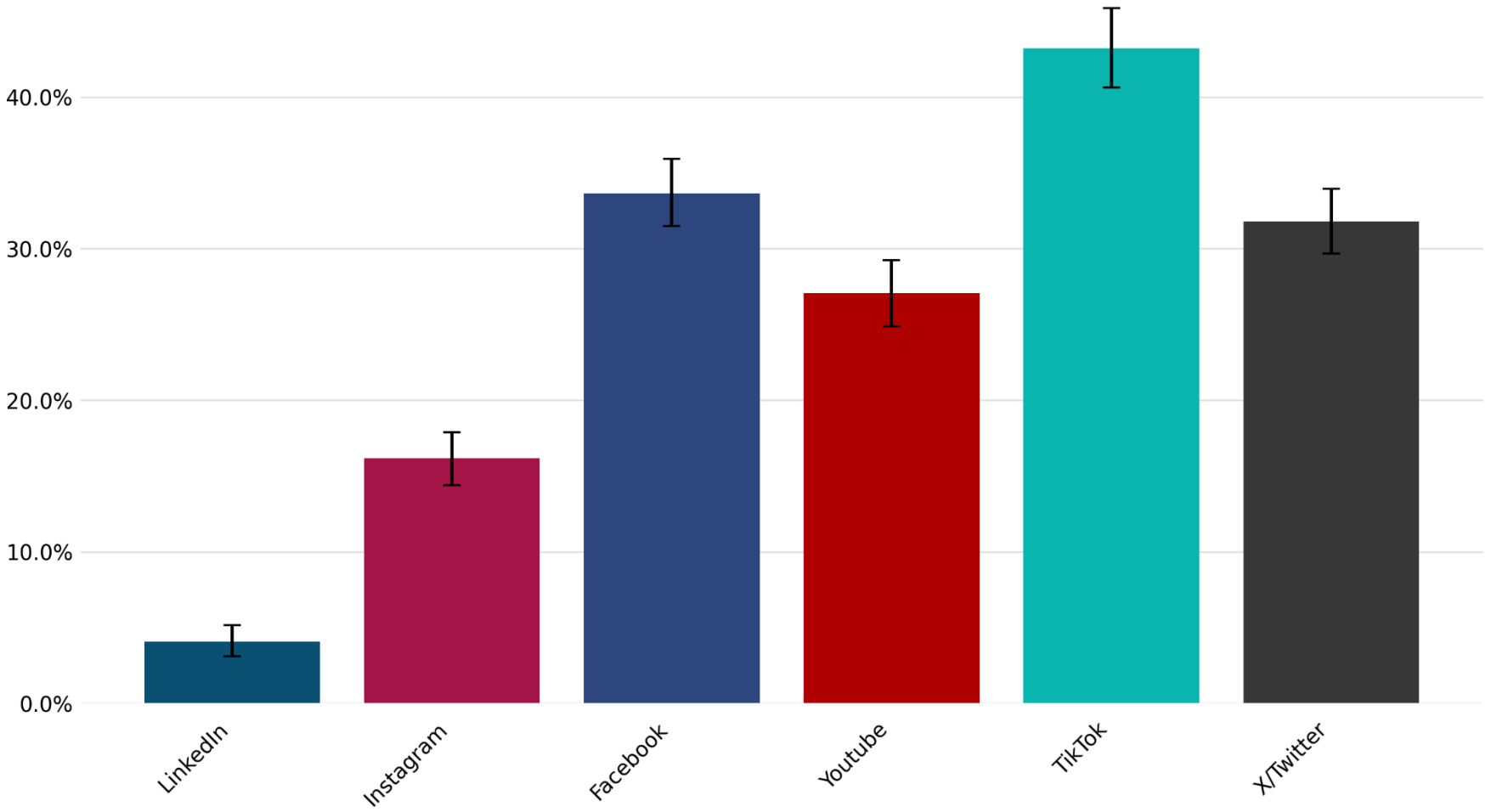


Figure 2.6 – Prevalence of *Problematic* content (defined by the grouping of *Mis/disinformation*, *Borderline* and *Abusive* content) across the six very large platforms, aggregated across all languages. The error bars represent the 95% confidence intervals measuring the uncertainty around each estimate, calculated using a bootstrapping method (See Appendix 5.1)

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[13-15], and accounts that share mis/disinformation at a given point in time tend to continue doing so in the future[15].

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 prioritise content from trustworthy sources while reducing the prominence of misleading or harmful content.

Platform	Prevalence [CI 95%]
LinkedIn	**4.1%** [3.1%, 5.2%]
Instagram	**16.2%** [14.4%, 17.9%]
Facebook	**33.6%** [31.5%, 35.9%]
YouTube	**27.0%** [24.9%, 29.2%]
TikTok	**43.2%** [40.6%, 45.9%]
X/Twitter	**31.8%** [29.7%, 33.9%]

Table 2.2 – Prevalence of *Problematic* content across the six very large platforms, aggregated across all languages (same values as in Figure 2.6). 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 behaviours 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 accounts' audience size, 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 (October 1 – October 31). After excluding accounts that mostly shared content deemed irrelevant according to the project's definition (see Appendix 5.2.2), 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. Recognising 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 to which category each account belonged by reviewing their posts over the study period.

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, at the beginning of the project, before the first data collection period. 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[16]. The "Fact-checkers' list" of accounts used in this second measurement period is the same as the one used in the first period[7].

C. COMPARISON AND MERGING OF THE TWO LISTS

Upon comparing the two lists, we note 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 identified accounts of politicians or political parties, categorizing them as 'Political'. These accounts were excluded from the primary analysis presented in the Results below (except for Figure 2.10).

2.2.2 Results

A. ACCOUNTS AUDIENCE SIZE

The first observation is that high-credibility accounts consistently have larger audiences than low-credibility accounts. As shown in Figure 2.7, 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.

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 organisations in the high-credibility list and prominent "influencers" in the Neither category. In contrast, low-credibility sources display less variability in their follower counts, with a narrower distribution, indicating a more homogeneous audience size.

Similar results are obtained when we analyse the median number of followers per group instead of the average, showing the differences are not driven by a handful of very large channels (see Appendix 5.2.5).

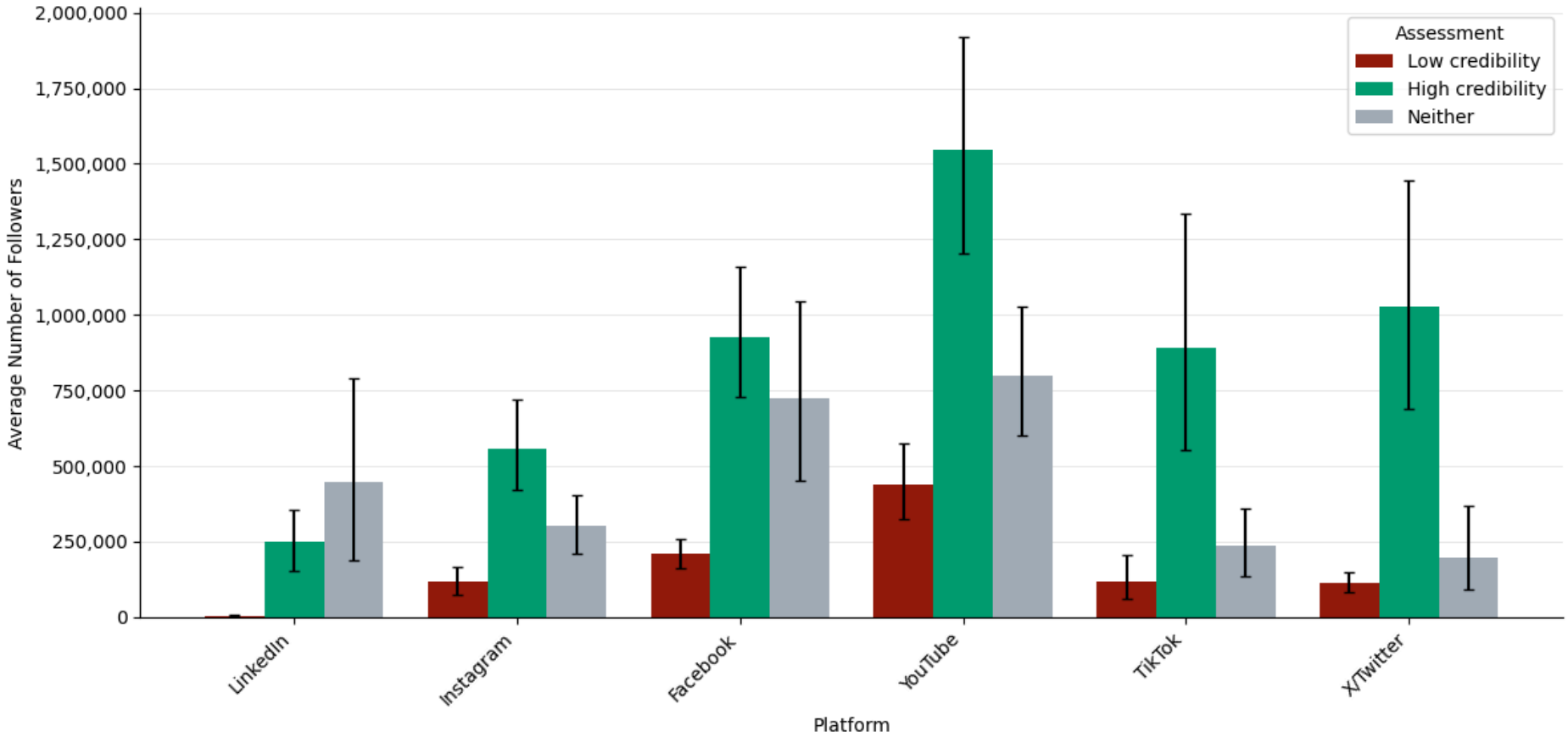


Figure 2.7 – Average number of followers for accounts classified as High-credibility, Low-credibility, and Neither on each platform. Error bars represent 95% confidence intervals, calculated using a bootstrapping method (see Appendix 5.1.4).

Platform	High-credibility	Low-credibility	Neither
LinkedIn	**249** [154, 356]	**5** [2.0, 8.8]	**446** [187, 791]
Instagram	**557** [421, 720]	**116** [73, 168]	**303** [211, 405]
Facebook	**926** [728, 1 157]	**208** [160, 260]	**726** [452, 1 044]
YouTube	**1 546** [1 203, 1 919]	**440** [324 , 578]	**797** [599, 1 028]
TikTok	**893** [552, 1 335]	**117** [61, 207]	**234** [137, 361]
X/Twitter	**1 029** [688, 1 444]	**115** [84, 151]	**198** [90, 368]

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 2.7. 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).

B. ACCOUNTS' ENGAGEMENT RATES: THE 'MISINFORMATION PREMIUM'

While followership provides insight into audience size of high- and low-credibility accounts, 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 (Figure 2.8). Normalising 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.

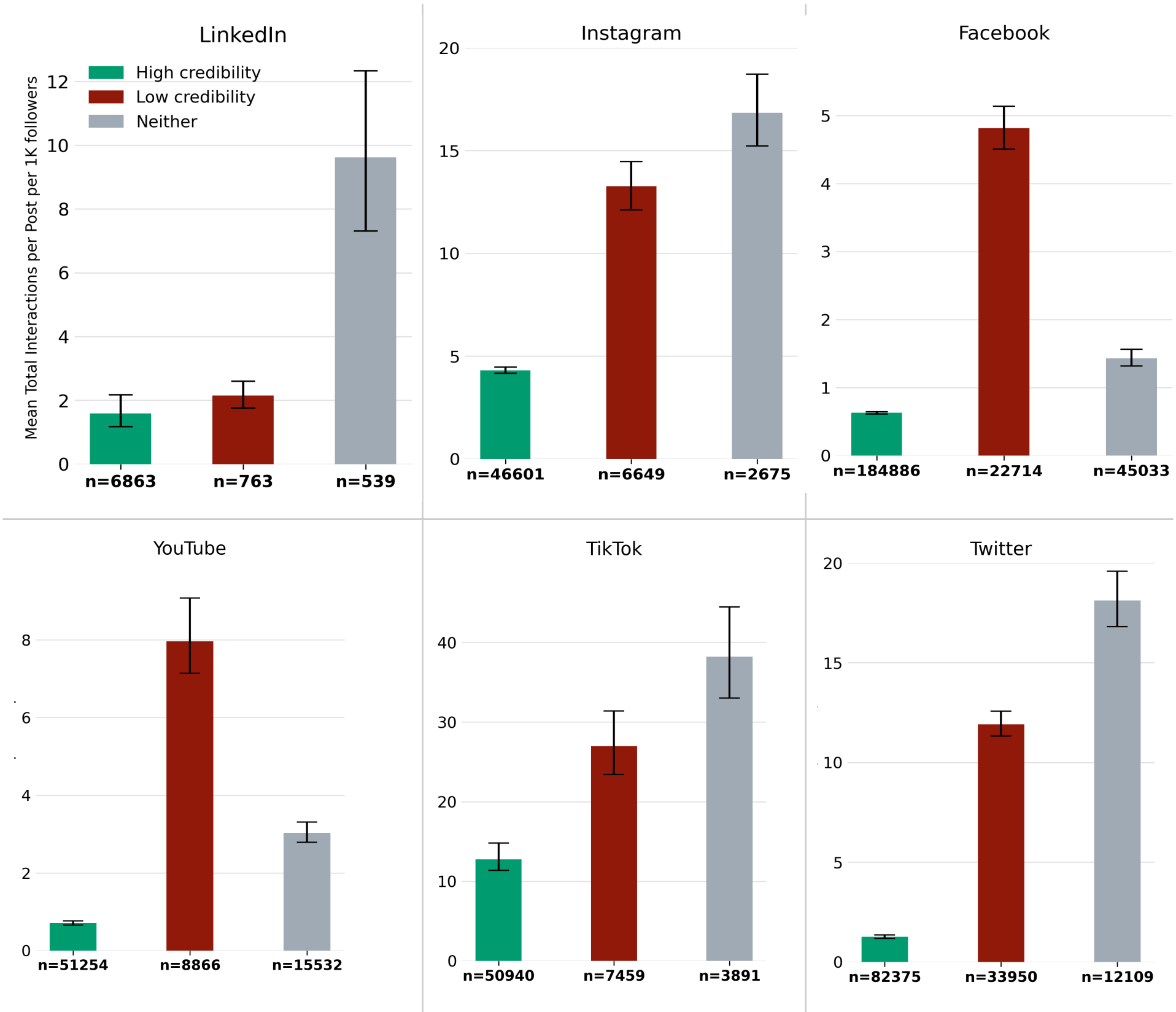


Figure 2.8 – Average number of interactions per post per 1 000 followers for accounts classified as High-credibility, Low-credibility, and Neither on each platform. Error bars represent 95% confidence intervals, calculated using a bootstrapping method (see Appendix 5.1.4 for details).

Figure 2.8 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 2 interactions per post per 1 000 followers on LinkedIn and 5 on Facebook, while on TikTok, this figure is much higher at about 27.

Platform	High-credibility	Low-credibility	Neither
LinkedIn	**1.6** [1.2, 2.2]	**2.2** [1.8, 2.6]	**9.6** [7.3, 12.3]
Instagram	**4.3** [4.2, 4.5]	**13.3** [12.1, 14.5]	**16.8** [15.2, 18.7]
Facebook	**0.62** [0.61 - 0.64]	**4.8** [4.5 - 5.1]	**1.4** [1.3 - 1.6]
Youtube	**0.7** [0.66, 0.76]	**8** [7.1, 9.1]	**3.0** [2.8, 3.3]
TikTok	**12.7** [11.4, 14.8]	**27** [23.4, 31.4]	**38.2** [33.0, 44.5]
X/Twitter	**1.26** [1.19, 1.35]	**11.9** [11.3, 12.6]	**18.1** [16.8, 19.6]

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 2.8. 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).

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 2.9). On YouTube, low-credibility accounts receive more than 11 times the engagement of high-credibility accounts. X/Twitter shows a similar pattern, exhibiting a ratio of about 10x, followed by Facebook with a ratio of 9x.

Instagram and TikTok show more moderate but still significant ratios of approximately 4x and 2x, respectively. The lowest ratio is observed on LinkedIn (1x), which is the only platform where low-credibility accounts do not outperform high-credibility accounts in interactions per post per follower.

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 was calculated by aggregating the number of comments, shares and likes of each post.

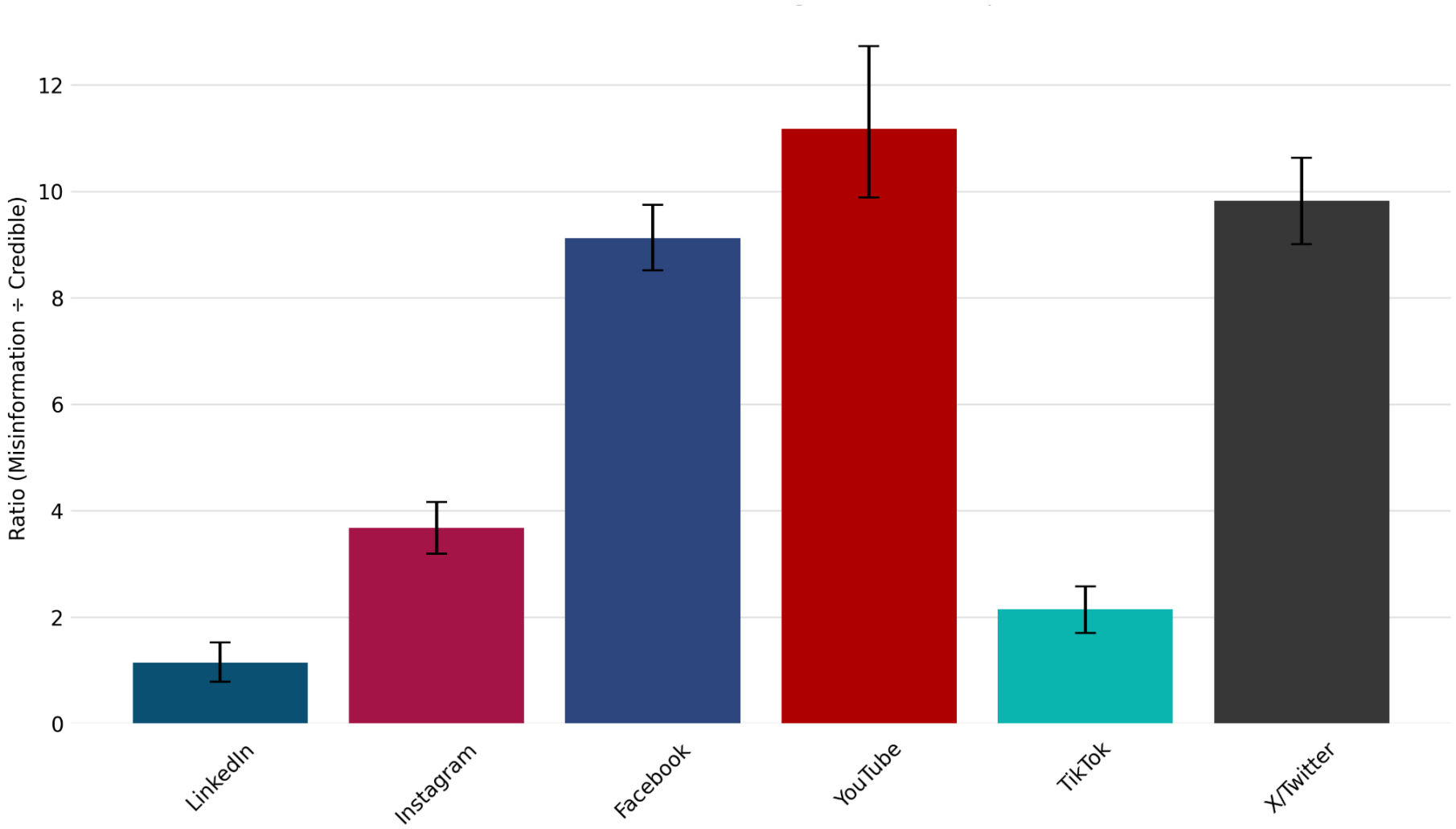


Figure 2.9 – Misinformation premium, i.e., the ratio of the average number of interactions per post per 1 000 followers for accounts classified as Low-credibility to the same number for High-credibility accounts, for each platform. Error bars represent 95% confidence intervals, calculated using a bootstrapping method (see Appendix 5.1.4).

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.

For a comparison of the results presented here and in the previous measurement period, see Appendix 5.4.2-B.

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 2.10 shows that the platform with the highest proportion of low-credibility accounts in the Top 50 is TikTok, with about 34%, followed by X/Twitter with 26%. On Facebook and YouTube, about 23% of accounts in the Top 50 are low-credibility. The platform with the smallest proportion of low-credibility accounts is LinkedIn (2%).

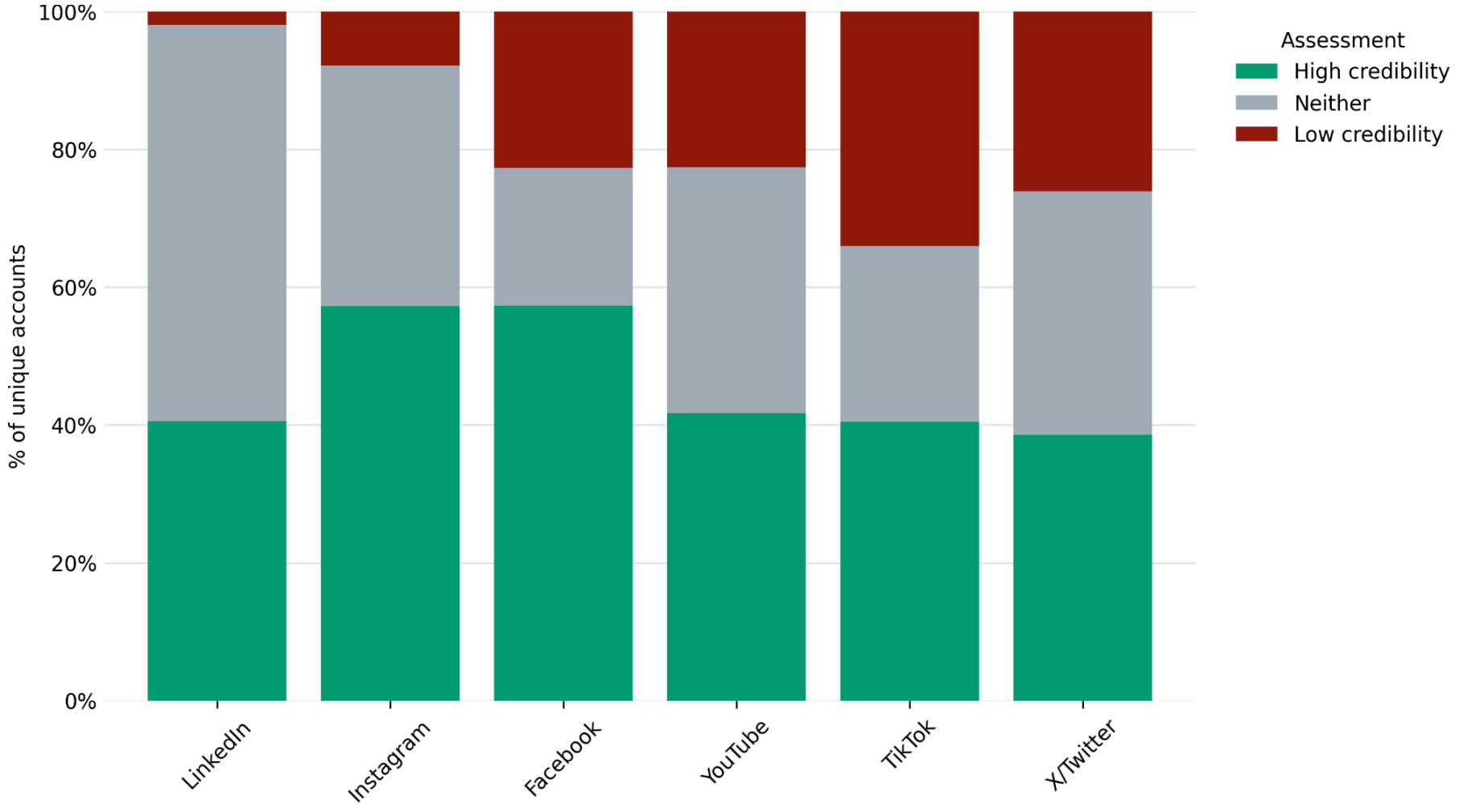


Figure 2.10 – Share of accounts in the Top 50 that are Low-credibility, High-credibility and Neither for each platform. ('Political' accounts are included in this figure.)

2.3 MONETISATION of MIS/DISINFORMATION

Beyond measuring how prevalent mis/disinformation is on platforms and how influential its sources are, we need to investigate whether platforms financially incentivise its dissemination. Commitment 1 of the Code of Conduct on Disinformation sets out a series of Measures aimed at reducing the economic incentives that sustain disinformation ecosystems. These include:

- adopting policies to avoid placing advertising next to mis/disinformation content;

- ensuring that accounts which systematically violate platform policies cannot benefit from monetisation programmes; and
- providing greater transparency, including enabling independent scrutiny of the effectiveness of such measures.

Although several platform Signatories (except LinkedIn) unsubscribed from certain demonetisation-related measures in early 2025, the underlying question remains unchanged: to what extent are platforms funding actors that repeatedly disseminate mis/disinformation? Disinformation does not only spread because of engagement dynamics, it is often incentivised by revenue-generating models, including advertising revenue sharing, creator funds, sponsored content, and other platform-based monetisation schemes. As such, a robust Code monitoring framework should include a Structural Indicator on monetisation.

2.3.1 Data Access

As pointed out in EDMO's second report, a methodologically sound Structural Indicator on demonetisation requires data that is, so far, inaccessible to outside researchers using publicly available data[6].

Since November 2025, the European Commission has activated the DSA Data Access Portal, a system designed to operationalise researcher access to data from Very Large Online Platforms (VLOPs).

In the context of this project, we submitted six separate data access requests through this mechanism on 12 January 2026, seeking two data points that are essential for constructing a methodologically sound Structural Indicator on monetisation:

- the total amount of revenue generated by a given account from the platform, across monetisation streams (e.g., ad-revenue sharing, tipping, creator partnerships, on-platform shops);
- whether a specific piece of content contributed to the creator's revenue stream.

These data points are directly linked to financial transfers and should therefore be readily available to platforms. At the time of writing, we have not received a response to these requests, nor have we been contacted by any of the platforms concerned.

2.3.2 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 monetisation features, such as brand partnerships (disclosed or undisclosed), tipping, or subscriptions, were left out, but should be requested in future iterations.

In most cases, access to ad-revenue sharing programmes depends on two types of criteria: (i) meeting minimum activity and audience thresholds (e.g., number of videos posted, cumulative views, or watch time), and (ii) being in good standing under the platform's community guidelines.

To isolate, as far as possible, the second dimension, where the impact of disinformation-related enforcement should materialise, we adopted a comparative approach. Specifically, we distinguished between high-credibility and low-credibility accounts. Starting from the Sources of Mis/disinformation dataset described in Section 2.2.1, we retained only accounts that had published at least one monetisation-eligible piece of content during the September–November 2025 period. We then excluded accounts that did not meet the publicly observable activity and audience thresholds, leaving only accounts that were plausibly eligible for monetisation (see Appendix 5.3 for platform-specific details).

We hypothesised that, under properly functioning demonetisation systems, eligible high-credibility accounts would be monetised to a large extent, while low-credibility ones would not be monetised.

2.3.3 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 monetisation status and were consequently left out (see Appendix 5.3 for discussions).

Table 2.5 summarises our results, showing the number of channels or accounts in our dataset, the number of eligible accounts, and the proportion of eligible accounts for which we have been able to confirm monetisation.

In absolute terms, we observe that none of the platforms is fully successful in ensuring that low-credibility accounts do not receive a share of ad revenue. However, we observe substantial differences: on Facebook, around 22% of eligible low-credibility accounts appear monetised; by contrast, on YouTube, 81% of eligible low-credibility channels are monetised. High- and low-credibility actors are monetised at similarly high rates on YouTube (90% vs 81%), whereas Facebook shows a larger gap (51% vs 22%). These results are consistent with the ones we observed for the first measurement period[7].

Platform	High-credibility accounts			Low-credibility accounts		
	Nr. of accounts	Eligible accounts	Monetized accounts (% of eligible)	Nr. of accounts	Eligible accounts	Monetized accounts (% of eligible)
Facebook	190	144	73 **(51%)**	134	51	11 **(22%)**
YouTube	111	111	100 **(90%)**	107	103	83 **(81%)**

Table 2.5 Number and proportion of high- and low-credibility accounts likely benefitting from ad-revenue sharing with different services on Facebook and YouTube

2.4 AI-GENERATED DISINFORMATION

With the rapid expansion of publicly accessible generative AI tools, we have observed a growing volume of AI-generated content on social media platforms, including content designed to mislead users. Several studies have documented a rise in such material, which, owing to its capacity for virality and low production cost, has become an effective means of attracting large numbers of views and engagements[e.g. 17]. The flooding of social media with synthetic content is often referred to colloquially as "AI slop"[18]. Some platforms incentivise this phenomenon through creator programmes that reward accounts producing highly viewed content, regardless of its quality or authenticity.

While much of this content portrays humorous or absurd scenarios, a growing body of evidence suggests that AI-generated material can also pose significant risks to civic discourse. These risks include AI-generated personas impersonating medical professionals[19-20], fabricated textual claims aimed at political manipulation, and other forms of misleading synthetic content[21,22]. As the depiction of events and individuals becomes increasingly realistic, synthetic media is becoming progressively harder for users to distinguish from authentic content.

In light of these developments, and the broader risks associated with deceptive synthetic media, the need to define and systematically measure a Structural Indicator of AI-generated mis/disinformation is becoming increasingly important. Our consortium has therefore decided to replace the indicator on cross-platform aspects of disinformation, which was

measured in the first reporting period, with one focused on AI-generated mis/disinformation. We consider the most critical dimensions to document to be: the proportion of mis/disinformation in our data sample that is AI-generated, and whether such content is labelled as such on each platform.

The section below provides an overview of VLOP policies with respect to what types of AI-generated content are permitted and how platforms handle the labelling of such content.

2.4.1 Platform Policies

At present, no specific regulatory provision explicitly requires platforms to systematically label AI-generated content. However, within their own terms and conditions, particularly under authenticity, integrity, or transparency policies, all six VLOPs state that synthetic or manipulated content is prohibited where it is intended to deceive users or cause public harm. In practice, this means that AI-generated or significantly altered content is not banned outright, but becomes a policy violation when used in a misleading manner.

Additionally, Meta, YouTube, and TikTok have adopted specific policies on AI-generated content, either requiring users to label such material or stipulating that their own detection tools will do so. As of the time of writing, no public documentation provides detailed information about the accuracy of these detection systems or the prominence of labels as they appear in users' feeds.

Examining each platform's approach in turn reveals important differences in scope and enforcement.

Under Meta's policy[23,24], content identified as AI-generated may receive a label called *AI Info*. This label can be applied in two ways: when users voluntarily disclose that their content was created using AI tools, or when Meta's own systems detect it as such. Critically, the *AI Info* label does not surface automatically in a user's feed: accessing it requires navigating a post's additional menu, an extra step that most users scrolling through content are unlikely to take. These limitations have not gone unnoticed: in the context of the ongoing conflict between Israel and Iran, Meta's Oversight Board called on the platform to develop more reliable and visible mechanisms for identifying AI-generated content[25].

TikTok's Community Guidelines take a more prescriptive approach. Creators are required to label content that is fully generated or significantly edited using AI, particularly when it includes realistic images, audio, or video. Unlike Meta's *AI Info*, this label appears directly on the video itself, where it is clearly visible to users. In addition to this disclosure requirement, TikTok states that it may automatically apply AI-generated labels to content it identifies as synthetic. The platform also explicitly prohibits misleading AI-generated content that

misrepresents real events or individuals, linking AI transparency directly to its broader integrity and misinformation policies[26,27].

YouTube's approach, like Meta's, is limited in terms of label visibility: the disclosure appears in the video description rather than on the video player itself, requiring an additional step from the user. A notice is displayed when a video is created using YouTube's own generative AI tools, when the user manually discloses AI use, or when valid content credentials metadata indicates that the video is entirely AI-generated. While YouTube references content credentials as a signalling mechanism, its public policies do not indicate that the platform systematically detects and labels AI-generated content independently of user disclosure or embedded metadata. Like TikTok, YouTube explicitly prohibits manipulated content that misrepresents real events or individuals under its Misinformation Policies[28].

X/Twitter, by contrast, lacks a dedicated policy specifically addressing AI-generated content. Synthetic or manipulated media instead fall under the platform's Authenticity Policy, which prohibits deceptive media likely to cause public harm. AI-generated content that does not directly result in public harm appears to be permitted, and there is no explicit requirement to label or otherwise disclose that content was generated using AI tools[29].

LinkedIn similarly does not maintain a standalone policy on AI-generated content. Synthetic or manipulated material is addressed within broader rules prohibiting misleading or deceptive conduct. As with X/Twitter, there is no clear obligation to label AI-generated content as such, unless it violates general platform standards relating to misinformation or authenticity[30].

2.4.2 Proportion of AI-Generated Mis/Disinformation

Building on the content labelled by fact-checkers in Section 2.1.1 C, we introduced an additional annotation layer to assess the role of generative AI in the misinformation ecosystem. For each piece of content identified as mis/disinformation, fact-checkers were asked to determine: (i) whether the content included AI-generated elements, such as synthetically generated images or videos; and (ii) whether the platform or the content creator had clearly labelled it as AI-generated. This approach allows us to estimate the prevalence of AI-generated mis/disinformation within the broader set of misleading content identified in our sample.

While our sample contains a mix of post types (text-based as well as image and video content) our analysis of AI-generated content focuses specifically on images and videos. Text-based AI-generated content is considerably harder to identify with the naked eye than visual content, making images and videos more suitable for this type of assessment. It is

important to note that purely text-based mis/disinformation is not excluded from our overall assessment: the proportion of AI-generated content is calculated relative to the total number of mis/disinformation posts on each platform, regardless of format.

Figure 2.11 shows that TikTok (24%) and YouTube (19%) exhibit the highest proportions of AI-generated content within the mis/disinformation sample. This finding is likely explained in part by the fact that both platforms are exclusively video-based. Facebook follows at 7%, while X/Twitter (4.4%) and Instagram (2.6%) display lower proportions of AI-generated misleading content. The platforms rely more heavily on text-based or mixed media formats, which may partly account for these lower figures given that only visual AI-generated content was detectable through our method. LinkedIn showed no instances of AI-generated mis/disinformation within our sample.

Of all pieces of mis/disinformation identified as containing AI-generated elements, only 16.5% were visibly labelled as synthetic or AI-generated by either the platform or the content creator. The platform-level breakdown reveals striking disparities. On TikTok, 14% of AI-generated mis/disinformation posts carried a label. On Facebook, this proportion dropped to just 1.8%, and on YouTube to 0.9%. No labels were observed on AI-generated mis/disinformation on any of the remaining platforms in our sample.

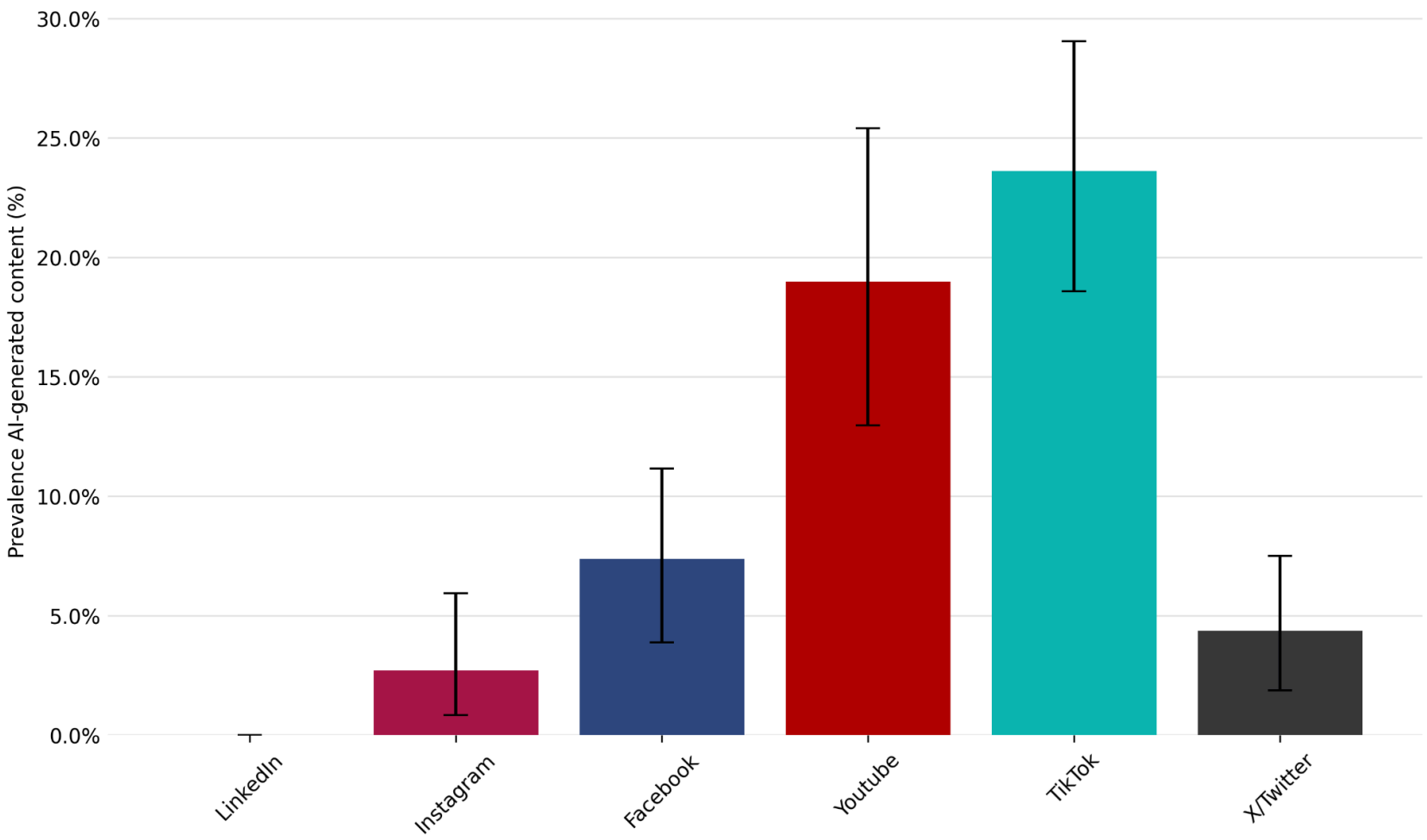
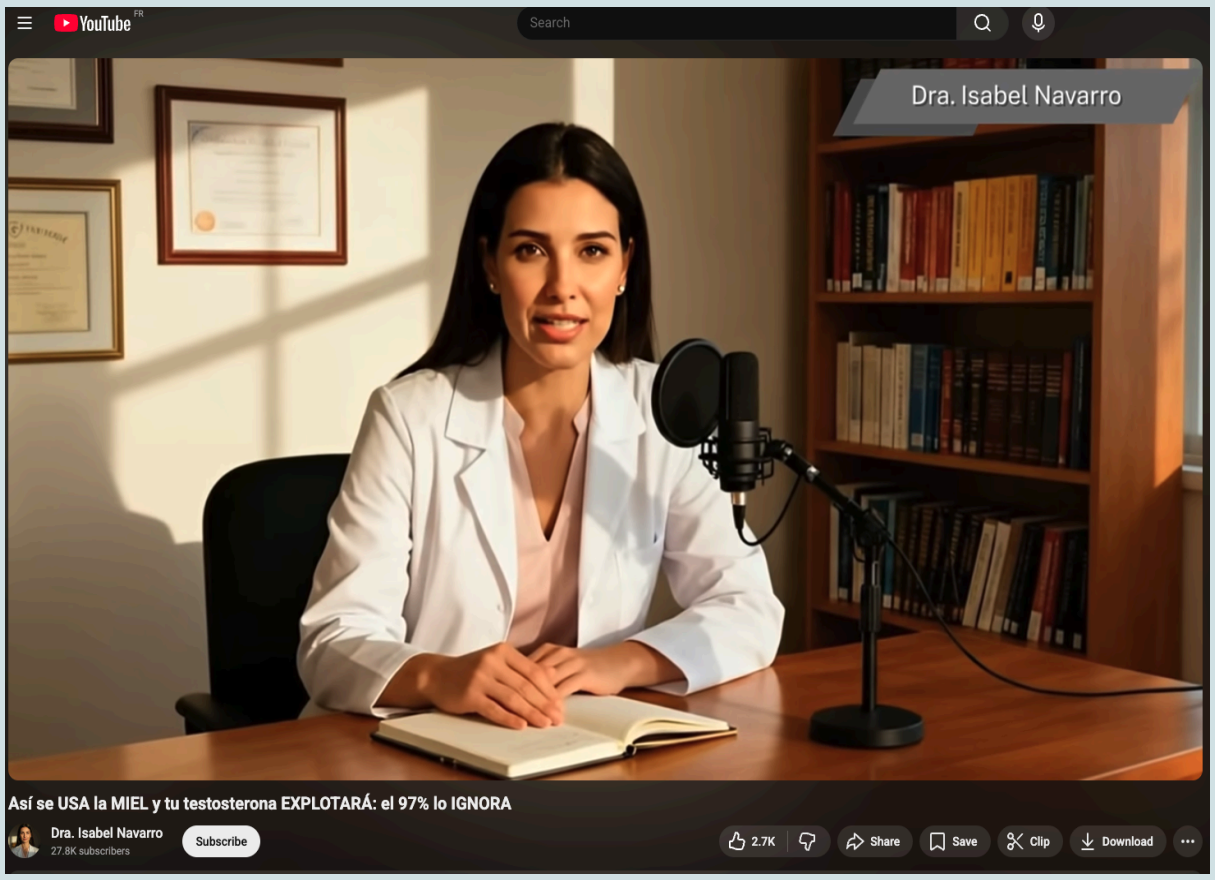


Figure 2.11 – Proportion of content containing AI-generated images or videos within the mis/disinformation sample across the six very large platforms, aggregated across all languages. The error bars represent the 95% confidence intervals measuring the uncertainty around each estimate, calculated using a bootstrapping method (see Appendix 5.1.4).

Examining the thematic distribution of AI-generated mis/disinformation reveals a strong concentration in the health domain, particularly on the platforms where such content is most prevalent. Figure 2.12 presents an illustrative example drawn from our sample of AI-generated misinformation. On YouTube and TikTok, health-related content accounts for 61% and 57% of AI-generated mis/disinformation respectively. The remaining thematic distribution differs between the two platforms: on TikTok, migration and the Russia–Ukraine war account for most of the non-health AI-generated mis/disinformation, while on YouTube, migration and national politics are more prominent.

In terms of reach, AI-generated mis/disinformation in our sample accumulated approximately 34 million views, with TikTok accounting for 69% of these, followed by YouTube at 23%.



Illustrative example drawn from our sample. The case involves a YouTube video featuring a self-proclaimed "Dr. Isabela Navarro" (an AI-generated avatar) claiming that honey can increase testosterone levels in humans. While some studies have explored this effect in rats, no solid evidence supports such claims in humans[31].

At the time of analysis, the channel, which is monetised and posts exclusively AI-generated videos featuring the same fictitious doctor, had **27 800 subscribers**, and the video had accumulated **61 700 views**.

No AI label indicated that "Dr. Navarro" was a synthetic persona, illustrating both the failure of YouTube's detection and labelling mechanisms and the ease with which AI-generated content can fabricate an appearance of medical authority, reinforced here by visual cues such as a white coat and fake diplomas.

Figure 2.12 – An AI-generated video from the annotated sample labeled as mis/disinformation, featuring an AI avatar posing as a medical professional.

2.5 AUDIENCE GROWTH

One of the purposes of Structural Indicators is to enable tracking of how the disinformation landscape evolves over time. This section introduces an additional indicator designed to monitor whether the audiences of high- and low-credibility accounts are growing, and whether their growth rates differ. This should shed light on whether platform dynamics are systematically favouring accounts that repeatedly share mis/disinformation, or the reverse.

2.5.1 Methodology

To monitor the evolution of high- and low-credibility accounts over time, we established a tracking system building on the dataset compiled for the first SIMODS report[7]. Specifically, we monitored a set of accounts comprising: (i) the Top 50 accounts identified from content collected during the first measurement period, and (ii) the Fact-checkers' list, which remained unchanged between the first and second reporting waves. For this set of accounts, we tracked changes in audience size and engagement per post between the two data collection periods.

A. FOLLOWERSHIP GROWTH

Due to data collection constraints, follower counts were not available at identical points in time for all accounts and platforms across the monitoring period (March 2025 to February 2026): for some accounts the earliest available measurement dates from March 2025, while for others it is June 2025. To account for this, relative follower growth ΔF was calculated using the first (F_i) and last (F_f) available data points for each account, normalised by the number of months elapsed between the two measurements (N_m), according to the following formula:

$$\Delta F = \frac{F_f - F_i}{F_i} \times \frac{100}{N_m}$$

This yields an average monthly follower growth rate for each account, which is then averaged across all accounts within the low-credibility and high-credibility groups respectively.

B. ENGAGEMENT GROWTH

Another metric we calculated from the aforementioned dataset is the average change in engagement per account. Given the challenges inherent in collecting data at scale across a large number of accounts, it was not possible to retrieve all posts published by every account in the sample.

To calculate the average change in engagement per account, posts were split into two periods: Period 1 (March–June 2025) and Period 2 (September 2025–February 2026). To mitigate the influence of outliers, only accounts with a minimum of five posts in each period were retained. As a result, the number of accounts used to measure this metric may differ from those used in the Audience Growth analysis.

Engagement growth is defined as the difference between the average number of engagements per post per account in Period 1 and the average number of engagements per post per account in Period 2.

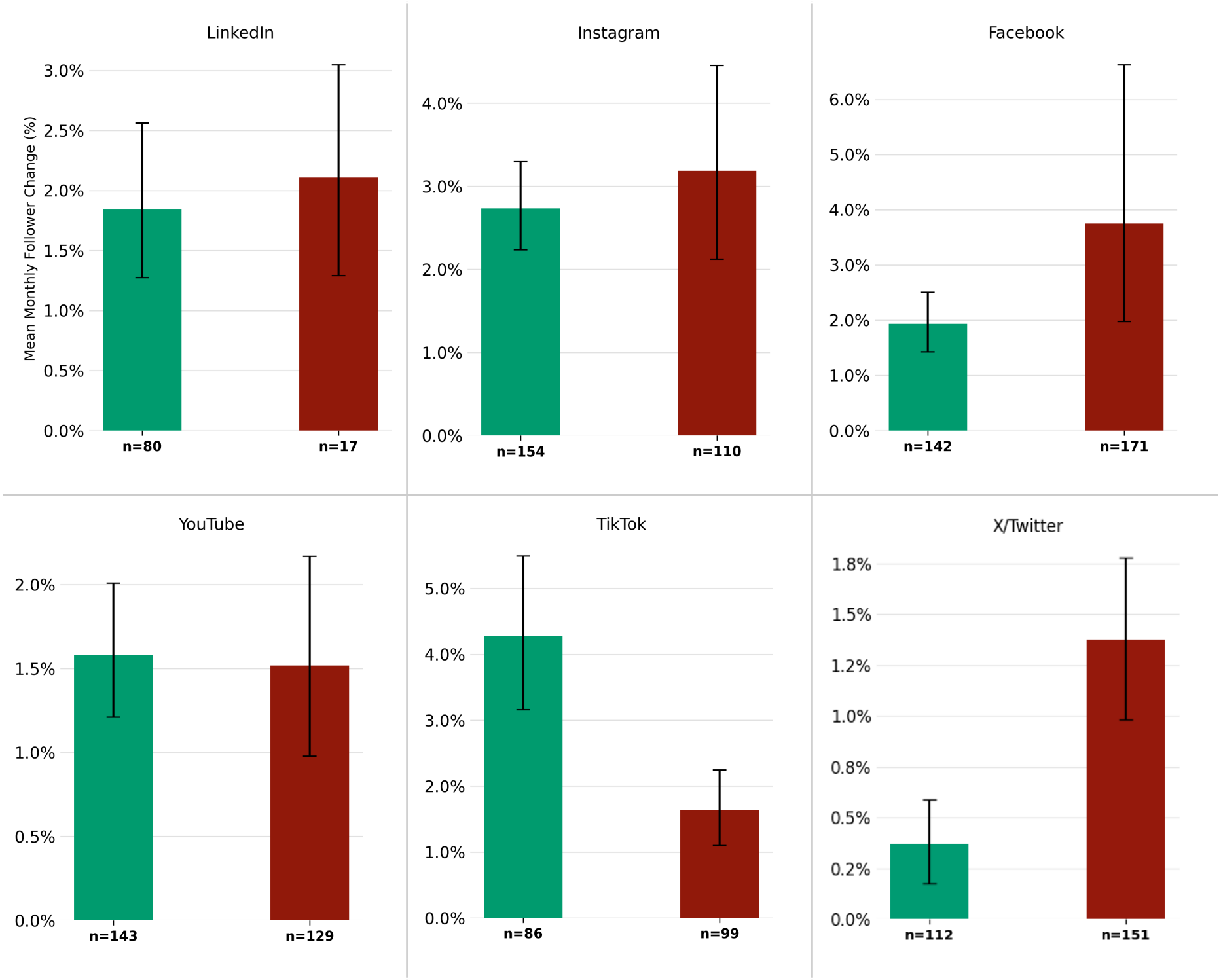


Figure 2.13 – Average monthly follower growth rate for low- and high-credibility accounts between the first and second measurement periods. The number of accounts serving as the basis for the calculation is indicated at the bottom of each bar. The error bars represent the 95% confidence intervals measuring the uncertainty around each estimate, calculated using a bootstrapping method (see Appendix 5.1.4).

2.5.2 Results

As shown in Figure 2.13, across all platforms except TikTok and X/Twitter, there is no statistically significant difference in follower growth between high- and low-credibility accounts: both groups are gaining followers at broadly comparable rates: approximately 2% average monthly growth on LinkedIn, 3% on Instagram, and 1.5% on YouTube. On

Facebook, growth rates are also not statistically distinguishable between the two groups, though this is largely attributable to the high variability in growth rates among low-credibility accounts; a larger sample of such accounts would be needed to reduce uncertainty and potentially confirm the higher growth rate suggested by the point estimate.

On X/Twitter, low-credibility accounts experienced significantly faster audience growth, with a monthly increase of approximately 1.4%, around 3.5 times the rate observed for high-credibility accounts (0.4%). TikTok shows the reverse pattern: high-credibility accounts grew at an average of 4.3% per month, compared to just 1.6% for low-credibility accounts. Taken together, these results suggest that X/Twitter is the platform where low-credibility accounts are gaining ground most rapidly relative to their high-credibility counterparts.

As for the evolution of engagement per post per follower, our results show that there are no statistical differences between high- and low-credibility accounts that have been tracked continuously between the two periods.

Platform	High-credibility	Low-credibility
LinkedIn	**1.8** [1.3, 2.6]	**2.1** [1.3, 3.1]
Instagram	**2.7** [2.2, 3.3]	**3.2** [2.1, 4.5]
Facebook	**1.9** [1.4, 2.5]	**3.8** [2.0, 6.6]
Youtube	**1.6** [1.2, 2.0]	**1.5** [1.0, 2.2]
TikTok	**4.3** [3.2, 5.5]	**1.6** [1.1, 2.2]
X/Twitter	**0.4** [0.2, 0.6]	**1.4** [1.0, 1.8]

Table 2.6 – Average monthly follower growth rate for low- and high-credibility accounts between the first and second measurement periods, as displayed in Figure 2.13. 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).

3. Recommendations

The findings presented in this report, now spanning two consecutive measurement periods, reveal persistent structural patterns in the presence and amplification of mis/disinformation on large online platforms in Europe. The consistency of results across time reinforces their credibility and sharpens the case for action. The following recommendations are directed at three distinct audiences: the European Commission and DSA enforcement bodies, the platforms themselves, and funders of independent research and monitoring.

TO THE EUROPEAN COMMISSION AND DSA ENFORCEMENT BODIES

The integration of the Code of Conduct on Disinformation into the DSA framework, effective July 2025, marks a significant step. The evidence presented in this and the first SIMODS report now provides regulators with independent, comparable baselines against which platform compliance can be assessed. We recommend the following actions.

1) Enforce meaningful researcher data access.

Despite the provisions of DSA Article 40, only one out of six platforms studied supplied the random sample requested under Article 40.12. The European Commission should treat non-compliance with researcher data access requests as an enforcement priority, not a secondary concern. Without reliable access to platform data, independent auditing of systemic risks, as explicitly foreseen under the DSA, remains impossible.

2) Use Structural Indicators as formal compliance benchmarks.

The metrics developed by the SIMODS project, including the prevalence of mis/disinformation and the misinformation premium, were explicitly designed to be comparable across platforms and stable over time. With two measurement waves now completed, these indicators are mature enough to serve as formal benchmarks in DSA auditing and compliance assessments. We encourage the European Commission to integrate them, or equivalent methodologies, into its enforcement framework.

3) Address the monetisation of disinformation through binding disclosure obligations.

Meaningful measurement of whether platforms are honouring their demonetisation commitments is currently not possible for the VLOPs studied, due to the opacity of their

monetisation systems. Given that financial incentives are a core driver of disinformation production and amplification, the Commission should require all VLOPs to disclose monetisation status at the account level to accredited researchers, as part of their obligations under the DSA.

4) Mandate consistent labelling of AI-generated content across all platforms.

Our findings show that a significant and growing share of mis/disinformation posts are AI-generated, and that the vast majority carry no label. Platform policies on AI-generated content disclosure vary substantially: some require creator disclosure, others apply automated detection, and others have no dedicated policy at all. This inconsistency leaves users without the basic contextual information needed to evaluate what they are seeing. The European Commission should establish a harmonised, enforceable labelling standard across all VLOPs. This is not merely a technical measure; it is a necessary step in accompanying the public through a period of rapid change in the nature of online content, helping citizens develop the literacy and habits needed to navigate this new environment.

TO THE PLATFORMS (VLOPS)

The data shows that across all platforms studied, low-credibility accounts consistently receive disproportionate engagement relative to their audience size, a pattern we term the "misinformation premium". This premium is an indicator of algorithmic amplification of low-quality and misleading content, and its persistence across two measurement periods suggests it is structural rather than incidental. We recommend the following actions.

1) Reform recommendation and amplification systems to reduce the misinformation premium.

On most platforms studied, a post from a low-credibility account generates substantially more interactions per follower than one from a credible source. Platforms should audit their recommendation algorithms and content distribution systems to identify and correct the mechanisms that produce this effect, and report on their findings transparently.

2) Honor demonetisation commitments in practice, and address the incentivisation of AI-generated slop.

Our analysis, where data permitted, shows that low-credibility accounts can access monetisation programmes at rates that are either comparable to, or only modestly lower than, those of high-credibility sources. Platforms should implement content-quality checks as part of monetisation eligibility criteria, and establish robust mechanisms to remove monetisation access from accounts that repeatedly share misleading content. This concern is made more acute by the rapid proliferation of AI-generated low-quality content: several platforms currently remunerate producers of such content through their creator programmes, thereby actively incentivising its creation. Platforms should extend their demonetisation policies to explicitly cover AI-generated content that serves no informational value and is produced primarily to exploit algorithmic amplification.

3) Implement consistent, platform-wide labelling of AI-generated content.

Several platforms have adopted policies on AI-generated content disclosure, but implementation is inconsistent and coverage is partial. Platforms should ensure that labelling is applied systematically, including through automated detection, and not only when creators voluntarily disclose it. This is a basic requirement for users' right to be informed.

4) Provide standardised datasets to the research community.

As called for in the first SIMODS report, platforms should work with regulators and researchers to define and publish standardised datasets that enable independent, reproducible, cross-platform measurement of Structural Indicators. Ad hoc API access, subject to platform-determined rate limits and scope restrictions, is not a substitute for structured data sharing. These datasets should include, at minimum, content-level engagement metadata and account-level monetisation status for researchers accredited under DSA Article 40 procedures.

TO FUNDERS OF INDEPENDENT RESEARCH AND MONITORING

The robustness of the findings presented in this and the first SIMODS report, including the consistency of results across two measurement periods, demonstrates that the methodology is sufficiently mature to underpin a sustained, long-term monitoring programme. A single report, or even two, captures a snapshot; what regulators and the

public need is a longitudinal tracking system capable of detecting deterioration or progress over time.

We therefore call on funders, including the European Commission, national governments, and philanthropic foundations, to support the **establishment of a permanent, independent monitoring panel for online disinformation in Europe**, built on the Structural Indicators framework developed by EDMO and operationalised by the SIMODS project. Such a panel would allow for the timely detection of emerging trends, provide a consistent evidence base for regulatory action, and ensure that enforcement decisions are grounded in rigorous, comparable data rather than platform self-reporting.

At the national level, we also encourage authorities to consider how monitoring and fact-checking capacity can be embedded into broader information integrity strategies. Given that health misinformation constitutes the largest single category of misleading content identified in our study, and is particularly concentrated on certain platforms, we **encourage Member States to consider developing national strategies to protect their populations from the harms arising from digital platforms**. France's *Stratégie nationale de lutte contre la désinformation en santé*, launched in January 2026 by the Ministry of Health, could serve as a source of inspiration and be adapted to local contexts across different countries.

Investing in media literacy programmes targeted at the platforms and topics where problematic content is most prevalent remains equally essential. Fact-checking and platform enforcement address the supply side; media literacy addresses the demand side, building citizens' capacity to critically evaluate the information they encounter and reducing the audience susceptibility on which disinformation ecosystems depend.

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, TikTok and YouTube 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 (2026) Second Measurement of the State of Online Disinformation in Europe on Very Large Online Platforms. Second report of the SIMODS project (Structural Indicators to Monitor Online Disinformation Scientifically). Science Feedback.

5. Appendices

5.1 METHODOLOGY FOR THE PREVALENCE INDICATOR

5.1.1 Data collection

To collect data that reflects the diverse types 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 national 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 to strike 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 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 in the analysis.

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 to request a random sample of 200 000 posts per language, efforts that started with the first iteration of the SIMODS Report on 19 December 2024.

Following our initial outreach, we sent several follow-up reminders to all platforms, the latest being on 7 October 2025, but as of the writing of this report, we have not received a response.

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 and a second application for the second iteration of the project, but as of February 2026, we had not received a reply.

Out of the six platforms, only LinkedIn provided the requested dataset, which 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 for each category of ‘personal’ and ‘business’ accounts (so up to 400 000 posts per language in total). The exact number varied depending on data availability in each respective country.

TikTok and YouTube granted access to their research APIs, which enabled data collection through keyword-based queries. However, neither API allows for the extraction of a platform-generated random sample. Instead, data retrieval is limited to content identified via keyword searches or filtered by publication date.

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 that 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 interaction 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 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 interaction metrics and view counts.

C. YOUTUBE

Data collection for Youtube was conducted using two methods, leveraging the methodologies of our consortium partners and the YouTube API access which we received in time for the second data collection.

Firstly, CheckFirst collected data daily, using their 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. While content from other platforms was restricted to posts published during the data collection period (1 and 31 October 2025), for YouTube we filtered content by publication date during post-processing to ensure that only posts from 1 July 2025 to 31 October 2025 were included in the analysis. This decision reflects the longer content lifespan and engagement cycle of YouTube videos compared to other platforms.

In addition to the primary search results, CheckFirst's 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.

Secondly, access to YouTube's API enabled us to retrieve content based on keyword queries within the defined time window (1–31 October 2025), consistent with the approach used for other platforms. We collected this data retroactively, allowing for the content to reach its audience, and capture an accurate representation of the engagement these videos received on the platform. The queries were geolocated to the four countries included in our analysis, meaning only videos that can be viewed by people located in these countries are included. The datasets obtained through keyword-based retrieval were then merged, duplicates were removed, and a consolidated dataset was constructed for analysis.

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 from 1 July to 31 October 2025 was included in the final dataset, matching our approach with YouTube.

To enhance robustness and broaden coverage, we complemented this approach with data collected via the TikTok Researcher API. This allowed us to retrieve keyword-based content published between 1 and 31 October 2025, filtered by their availability from the four countries included in our analysis. The dataset obtained through the TikTok API was merged with the data collected by CheckFirst, duplicates were removed, and the final consolidated dataset was analysed.

5.1.2 Data Processing & Sampling

A. DATA CLEANING & PROCESSING

The resulting dataset from the data collection period, spanning 1 to 31 October, comprised a total of 3.3M posts across all platforms and target languages, after duplicates were

removed. For YouTube, posts collected by CheckFirst published between 1 July and 31 October 2025 were retained, to compensate for the more limited amount of data on this platform, 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.

Across the full dataset of 3.3 million posts, spanning six platforms and four languages, the combined total reached 18 billion views (Table 5.1), reflecting the overall reach of the content analysed.

Platform	# Views (entire dataset collected)	# Posts (entire dataset collected)
LinkedIn	4 713 686 000	1 275 120
Instagram	3 354 197 000	142 517
Facebook	1 965 926 000	250 643
YouTube	4 364 608 000	59 077
TikTok	2 486 712 000	106 722
X/Twitter	725 119 000	1 457 705

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. Various prompts were tested during the first iteration of the SIMODS project. Based on those results, we retained the same prompt and model configuration for the present measurement period to ensure methodological consistency and comparability over time.

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 labelled as *Irrelevant* or *Geographically Irrelevant* was 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 annotated sample, thus reflecting what users are actually seeing on the platform. As shown in Table 5.2, the resulting sample accounts for a total of 1.9 billion views, with 604 million views coming from TikTok alone. Second, it helps mitigate potential biases introduced by platform-specific search algorithms, which may be influenced by personalisation 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.

5.1.3 Annotation

To improve the precision of our confidence intervals relative to the first report, we sampled 600 pieces of content per platform and per language, of which only those not flagged as

"Irrelevant" counted towards the 500. This allowed us to increase the number of posts that are deemed "relevant", and thus taken into account in the calculation of prevalence. Each sample 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.

Platform	Total number of views in the random sample
Facebook	280 839 000
Instagram	396 464 000
YouTube	514 444 000
X/Twitter	71 820 000
TikTok	604 264 000
LinkedIn	43 667 000

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

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, a cross-verification process was implemented. For each platform and language, a first fact-checker was responsible for annotating the full sample of posts, while a second fact-checker independently reviewed a randomly selected 20% subset of the same sample. Both fact-checkers worked 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 labelled, 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 Appendix 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 labelled 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 falls 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 labelled *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 centred on events in Latin America or Francophone Africa, for instance, 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 labelled 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 labelled 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 piece of 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 categorisation 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 85%. A notable 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* 39 times, while the second rated it as *Unverifiable*, and the second fact-checker rated content as *Credible* 99 times while the first rated it as *Unverifiable*.

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

Once the data sample was fully labelled, 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 labelled 68 pieces of content as *Abusive*. The final label agreed in 56 cases (82%), 6 were relabeled *Unverifiable* (9%), 5 *Borderline* (7%), and 1 *Credible* (1.5%). In the bootstrapping process, at each iteration, we therefore randomly swapped the *Abusive* label with *Borderline* with probability 7%, with *Unverifiable* with probability 9%, with *Credible* with probability 1.5% and left it unchanged with probability 82% (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* content 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. The highest levels of *Problematic* content as compared to *Credible* content are found on TikTok, YouTube and X/Twitter, where it represents more than two-thirds (64%), 56% and a bit more than half (54%) respectively.

B. PREVALENCE OF MIS/DISINFORMATION + BORDERLINE

As we have explained above, factually accurate content 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 labelled as *Misinformation* or *Borderline*, relative to the total number of posts labelled 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 40.5% of posts falling into these categories. It is followed by Facebook at 29% and YouTube at 26%. LinkedIn displays the lowest prevalence of such content at 4%, indicating a significantly more limited presence of misleading content on this platform.

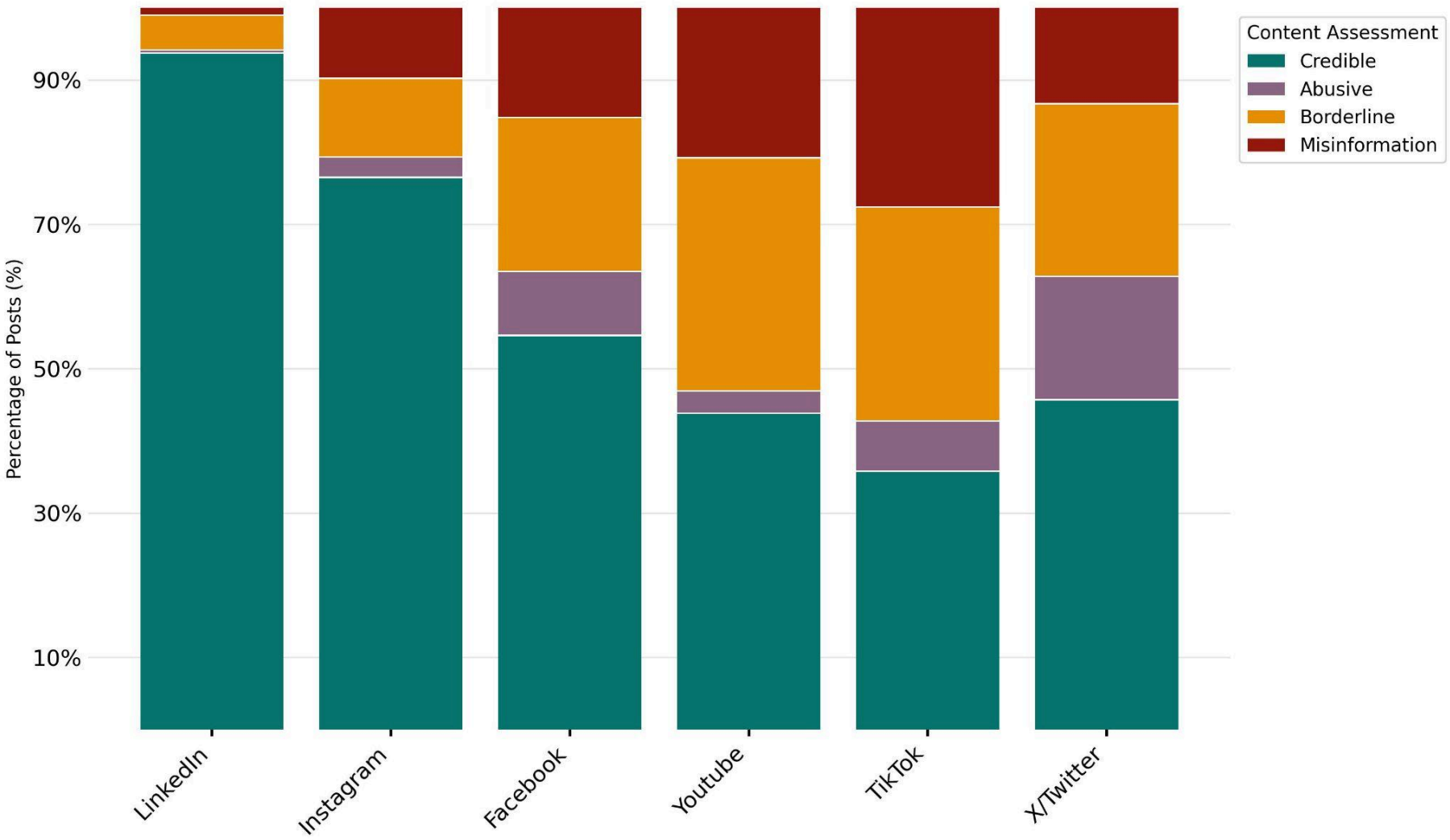


Figure 5.2 – Percentage of posts belonging to each category for each of the six very large online platforms, same as Figure 2.2 but excluding *Unverifiable*.

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, prevalence is not statistically different across four platforms, with values ranging from ~8% to ~12% on all platforms except LinkedIn and X/Twitter. Spain reveals a similar pattern with no statistical differences between several platforms; TikTok appears to be an exception with a prevalence of ~17%.

By contrast, France shows more contrasted results: TikTok has the highest prevalence (~39%), followed by X/Twitter and Facebook (~21%). In Slovakia, TikTok, Facebook and

YouTube exhibit the highest prevalence (~17% to ~25%), contrasting with the situation on X/Twitter, Instagram and LinkedIn where prevalence is markedly lower.

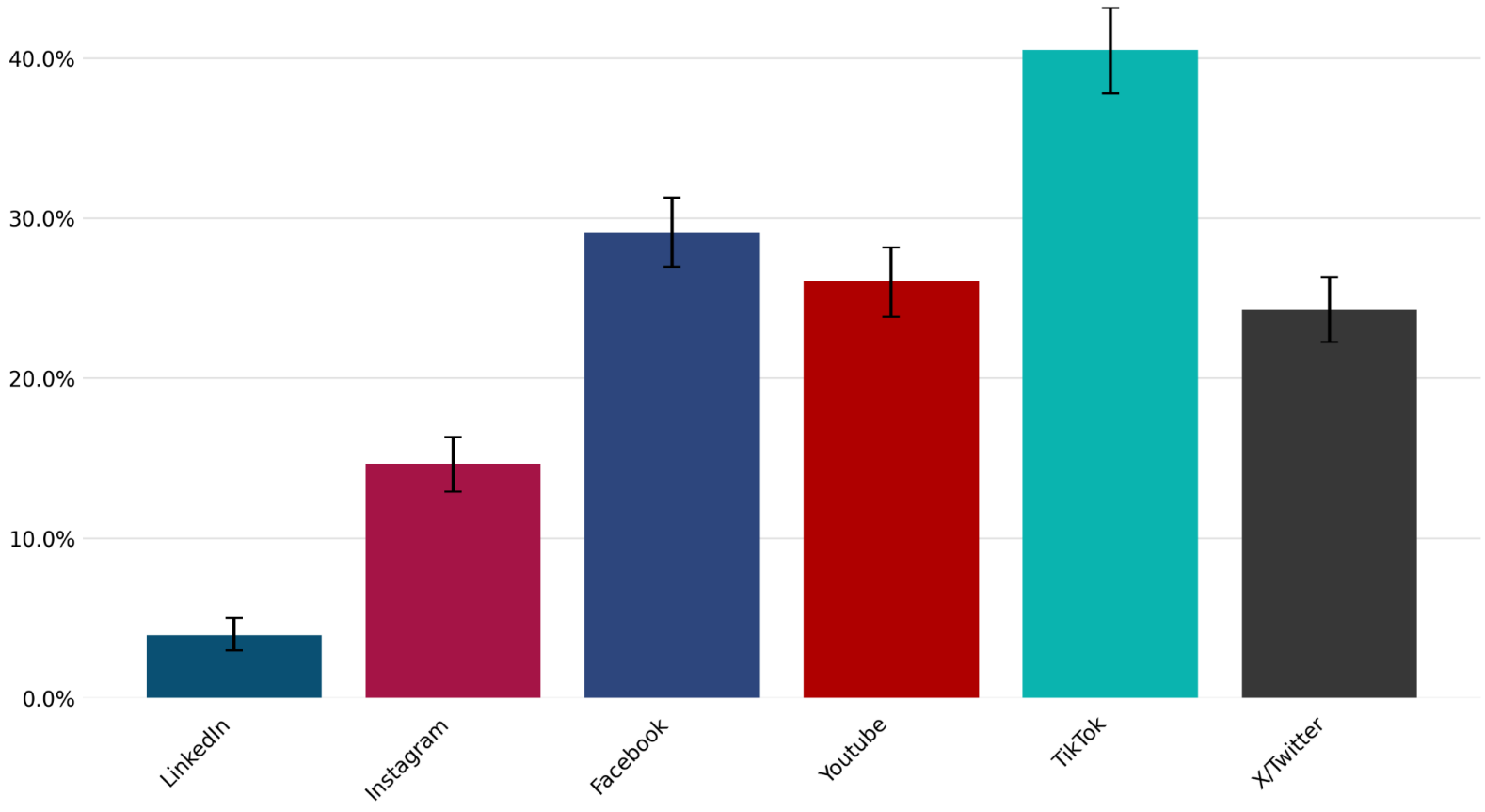


Figure 5.3 – Prevalence of misinformation and borderline content across the six very large platforms, aggregated across all languages. The error bars represent the 95% confidence intervals measuring the uncertainty around each estimate, calculated using a bootstrapping method (see Appendix 5.1.4).

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[16].

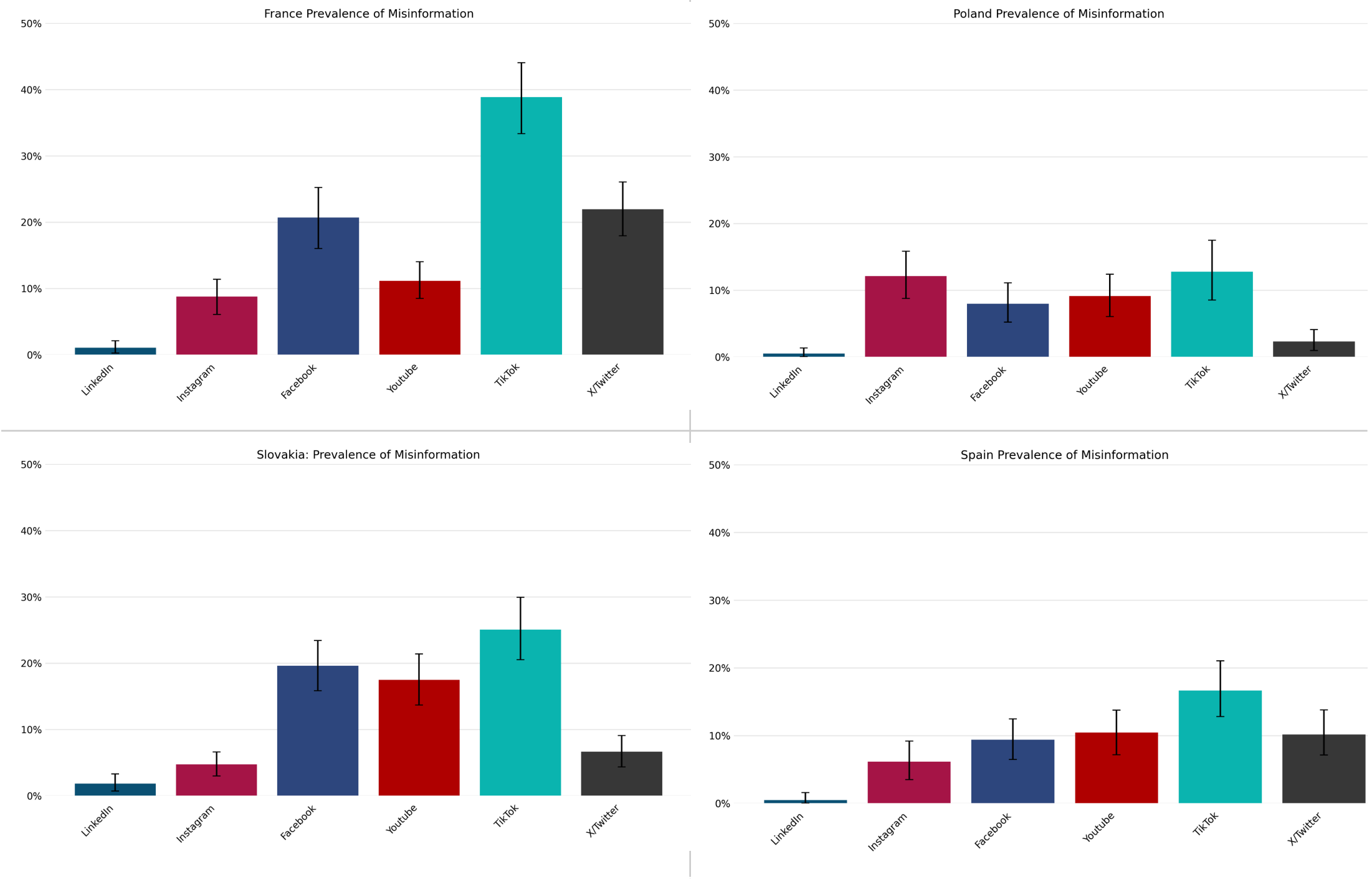


Figure 5.4 – Misinformation prevalence for each country across the six very large platforms per language. The error bars represent the 95% confidence intervals measuring the uncertainty around each estimate, calculated using a bootstrapping method (see Appendix 5.1.4).

B. THE TOP 50 LIST APPROACH

In another approach, we leveraged the keyword-based dataset used in the Prevalence Indicator (Section 2.1) 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 (1 October to 31 October 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.

C. POSTS COLLECTION

For the Top 50 list, we collected all posts published during the data collection period. X/Twitter constituted an exception, as its interface does not allow retrieval of posts sufficiently far back in time. Instead of collecting data for October 2025, we collected posts from January 2026 as a proxy window. For the fact-checkers' recommendation list, identical to that used in the first SIMODS report, we monitored and collected posts and associated metadata from October 2025 through January 2026 (inclusive). Data collection relied on third-party tools (Bright Data, Apify) as well as the official APIs of YouTube and TikTok. We retrieved both post content and associated metadata, including engagement metrics, follower counts, and available media (images and videos). Due to platform constraints, view counts per post were only available for YouTube, TikTok, and X/Twitter.

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 labelled 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 content.

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.

Fact-checkers reviewed these accounts and their posts and assigned credibility labels based on the definitions outlined above. An account was classified as low-credibility if it contained two or more posts identified as *Misinformation*. An account was classified as *Credible* if at least 95% of its posts consisted of trustworthy, informative or scientific content. Accounts that met neither threshold were categorised as *Neither*.

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 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. 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, there were no misinformation accounts in our Top 50 dataset.

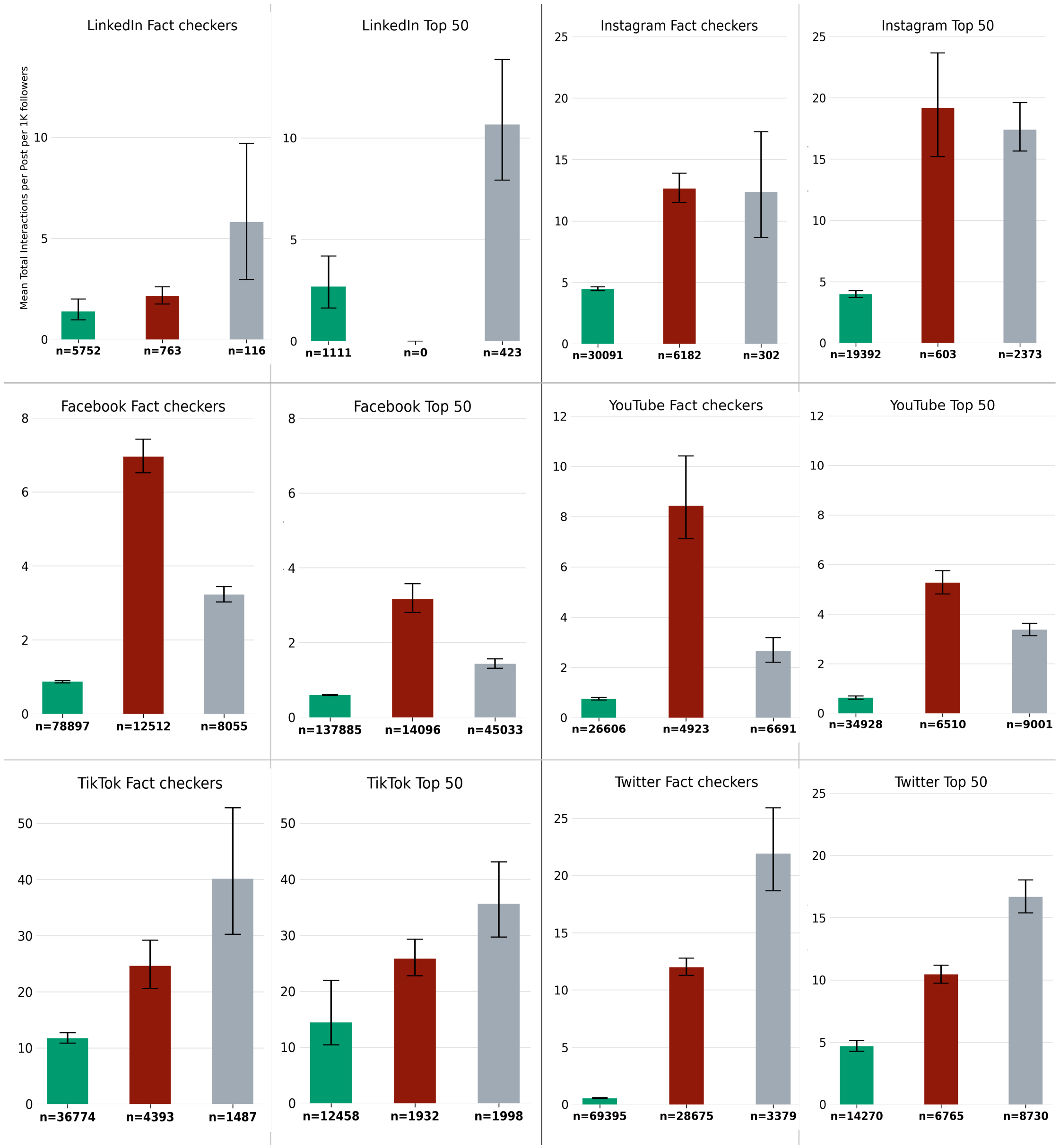


Figure 5.5 – Average number of interactions per post per 1 000 followers for accounts classified as High-credibility and Low-credibility on each platform in the two different sources datasets (the Top 50 and the Fact-checkers' list). Error bars represent 95% confidence intervals (see Appendix 5.1.4).

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. In this section, 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 5.6 indicates that low-credibility accounts receive more views per post per 1 000 followers than high-credibility accounts across all observable platforms.

On YouTube, low-credibility accounts obtain approximately five times more views per post per 1 000 followers. On TikTok, the ratio is roughly two-to-one. On X/Twitter, low-credibility accounts reach audiences approximately five times larger, relative to their follower base, than 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 premium is 8× on YouTube versus 5× for its views-based equivalent, while on X/Twitter it is 10× versus 5×. On TikTok, the interaction-based and view-based misinformation premiums are the same, at about 2×.

5.2.5 Accounts' Audience Size

To better characterise the typical size of accounts in our dataset, we report the median number of followers, in addition to the mean. As shown in Figure 5.7 on all platforms credible accounts have substantially larger subscriber bases than misinformation accounts when considering the median. For example on YouTube, credible accounts have a median of 500 000 subscribers, compared to 170 000 for misinformation accounts. This pattern is consistent with the comparison based on the average number of followers presented in Figure 2.7(Section 2.2.2 A).

Taken together, these results indicate that credible accounts tend to have larger audiences than misinformation accounts, both in terms of typical account size and overall distribution. Therefore, our results show that the difference is not driven solely by a handful of very large

legacy media outlets, which would especially affect the average, but reflects a broader structural gap in subscriber bases.

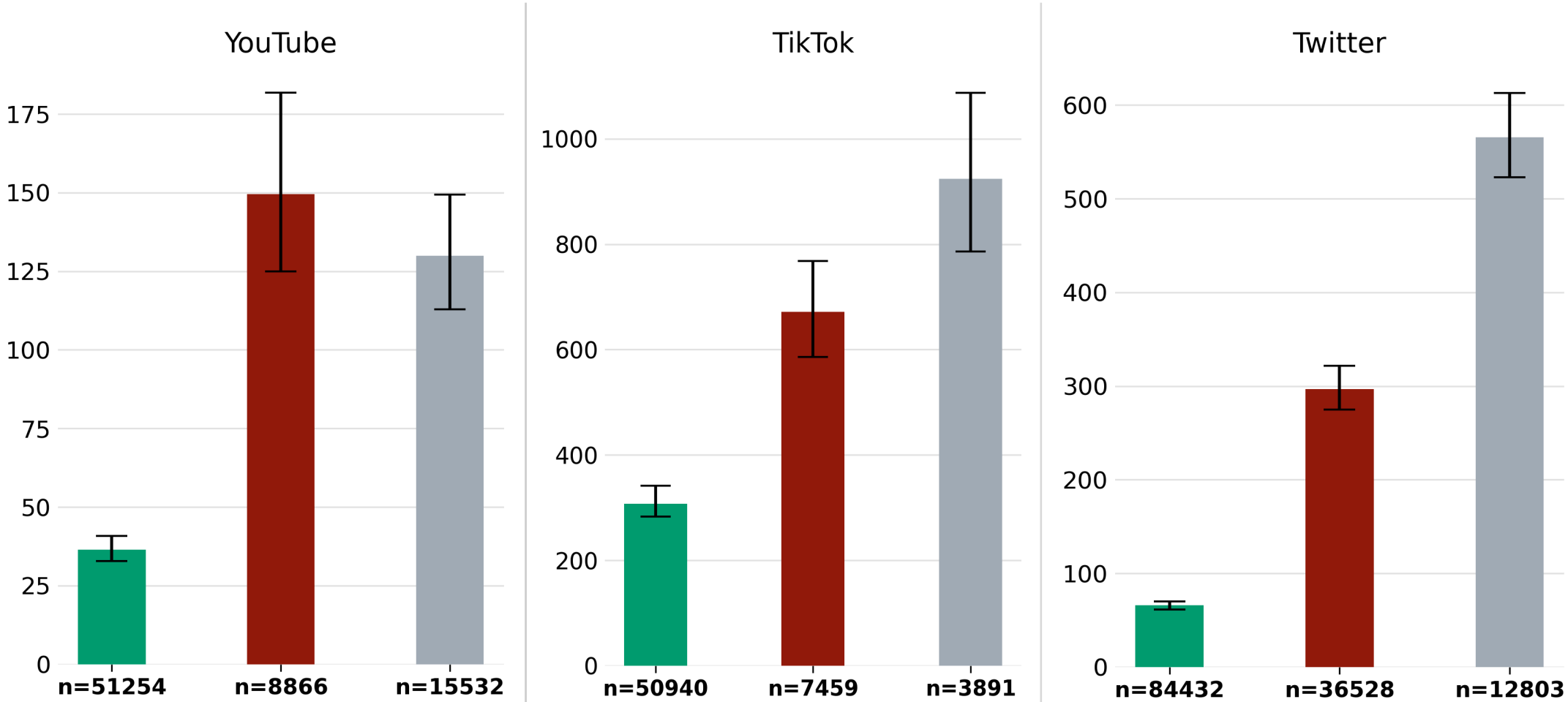


Figure 5.6 – Average number of views per post per 1 000 followers for accounts classified as High-credibility (green), Low-credibility (red), and Neither (grey) on the three platforms where the number of views was available. Error bars represent 95% confidence intervals, calculated using a bootstrapping method (see Appendix 5.1.4).

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 TikTok, the quartiles were defined as follows:

- Q1: 0 – 37 000;
- Q2: 37 000 – 130 000;
- Q3: 130 000 – 360 000;
- Q4: 360 000 – 2 600 000.

These thresholds were applied to both low- and high-credibility accounts. High-credibility accounts exceeding 2.6 million followers were excluded from the comparison to ensure comparability. LinkedIn was treated separately, as limited dispersion in follower counts allowed only two groups.

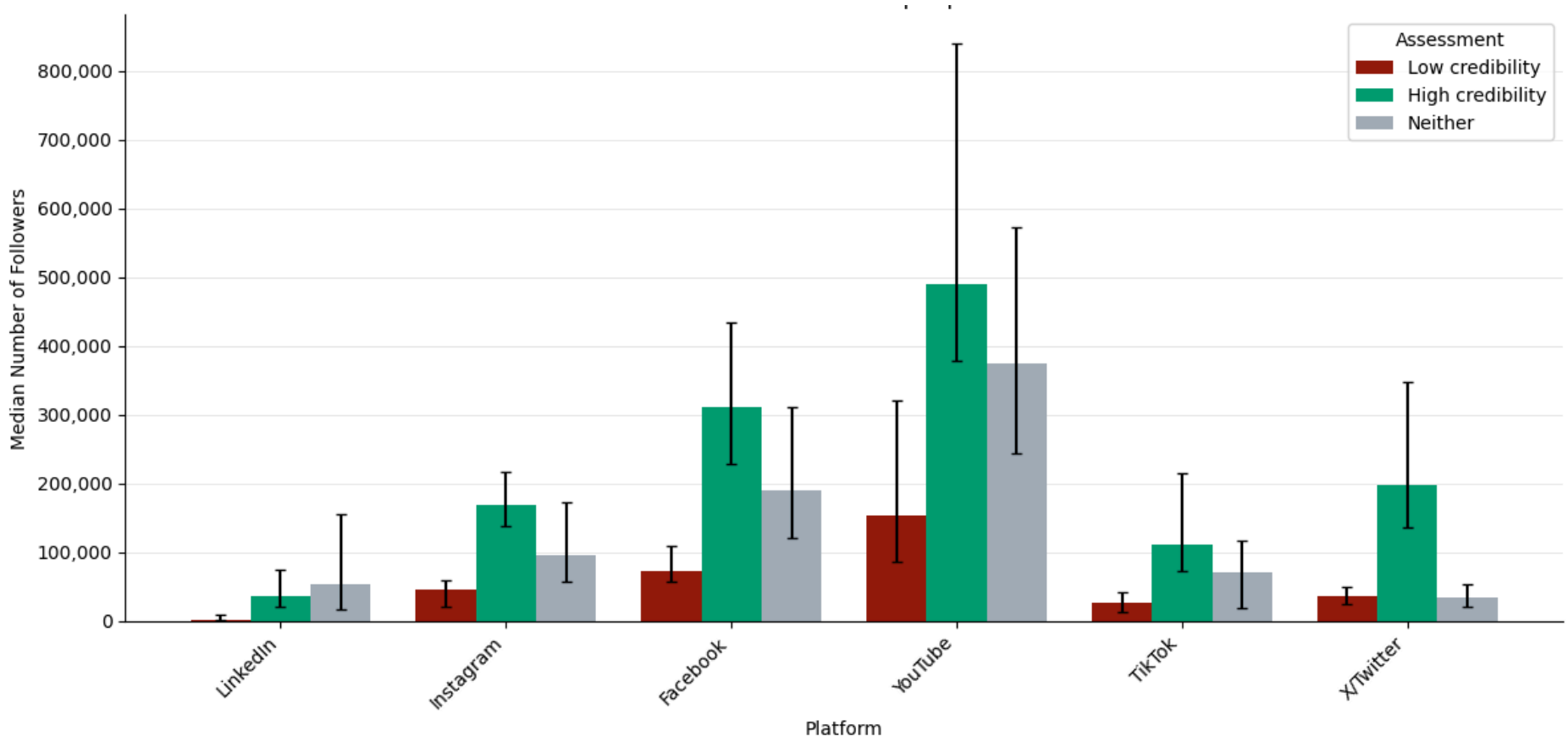


Figure 5.7 Median number of followers for accounts classified as High-credibility, Low-credibility, and Neither on each platform. Error bars represent 95% confidence intervals, calculated using a bootstrapping method (see Appendix 5.1.4).

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 5.3 confirms the pattern observed in Figure 2.8: across nearly all platforms and quartiles, low-credibility accounts generate significantly higher engagement than high-credibility accounts.

Two exceptions emerge. First, on LinkedIn, no statistically significant difference is observed, consistent with earlier findings. Second, on TikTok, low-credibility accounts outperform high-credibility ones only in the fourth quartile, while results are statistically indistinguishable in the first, second and third quartiles. These contrasts explain why the overall misinformation premium on TikTok remains comparatively moderate (approximately 2×).

Platform		Q1	Q2	Q3	Q4
Instagram	High-credibility:	**5.4** [4.3, 6.7]	**4.3** [3.8, 4.8]	**4.9** [4.7, 5.1]	**4.8** [4.5, 5.1]
	Low-credibility:	**24.8** [20.8, 29.2]	**6.4** [5.7, 7.1]	**21.5** [19.1, 24.3]	**7.3** [6.5, 8.1]
Facebook	High-credibility:	**1.5** [1.3, 1.6]	**1.0** [0.96, 1.05]	**0.62** [0.59, 0.64]	**0.42** [0.40, 0.44]
	Low-credibility:	**12.2** [10.9, 13.6]	**5.4** [5.0, 5.8]	**3.5** [3.25, 3.80]	**1.78** [1.67, 1.9]
X/Twitter	High-credibility:	**5.0** [4.2, 5.9]	**2.8** [2.6, 3.0]	**1.9** [1.7, 2.2]	**0.20** [0.19, 0.21]
	Low-credibility:	**14.1** [12.5, 16.0]	**16.2** [14.7, 17.6]	**12.5** [11.8, 13.2]	**5.7** [5.5, 5.9]
YouTube	High-credibility:	**4.6** [4.0, 5.3]	**1.2** [1.1, 1.3]	**0.62** [0.57, 0.68]	**0.24** [0.22, 0.26]
	Low-credibility:	**23.5** [20.7, 27.6]	**7.9** [7.5, 8.3]	**1.5** [1.3, 1.6]	**0.7** [0.6, 0.8]
TikTok	High-credibility:	**46.4** [24.2, 85.9]	**25.6** [19.2, 35.2]	**18.2** [15.6, 20.9]	**9.0** [8.5, 9.4]
	Low-credibility:	**43.3** [31.1, 59.2]	**26.3** [22.0, 30.9]	**17.9** [15.5, 20.6]	**20.0** [17.0, 23.8]
LinkedIn	High-credibility	**7.7** [5.1, 11.1]	**1.95** [1.60, 2.33]	No Data	No Data
	Low-credibility	**3.5** [2.8, 4.4]	**0.69** [0.5, 0.93]	No Data	No Data

Table 5.3 - 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. 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).

5.3 METHODOLOGY FOR THE MONETISATION 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 monetisation and follow our Partner Monetisation Policies". Those lists refer specifically to accounts whose video content can be used for monetisation (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 monetisation (subscription, Meta Stars, Facebook Content Monetisation program, branded content), these lists are not exhaustive and can only be considered indicative of the broader phenomenon of the monetisation of disinformation on Facebook.

Starting from the Facebook accounts identified in Section 2.3 and assigned a credibility label, we applied additional eligibility criteria. Because the partner-publisher lists focus on video monetisation, pages or profiles that had not recently published videos or reels, or that had reached only minimal audiences on such content, were marked as ineligible[1].

In July 2025, Facebook began rolling out its Facebook Content Monetisation Program[32], intended to consolidate several monetisation features. From September 2025 onward, this transition was expected to phase out the in-stream Ads and Reels ads programmes on which the partner-publisher lists are based. At the time of analysis, however, these lists were still being updated.

It is important to note that the partner-publisher lists do not capture the full range of monetisation mechanisms available on Facebook. Creators, whether high- or low-credibility, may monetise through other formats (e.g., photos, text posts, or stories), which are not reflected in the lists and may require lower production effort than video content[33].

A data access request under DSA Article 40.4 was submitted on 12 January 2026. At the time of writing, no response has been received.

5.3.2 Instagram

Instagram does not offer meaningful data to track account-level monetisation. Transparency has taken a step back as Instagram stopped publishing its “partner-publisher lists” that detailed the accounts that were eligible for monetisation in H2 2025. A data access request under DSA Article 40.4 was submitted on 12 January 2026. At the time of writing, no response has been received.

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 monetisation on X/Twitter. A data access request under DSA Article 40.4 was submitted on 12 January 2026. At the time of writing, no response has been received.

[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.

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 monetisation on LinkedIn. A data access request under DSA Article 40.4 was submitted on 12 January 2026. At the time of writing, no response has been received.

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 to 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% 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 the 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 monetisation on TikTok. A data access request under DSA Article 40.4 was submitted on 12 January 2026. At the time of writing, no response has been received.

5.3.6 YouTube

As with Facebook, we screened YouTube channels associated with high- and low-credibility actors (see Section 2.3) to determine which were eligible for monetisation based on publicly observable criteria, excluding any assessment related to compliance with YouTube’s community guidelines.

Eligibility was defined using two thresholds:

- more than 1 000 subscribers;
- more than 4 000 hours of watch time over the past 12 months (excluding Shorts).

As watch time is not directly observable, we approximated it using the same method as for Facebook: the sum of views on videos published in the last 12 months multiplied by video duration, divided by two to account for early drop-off.

Because official databases do not disclose monetisation status, we inferred it by examining the last ten videos published by each channel. If at least three displayed advertisements (either before or during the video playing, or in the top-right-hand corner of the video), the channel was considered monetised.

5.4 COMPARISON OF THE RESULTS ACROSS THE 2 REPORTS

The first full measurement of the SIMODS project was conducted between 17 March and 13 April 2025. During this period, we collected more than 2.6 million posts across the six VLOPs, using the standardised methodology described in Section 2.1.1.

The second measurement wave was conducted between 1 October and 31 October 2025. Applying the same keyword search framework, we collected approximately 3.3 million posts across the same platforms and languages.

Below, we compare the most prominent findings from the two measurement periods. The overall consistency of these results demonstrate the robustness of our methodology and its reproducibility. This supports the conclusion that our results reflect structural patterns across platforms and are not fundamentally driven by short-term variations.

5.4.1 Prevalence of Mis/Disinformation

Comparing the prevalence of misinformation across the two measurement periods suggests a high degree of stability in the overall magnitude of misinformation across platforms. At the aggregate level, the results from the two measurements are very consistent (Figure 5.8): the prevalence estimates remain within a comparable range, indicating that the indicator is capturing a structurally persistent phenomenon rather than short-term fluctuations.

That said, we observe some platform-level changes. On TikTok, misinformation prevalence increased from around 20% in March–April to 25% in October 2025. YouTube also registered a significant increase between the two periods, from around 8.5% to 12%. Values for LinkedIn didn't significantly decrease between the two periods (from around 2% to 1%), and the same can be said on other platforms.

The distribution of topics within our datasets remained fairly stable across the two measurement periods, with only minor variations observed. Health continued to be the dominant misinformation category in both waves, followed by the Russia–Ukraine war and national politics (Figure 5.9).

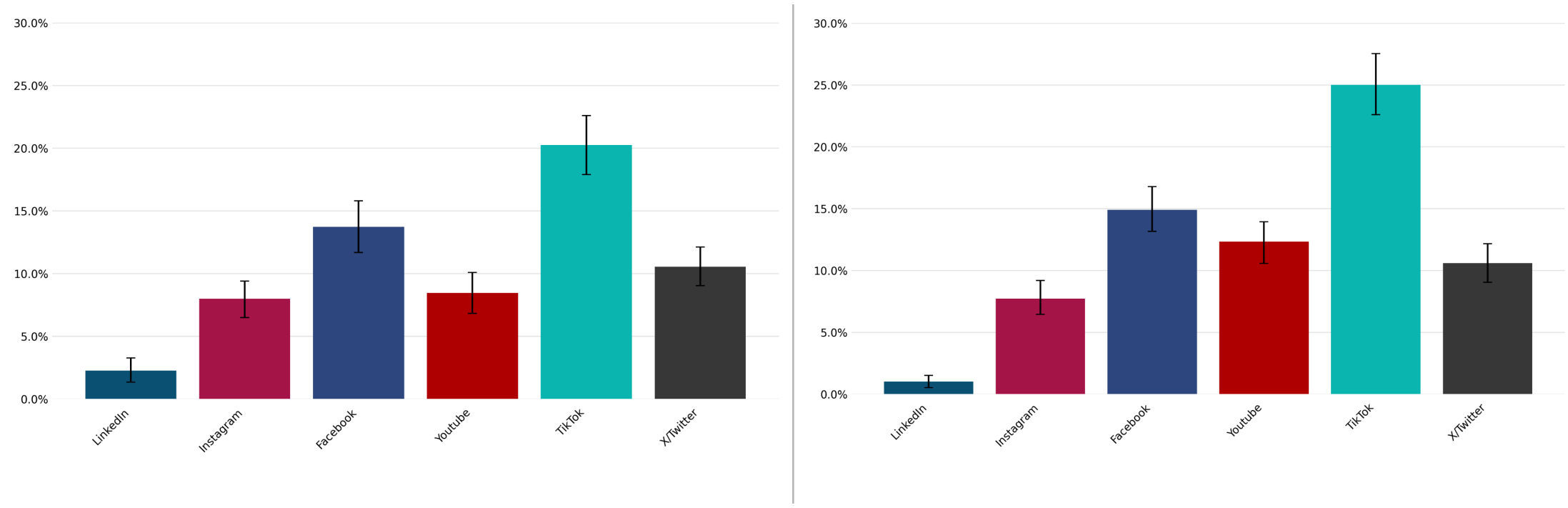


Figure 5.8 Comparison between the two time periods (March–April vs. October 2025): Prevalence of mis/disinformation across the six very large platforms, aggregated across all languages. The error bars represent the 95% confidence intervals measuring the uncertainty around each estimate, calculated using a bootstrapping method (see Appendix 5.1.4).

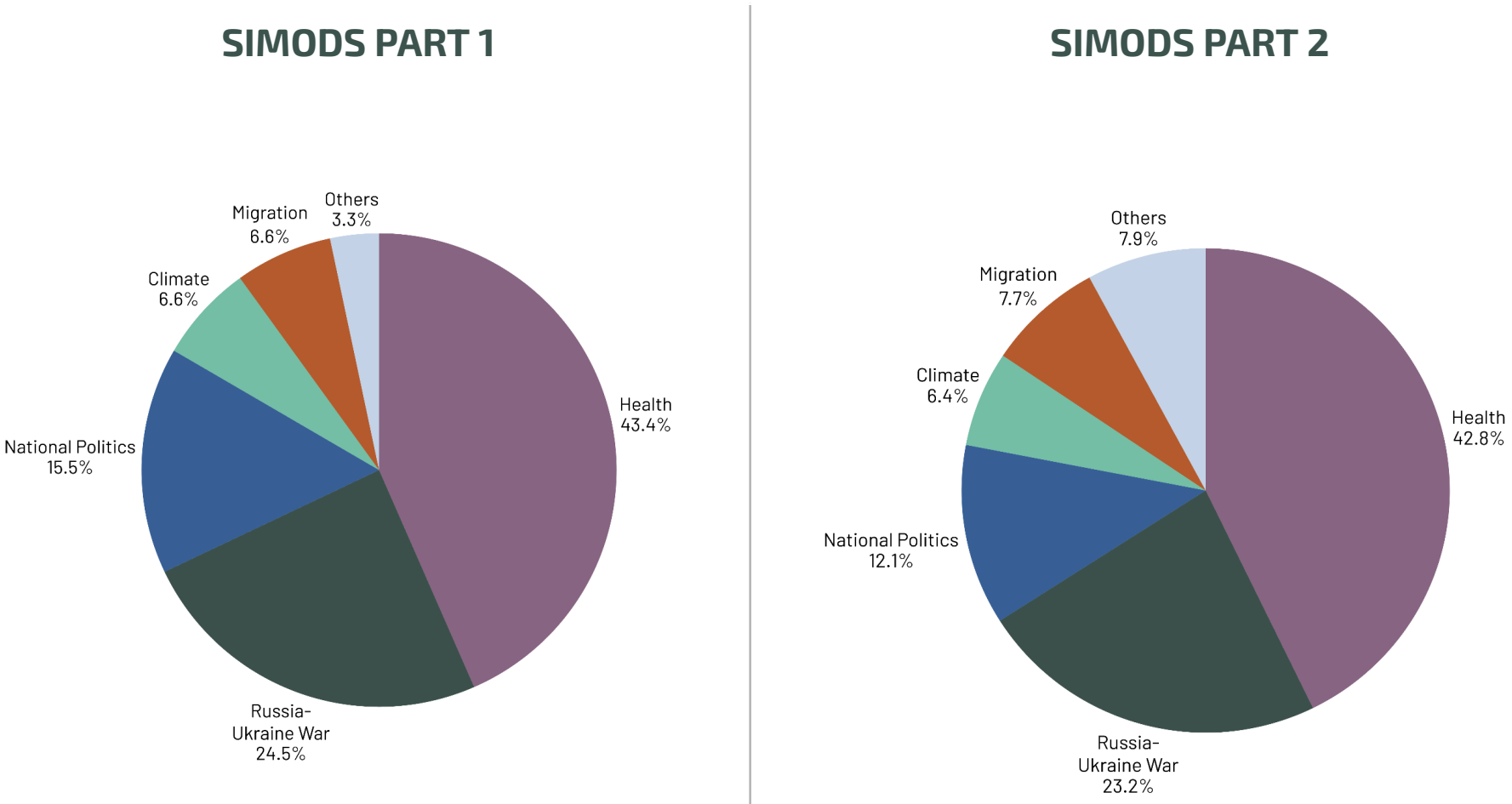


Figure 5.9 – Comparison between the two time periods (March–April vs. October 2025): Topic distribution of mis/disinformation posts across the studied data sample.

5.4.2 Sources of Mis/Disinformation

A. ACCOUNTS' AUDIENCE SIZE

In terms of audience size, follower counts remained broadly consistent across the two measurement periods. High-credibility sources continued to outperform low-credibility accounts, with the gap in magnitude between the two groups remaining relatively stable over time.

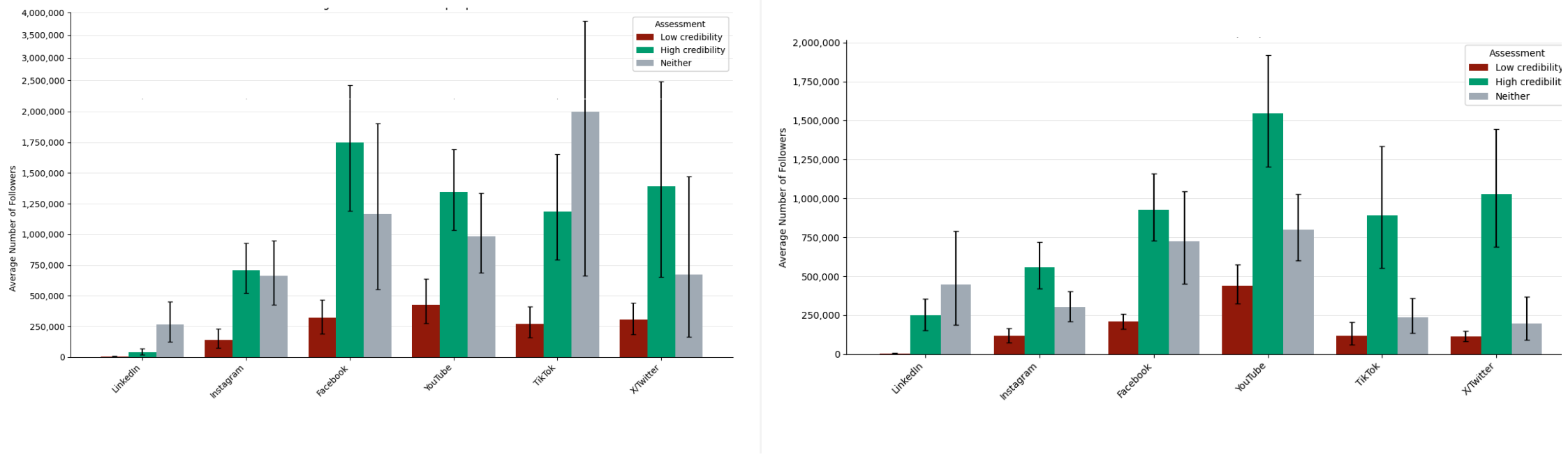


Figure 5.10 Comparison between the two time periods (March–April vs. October 2025): Average number of followers for accounts classified as High-credibility, Low-credibility, and Neither on each platform. The error bars represent the 95% confidence intervals measuring the uncertainty around each estimate, calculated using a bootstrapping method (see Appendix 5.1.4).

B. ACCOUNTS' ENGAGEMENT RATES: THE 'MISINFORMATION PREMIUM'

Looking at the metric *mean interactions per post per 1 000 followers*, which we use to calculate the ratio of engagement between low-credibility and high-credibility accounts, we observe some differences between the two measurement periods.

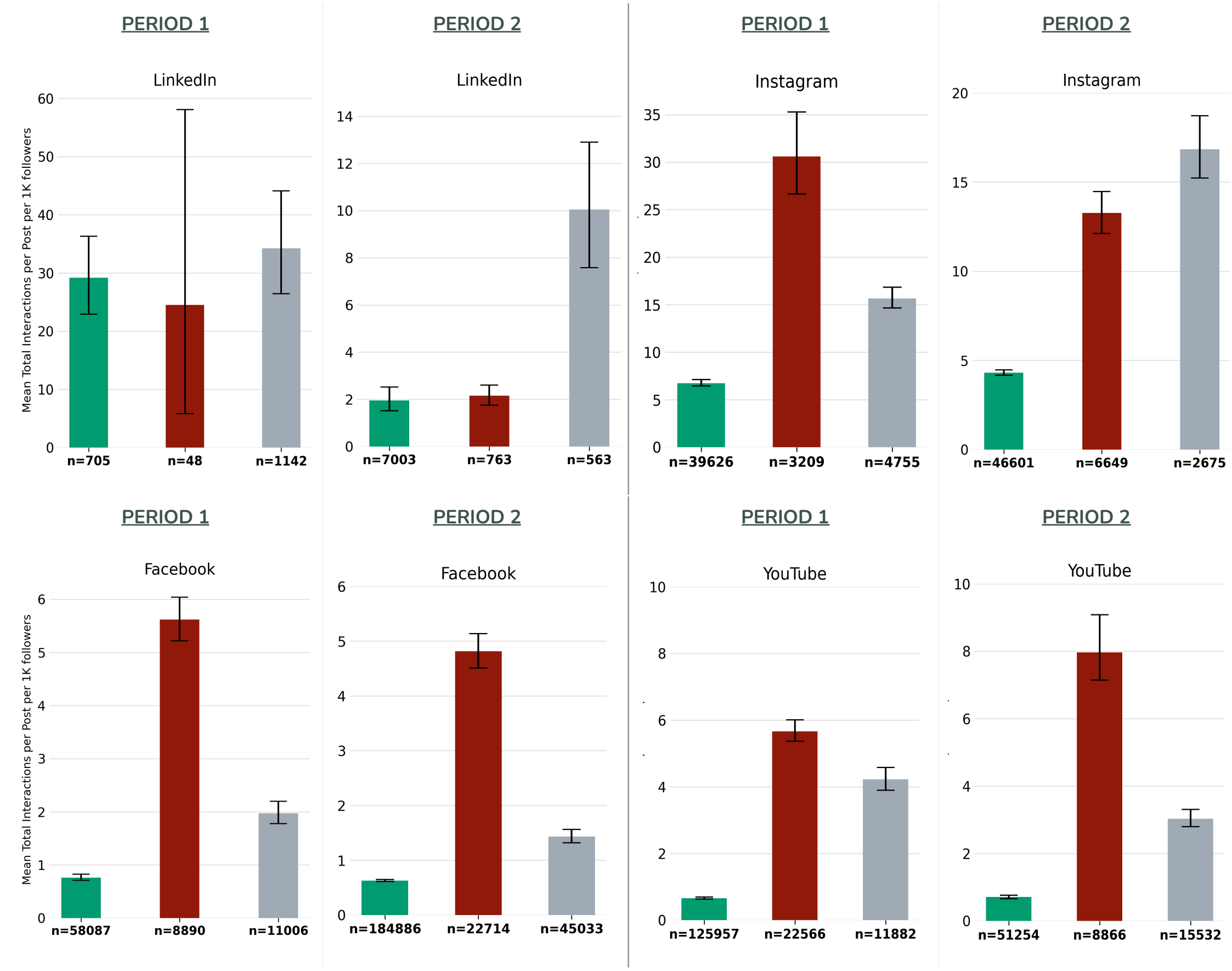
On platforms such as Facebook and YouTube, the mean number of interactions per 1 000 followers remained broadly consistent across the two measurement periods, with only limited variation. The most notable change was observed on YouTube, where the mean for low-credibility accounts increased from 5.5 to 8 interactions per post per 1 000 followers.

A noticeable difference can be observed on LinkedIn, with a tenfold drop in mean interactions per 1 000 followers. As shown in Figure 5.11, the number of posts (and of accounts) has increased substantially in the second period, providing a more representative

picture of engagement on LinkedIn across all three account types. In the first period, the limited number of accounts meant that engagement metrics were not representative, as reflected in the wide confidence intervals. In the second period, the greater diversity of the dataset enables a more robust estimation of these metrics, as evidenced by the narrower confidence intervals.

Regarding the “misinformation premium”, the values are similar between the two periods for most platforms (Figure 5.12). The most important, statistically significant, differences are observed on:

- X/Twitter, where the premium increased substantially from around 4 to around 10, and
- YouTube, where the premium increased from around 8.5 to around 11.



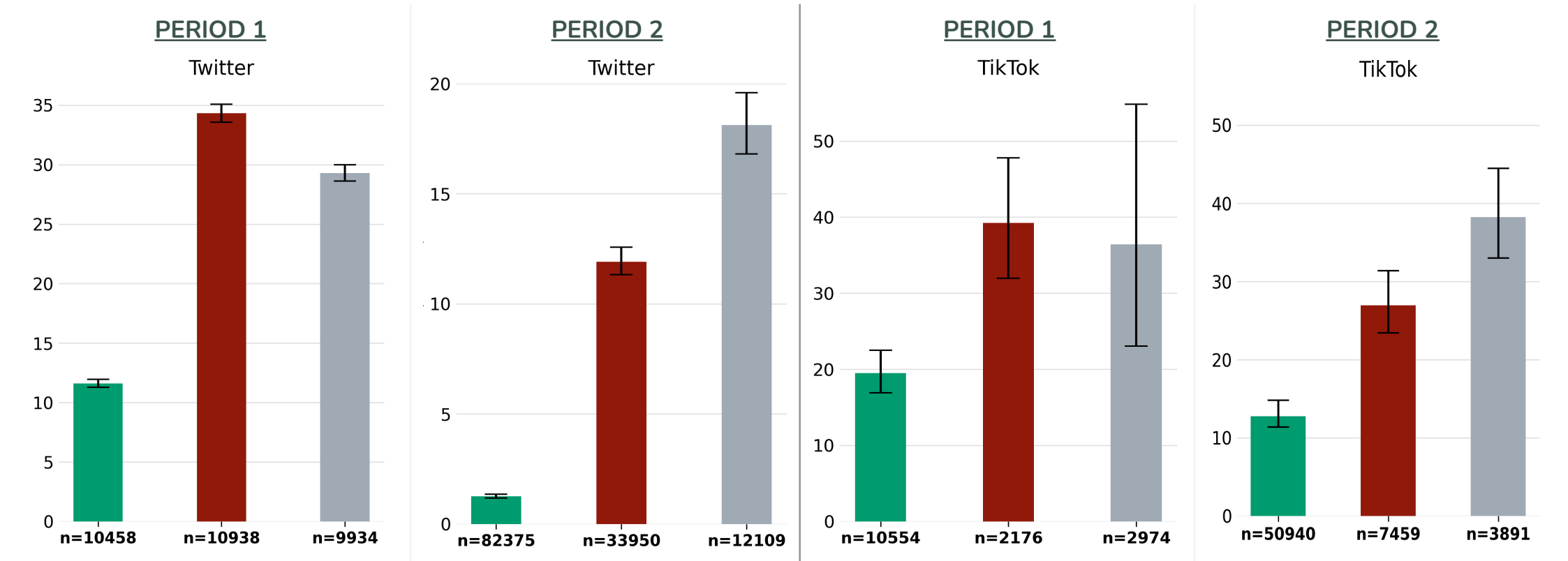


Figure 5.11 – Comparison between the two time periods (March–April vs. October 2025): Average number of interactions per post per 1 000 followers for accounts classified as High-credibility, Low-credibility, and Neither on each platform. The error bars represent the 95% confidence intervals measuring the uncertainty around each estimate, calculated using a bootstrapping method (see Appendix 5.1.4).

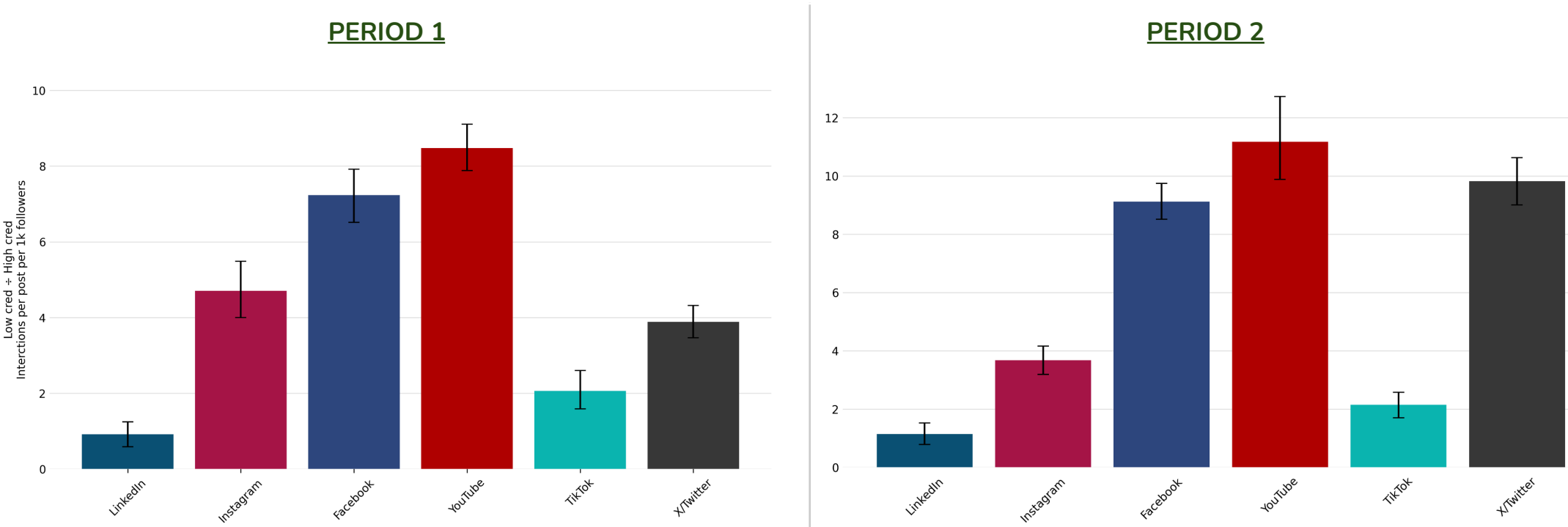


Figure 5.12– Comparison between the two periods (March-April vs. October 2025): Ratio of the average number of interactions per post per 1 000 followers for accounts classified as Low-credibility to the same number for High-credibility accounts. Error bars represent 95% confidence intervals, calculated using a bootstrapping method (see Appendix 5.1.4).

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] 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). Science Feedback https://science.feedback.org/wp-content/uploads/2025/09/SIMODS-Report-1.pdf

[8] 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

[9] 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

[10] Amarasinghe I, Romano S, Amidei J, Vincent EM, Kaltenbrunner A (2026) Uncertainty-Aware Estimation of Mis/Disinformation Prevalence on Social Media. (under submission) https://arxiv.org/abs/2603.11058

[11] 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

[12] Huang C & He G (2025) Text clustering as classification with LLMs. In Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. https://dl.acm.org/doi/abs/10.1145/3767695.3769519

[13] 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

[14] Guglielmi G (2024) Tiny number of "supersharers" spread the vast majority of fake news. Science https://www.science.org/content/article/tiny-number-supersharers-spread-vast-majority-fake-news

[15] 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

[16] 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/

[17] Stanusch N, Degeling M, Romano S, Çetin RB, Schüler M, Semenzin S (2025) AI-Generated Algorithmic Virality. arXiv preprint arXiv:2508.01042 https://arxiv.org/abs/2508.01042

[18] Koebler J (2025) AI Slop Is a Brute Force Attack on the Algorithms That Control Reality. 404 Media https://www.404media.co/ai-slop-is-a-brute-force-attack-on-the-algorithms-that-control-reality/

[19] Grochocka M (2026) Fake Doctors on Social Media: Alarming Business of Selling Unproved Dietary Supplements Online Circumventing Existing Regulations. Pravda Association https://pravda.org.pl/wp-content/uploads/2026/02/Fake-Doctors-_-Pravda-Association-Brief.pdf

[20] Teoh F (2026) Beware of AI-generated doctors giving health advice on social media: investigating the phenomenon on TikTok. Science Feedback https://science.feedback.org/beware-ai-generated-doctors-health-advice-tiktok/

[21] van Ess H (2026) Yearly Fact Check Intelligence Report. ImageWhisperer. Accessed March 2026. https://imagewhisperer.org/yearly-report

[Archived: https://web.archive.org/web/20260312142529/https://imagewhisperer.org/yearly-report]

[22] Maldita.es (2026) TikTok polarization industry: making money off disinformation with AI-generated videos of protests. https://maldita.es/investigaciones/20260126/protests-ai-tiktok-money-polarization/

[23] Meta (2024) Our Approach to Labeling AI-Generated Content and Manipulated Media. Meta Newsroom https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/

[24] Meta (2024) Labeling AI Content. Meta Transparency Center https://transparency.meta.com/governance/tracking-impact/labeling-ai-content/

[25] Meta Oversight Board (2026) Board Calls for New Rules on Deceptive AI During Conflicts. Oversight Board. March 2026 https://www.oversightboard.com/news/board-calls-for-new-rules-on-deceptive-ai-during-conflicts/

[26] TikTok (n.d.) AI-Generated Content Label. TikTok Creator Academy. https://www.tiktok.com/creator-academy/en/article/ai-generated-content-label

[27] TikTok (n.d.) Community Guidelines Overview. TikTok Creator Academy. https://www.tiktok.com/creator-academy/en/article/community-guidelines-overview

[28] YouTube/Google (n.d.) Disclosing Use of Altered or Synthetic Content. Google Support https://support.google.com/youtube/answer/10834785

[29] X (n.d.) Authenticity Policy. X Help Center https://help.x.com/en/rules-and-policies/authenticity

[30] LinkedIn (n.d.) Professional Community Policies. LinkedIn https://www.linkedin.com/legal/professional-community-policies

[31] Zamani M, Hajizadeh MR, Mosavat SH, Zamani M (2019) Mechanisms of honey on testosterone levels. Journal of Complementary and Integrative Medicine https://doi.org/10.1515/jcim-2018-0199

[32] Meta (n.d.) Facebook Content Monetization. Meta for Creators https://creators.facebook.com/tools/facebook-content-monetization

[33] Meta (n.d) Facebook Content Monetization. Streamlined Beta Program https://about.fb.com/news/2024/10/monetize-content-facebooks-new-streamlined-program/