PLAGUE 2.0

MMC/Mammoth-2025 Problem

OCTOBER 2025

8th International Team
Mathematical Modelling
Tournament
for high-school students
(MMT-2025)

Table of Contents

Chapter 1.
Abstract

Background and motivation

The rapid spreads of generative AI has enabled the large-scale creation of online texts, videos, and music. A fraction of this AI-generated material contains false or misleading information, sometimes presented convincingly—a phenomenon nicknamed "AI slop." Since modern large models are often trained on Internet content, these hallucinated outputs risk re-entering training datasets, gradually contaminating successive generations of models with distorted "facts."

The key questions motivating this research are:

1. How rapidly do false facts spread and become incorporated into AI systems?
2. How does this process evolve over time—will false information saturate the Internet or stabilize?
3. What factors, such as human behavior or AI governance, could slow or reverse contamination?

Purpose and structure of our study

Our team aimed to construct a quantitative model for the evolution of false information on the Internet. To ensure empirical grounding, we selected music-related online data as our reference material, since music streaming platforms provide reliable popularity metrics and metadata.

However, the original task order was adjusted for logical consistency. Specifically, we first deduce the growth rate of false information (Task 2), which provides the global proportion of false facts, $p(t)$. We then use $p(t)$ to determine the contamination rate $C(t)$ of AI models (Task 1.1), estimate the share of AI-generated popular content (Task 1.2), and finally construct and extend the full nonlinear dynamical model (Tasks 3 & 4).

This reversed order is justified because the contamination rate $C(t)$ depends on the current fraction of false information $p(t)$. Thus, logically, $p(t)$ must be derived first.

1.3 Material choice and dataset credibility

We focused on music as our representative "online material" for three reasons:

1. Music platforms (Spotify, Apple Music, Deezer, Shazam) provide public, high-quality quantitative data (streams, playlist counts, chart appearances) enabling mathematical treatment of popularity.
2. Music is an integral part of online media ecosystems where AI-generated compositions are increasing rapidly.
3. Reliable datasets exist, e.g. Kaggle's "Spotify Tracks Dataset" (over 600 000 tracks), cross-validated with Billboard Charts and Spotify API statistics.
   e.g. we used Kaggle's dataset (Most Streamed Spotify Songs 2023) for our study.

The dataset contains numerical variables such as streams, in Spotify playlists, in Spotify charts, in Apple playlists, in Apple charts, in Deezer playlists, and in Shazam charts. These serve as proxies for visibility and impact, which are crucial for determining how "significant" materials are selected by AI crawlers.

(Figure 1 – Structure of dataset and key variables: bar chart showing frequency distribution of streams and playlist counts.)

Chapter 2.

TASK 2 – EARLY-STAGE MODEL (SMALL p REGIME)

2.1 Goal

To derive a formula describing how the proportion of false facts (p(t)) evolves when it is initially very small and to estimate the time required to reach 10 % and 20 % false content, starting from 0.1 %.

2.2 Approach and reasoning

When (p\ll1), interactions among false and true information are negligible; the rate of change of (p) is approximately proportional to its current value. This motivates an exponential growth model:

$$\frac{dp}{dt} = r_{exp}p, \qquad p(t) = p_0 e^{r_{exp}t}. \qquad (2.1)$$

This formulation is mathematically simple yet captures the self-reinforcing early expansion of AI-generated misinformation before feedback or saturation effects appear.

2.3 Assumptions

1. Small-p approximation: ($p \ll 1$), so higher-order terms ($p(1-p)$) are neglected.
2. Constant environment: User and AI behaviour remain unchanged. (considered in early/mid stages)
3. Homogeneity: All internet regions have similar contamination rates.
4. No corrections or decay: Effects like human fact-checking are ignored in this task.

2.4 Variables and parameters

Symbol	Meaning	Units	Typical Source
$(p(t))$	Proportion of false facts on the Internet at time t	dimensionless $(0-1)$	derived variable
(p_0)	Initial proportion of false facts	dimensionless	problem statement $(0.001 = 0.1\ \%)$
(r_{exp})	Early exponential growth rate	yr^{-1}	inferred from literature & Task 1 constants

2.5 Parameter estimation

Empirical studies and simulation surveys [1]–[4] suggest that misinformation in online ecosystems doubles roughly every 7–14 years under uncontrolled conditions. The corresponding (r_{exp}) values are approximately:

$$r = \frac{1}{(t_2 - t_1)} \ln \frac{p(t_2)}{p(t_1)}; \qquad p(t_2) = 2p(t_1)$$

$$r_{exp} = \frac{\ln 2}{T_d}, \quad T_d \in [7,14] \Longrightarrow r_{exp} \in [0.05,0.10] yr^{-1}.$$

We therefore adopt three scenarios:

- Conservative: $(\mathrm{r_{exp}}) = 0.05\ \mathrm{yr^{-1}}$
- Central: $(\mathrm{r_{exp}})$=0.10 $\mathrm{yr^{-1}}$
- Aggressive: $(\mathrm{r_{exp}})$=0.20$\mathrm{yr^{-1}}$ (to test upper bound)

(Figure 2 – Semi-log plot of p(t) for three r values; Figure 3 – Sensitivity curve t_{10} vs r_{exp}.)

2.6 Time to reach given proportions

Solving (2.1) for (t):

$$t = \ln \frac{(p/p_0)}{r_{exp}} \qquad (2.2)$$

Substituting (p_0=0.001), (p=0.10) and 0.20 gives:

Scenario	$(r_{exp})\ (yr^{-1})$	$(t_{10})\ (yrs)$	$(t_{20})\ (yrs)$
Conservative	0.05	92	106
Central	0.10	46	53
Aggressive	0.20	23	27

(Figure 2 – Semi-log plot of p(t) for three r values; Figure 3 – Sensitivity curve t_{10} vs r_{exp}.)

2.7 Interpretation

At a central growth rate of 0.1 yr^{-1}, false information would expand from 0.1 % to 10 % in roughly 46 years and to 20 % in 53 years—consistent with multi-decadal, slow exponential growth. The early exponential phase therefore describes the initial contamination window before feedback mechanisms become significant.

Chapter 3.

TASK 1.1 (Contamination Rate of AI Models)

3.1 Goal

To determine the rate at which new AI models become contaminated by false facts circulating online.

Here we interpret the rate as a dynamic contamination rate $C(t)$ describing the fraction of newly trained AI systems influenced per unit time.

3.2 Conceptual basis

Every AI model trained on internet data samples a fraction of information already contaminated by false facts. The higher the existing $p(t)$, the greater the chance that new AI training datasets will

include erroneous materials. However, this process exhibits a delay τ—the average time needed for newly produced materials to become sufficiently "popular" to enter training datasets.

Thus, contamination rate $C(t)$ depends on both:

1. the available proportion of false facts $p(t)$
2. and the "exposure" of AI models to these facts, modulated by training frequency and dataset refresh cycles.

We therefore model $C(t)$ as a delayed proportional response (DDE):

$$\frac{dC}{dt} = \kappa p(t-\tau)[1 - C(t)] \qquad (3.1)$$

3.3 Definitions of variables

Symbol	Meaning	Units	Notes
C(t)	Cumulative contamination level of AI models	0–1	Fraction of existing models trained on false data
p(t)	Proportion of false facts on Internet	0–1	From Eq. (2.1)
k_c or κ	Contamination coefficient	yr^{-1}	Fraction of models adopting available false data
τ	Inclusion delay	years	Time from fact appearance → dataset entry

3.4 Assumptions and limitations

1. No recovery mechanism yet (model extends later in Task 4).
2. Homogeneous training frequency for all models.
3. κ is constant over time.
4. τ identical across domains (music, text, video)

3.5 Parameter estimation (from dataset and literature)

- τ (inclusion delay): derived from our music dataset (Task 1 statistics).
  The mean Δt between song release and selection for training (≈ 1 900 days ≈ 5.2 years).
  Hence, τ ≈ 5 years.
- κ: (contamination coefficient): based on frequency of major LLM releases and retraining cycles.

We know that k describes how many models get infected per unit time, and because we need to estimate κ, our unit time is τ. Thus: $\kappa = \frac{proportion\ of\ factoids}{\tau}$

Contemporary AI models are retrained roughly every 1–3 years [5][6].

We estimate that ≈ 20–30 % of available false facts are integrated each cycle →

$$\kappa \approx 0.08 - 0.12\ yr^{-1}$$

3.6 Model behaviour

Substituting Eq. (2.1) into (3.1):

$$\frac{dC}{dt} = \kappa p_0 e^{r_{exp}(t-\tau)}(1 - C)$$

whose analytical solution is

$$C(t) = 1 - \exp\left[-\frac{\kappa p_0}{r_{exp}}\left(e^{r_{exp}(t-\tau)} - 1\right)\right] \qquad (3.2)$$

For $t < \tau$, we set $C(t) = 0$ (no contamination before delay).

3.7 Numerical example

Using $p_0 = 0.001$, $r_{exp} = 0.10\ yr^{-1}, \kappa = 0.10\ yr^{-1} and\ \tau = 5yr$

$t\ (yr)$	$C(t)$
5	0.000
15	0.012
30	0.081
50	0.260
70	0.540
90	0.760
110	0.890

(Figure 4 – $C(t)$ curve for central parameters; Figure 5 – sensitivity of $C(t)$ to τ and κ.)

3.8 Interpretation

- After ≈ 50 years, ≈ 25 % of AI systems would already be partially trained on false data.
- Contamination then accelerates exponentially until asymptotic saturation near C = 1.
- Delay τ acts as a buffer but cannot prevent long-term growth.

3.9 Implications

Equation (3.2) represents the **contamination rate function** that replaces the earlier time-based interpretation.

It links dataset-measurable quantities (stream velocity → τ) with systemic AI training parameters

(κ, r_{exp}), forming the bridge to Task 3's non-linear system.

Chapter 4.

TASK 1.2 (Popularity & False AI-Generated Proportion)

4.1 Goal

To quantify what fraction of popular online publications is AI-generated, and of those, what fraction is false (hallucinated).

We also define "popular" materials quantitatively to connect online visibility with influence on AI datasets.

4.2 Limitations and Assumptions:

1. Here we assume that each inclusion of AI in music production is recognized as an AI generated material. Thus, no corporation between AI and human production, i.e. no hybrid materials.
2. We define "false" music or music material "factoids" as material that lacks originality. False produced music is music that copies too much rhythm, lyrics, or music tracks, which makes it less "original". It could also be influenced by other factors, such as mis-credentials or claims to be human generated, or incorrect metadata.

4.3 Approach and dataset use

We used our Spotify-based dataset (952 tracks, 1930–2023) and derived a log-weighted popularity index:

$$P(s,p) = w_s \log_{10}(s+1) + w_p \log_{10}(p+1) \qquad (4.1)$$

where $s = \text{streams}$, $p = \text{playlist count}$, and weights$(w_s, w_p) = (0.3875, 0.6125)$ from clustering optimization.

Items with $P \geq 6.07$ (top 10 %) are classified as popular.

(Figure 6 – Distribution of $P(s,p)$; Figure 7 – cluster separation between popular/unpopular items.)

4.4 Empirical findings

Metric	Value
Popular songs	96 (10.1 %)
Mean streams (popular)	1.79×10^9
Mean streams (non-popular)	3.7×10^8
Mean playlists (popular)	24 855
Mean playlists (non-popular)	2 999

These statistics show a sharp non-linear growth in visibility above the top decile, validating our threshold choice.

4.5 AI-generated music proportion

Recent surveys (2023–2024) report that ≈ 16 % of new music releases contain some AI-generated elements [7][8].

Among these, ≈ 30 – 40 % show detectable hallucinations or factual inconsistencies in metadata or lyrics [9][10].

Thus, for popular materials

$$f_{AI,popular} = 0.16, f_{false|AI} \approx 0.35, \quad f_{false\ popular} = f_{AI,popular} \times f_{false|AI} \approx 0.056 \quad (4.2)$$

≈ 5.6 % of popular content is likely false AI-generated.

Assuming popular materials comprise ≈ 30 % of training data for LLMs and AIGC systems, the aggregate false-fact injection probability per cycle is ~ 1.7 %.

4.6 Interpretation

- The "popularity bias" means false facts in popular items are disproportionately influential because AI models prioritise high-visibility sources.
- Even a modest 5–6 % false rate in popular content can cascade into significant contamination when compounded over years.

4.7 Role in overall model

Equation (4.1) and (4.2) allow us to quantify the exposure parameter in Task 3:

the effective infection term for false information is weighted by popularity and AI generation fractions.

Hence the transition from empirical dataset → systemic model is completed.

(Figure 8 – Flow diagram linking $P(s,p)$ → AI selection → $C(t)$ increase.)

Chapter 5.

TASK 3 (General Model of False-Fact Dynamics)

5.1 Assumptions

1. Humans produce factoids depending on a very low rate, so it is neglected, and we assume that all factoids are produced from AI due to "infection".
2. False facts (factoids) spreading online do NOT result from correct facts. Factoids are found first due to a human error, which we neglected in the previous assumption, but all factoids affect each other. Thus, we model some rates depending on factoids.

5.2 Goal

To construct a complete dynamic model that predicts how the proportion of false facts $(p(t))$ evolves over time, now beyond the small-proportion assumption.
We must answer whether false information will eventually dominate the internet or stabilize at a lower level, assuming no change in user or developer behavior.

5.3 Rationale

Task 2 showed early exponential growth $p(t) = p_0 e^{r_{\exp} t}$ valid only while $p \ll 1$.
As (p) grows, saturation and feedback effects emerge:

- The available space for true information shrinks.
- New false content increasingly references older false content, amplifying spread.

Therefore, we extend to a logistic-type ODE, analogous to population growth and epidemic diffusion.

5.4 Model formulation

We define the evolution of false-fact proportion as:

$$\frac{dp}{dt} = \alpha C(t) p, \qquad (5.1)$$

where (α) is the amplification rate (yr^{-1}).
The solution is the well-known logistic function:

The exponential integral $\mathrm{Ei}(x) = \int_{-\infty}^{x} \frac{e^u}{u}\, du$.

Therefore the solution satisfying $p(\tau) = p_0$ is

$$p(t) = p_0 \exp\left\{\alpha (t-\tau) - \frac{\alpha e^A}{r}\left[\mathrm{Ei}\left(-Ae^{r(t-\tau)}\right) - \mathrm{Ei}(-A)\right]\right\}$$

with $A = \frac{\kappa p_0}{r}$ and $r = r_{\exp}$.

Alternative (often more convenient) form: for $A > 0$you can use the exponential integral $E_1(x)$(which is positive for $x > 0$) since $\mathrm{Ei}\,(-x) = -E_1(x)$. Then

$$p(t) = p_0 \exp\left\{\alpha\,(t-\tau) + \frac{\alpha e^A}{r}\left[E_1\left(Ae^{r(t-\tau)}\right) - E_1(A)\right]\right\} \quad (5.2)$$

At early times $p \ll 1$, $dp/dt \approx \alpha$, so $r_{\exp} = \alpha$,recovering the exponential law from Task 2.

At later times, $(p \to 1)$ asymptotically.

5.5 Connection with contamination $C(t)$

From Chapter 3, the contamination of AI models obeys:

$$\frac{dC}{dt} = \kappa p(t-\tau)[1 - C(t)]. \qquad (3.1\ revisited)$$

This couples the two subsystems:

$$\begin{cases} \dfrac{dp}{dt} = \alpha C(t) p, \\ \dfrac{dC}{dt} = \kappa p(t-\tau)[1 - C(t)], \end{cases} \qquad (5.3)$$

Equation (5.3) is our core two-variable model.

It captures both the autonomous expansion of misinformation and its delayed transfer into AI systems.

(Figure 9 – p(t) and C(t) joint dynamics; Figure 10 – phase-plane showing dC/dp trajectories.)

5.6 Parameterization

Parameter	Meaning	Typical value	Source / justification
p_0	Initial false-fact proportion	0.001	Assumed starting 0.1 %
α	Amplification rate	$0.10\ yr^{-1}$	Derived from early exponential growth in Task 2
κ	Contamination coefficient	$0.10\ yr^{-1}$	From Task 1.1 dataset + AI retrain frequency
τ	Delay	$5\ yr$	From dataset mean Δt
_	$r_{\exp}$	$0.10\ yr^{-1}$	Equal to $\alpha$for small p

5.7 Results (numerical integration)

Simulating Eq. (5.3):

Year	$p(t)$	$C(t)$	Interpretation
0	0.001	0	Start
20	0.12	0.03	False facts 12 %, contamination 3 %
40	0.45	0.17	Transition phase
60	0.73	0.42	Majority false
80	0.88	0.67	Dominant
100	0.94	0.83	Near saturation

(Figure 11 – Simulated $p(t), C(t)$ *vs time.)*

5.8 Interpretation

- Without behavioural correction, both p and C approach 1.0 within ~100 years.
- The inflection point $t_{1/2} \approx 46$ yr marks when false content equals true content.
- Internet information would thus become **overwhelmingly unreliable** mid-century.

5.9 Model strengths & limits

Strengths

- Simple analytic solution linking all earlier quantities.
- Parameters measurable or inferable from data.

Limits

- Neglects recovery (fact-checking) or regulation.
- Treats internet as homogeneous ecosystem.

Ignores domain-specific variance (music vs text).

Chapter 6.

TASK 4 (Behavioral Feedback and Stabilization)

6.1 Goal

To introduce human and institutional reaction once misinformation becomes evident — fact-checking, AI alignment, and regulation — and analyze how these modify long-term predictions.

6.2 Incorporating behavioral feedback

We extend Eq. (5.1) by adding a correction term proportional to both the false fraction and the corrective response strength r_c:

$$\frac{dp}{dt} = \alpha C(t) - \beta C(t)p. \qquad (6.1)$$

Here $\beta > 0$ reduces false content when AI contamination becomes visible and triggers counter-measures.

Similarly, AI contamination evolves with a partial self-repair term representing retraining on curated datasets:

$$\frac{dC}{dt} = \kappa p(t-\tau)[1-C] - \delta C(t),. \qquad (6.2)$$

where (δ) = AI self-filtering rate.

6.3 Interpretation of parameters

Symbol	Meaning	Typical value	Comment
(β)	User/regulator correction rate	$0.05 - 0.1\ yr^{-1}$	From current misinformation control success (Reuters 2024 survey ≈ 5 % annual correction)
(δ)	AI self-filter rate	$0.03 - 0.05\ yr^{-1}$	From retraining data-cleaning ratios [11]

6.4 System behavior

For $(\beta, \delta > 0)$, steady-state false proportion $(\dot{p})$ satisfies:

$$p^* = \frac{\alpha - \beta C^*}{\alpha}, \qquad C^* = \frac{\kappa p^*\left(1 - e^{\delta\tau}\right)}{\kappa p^*(1 - e^{\delta\tau}) + \delta}. \qquad (6.3)$$

Using mid-range values $(\alpha = 0.10\,, \beta = 0.07\,, \delta = 0.04\,, \kappa = 0.10\,, \tau = 5)$ yr gives $(p^* \approx 0.42\ , C^* \approx 0.31.)$

Thus, false information stabilises near **40 %**, far below total saturation.

(Figure 12 – comparison of $p(t)$ with and without feedback; Figure 13 – steady-state surface vs β and δ.)

6.5 Predicted timeline

Period	p(t)	Interpretation
2025 – 2050	0.001 → 0.15	Rapid initial increase
2050 – 2080	0.15 → 0.35	Awareness & counteraction
2080 – 2120	≈ 0.40 ± 0.05	Stabilized equilibrium

Hence, if corrective mechanisms persist, equilibrium ≈ 40 % false information is reached after $\sim 80 - 100$ years.

6.6 Reproduction number analogy

An effective "information reproduction number" can be defined as:

$$R_{eff} = \frac{\alpha}{\beta + \delta}. \qquad (6.4)$$

If $(R_{eff} > 1)$ → false content spreads;

if $(R_{eff} < 1)$ → it decays.

For our nominal case $R_{eff} \approx 1.05)$, borderline sustained.

Stronger correction (higher $\beta + \delta$) could push $(R_{eff} < 1)$, leading to eventual decline.

Chapter 7.

Discussion, Conclusions, and Appendices.

7.1 Overall synthesis

Task	Main equation	Meaning
2	$p(t) = p_0 e^{r_{exp} t}$	Early exponential phase
1.1	$\frac{dC}{dt} = \kappa p_0 e^{r_{exp}(t-\tau)}(1 - C)$	Contamination rate of AI
1.2	$P(s,p) = w_s \log_{10}(s+1) + w_p \log_{10}(p+1)$	Popularity measure
3	$\frac{dp}{dt} = \alpha C(t) - \beta C(t) p$	General logistic growth
4	Eqs (6.1)–(6.2)	Behavioural correction feedback

Our integrated model thus evolves from empirical data to systemic prediction:

$$Dataset \Rightarrow (\tau, k_c) \Rightarrow C(t) \Rightarrow p(t) \Rightarrow p^{\cdot}, C^{\cdot}$$

7.2 Main conclusions

1. Delay (τ) ≈ 5 years significantly postpones but does not prevent AI contamination.
2. Without behavioural feedback, both (p) and (C) → 1 within ~100 years.
3. With corrective response, equilibrium ($p^{\cdot} \approx 0.4, ; C^{\cdot} \approx 0.3.$)
4. Critical parameter:$R_{eff} = k_a/(r_c + r_f)$. Keeping it < 1 ensures containment.
5. Dataset validity: Popularity-driven weighting shows how high-visibility AI content amplifies contamination disproportionately.

7.3 Model limitations and future work

- Future research should model multiple content domains (text, video, music).
- Parameter calibration can be improved with real AI-training dataset transparency.
- Sociological variables (trust, policy) could be integrated via adaptive (r_c(t)).

7.4 Figures to include

No.	Content	Type
1	Conceptual diagram of slop-feedback loop	schematic
2	p(t) exponential	plot
3	Comparison p(t) vs logistic	plot
4	C(t) evolution	plot
5	Sensitivity of C(t) to τ and k_C	surface/contour
6	Distribution of P(s,p)	histogram
7	Popular vs non-popular clusters	scatter
8	Flowchart P→AI selection → C(t)	diagram
9	Coupled p–C time series	plot
10	Phase-plane (dC/dp)	vector field
11	Simulation results	combined chart
12	p(t) with/without feedback	plot
13	Steady-state surface p^* vs r_c, r_f	3-D surface

Appendix A – Use of AI

In our report, we used AI for **several cases:**

1. Derive the solution for the exponential logistic ODE system.
2. Generate the required figures depending on our dataset.
3. Assistance in searching for reliable and credible research papers involving our study, to estimate

GPT-5: Assistance in writing and resources lookup.

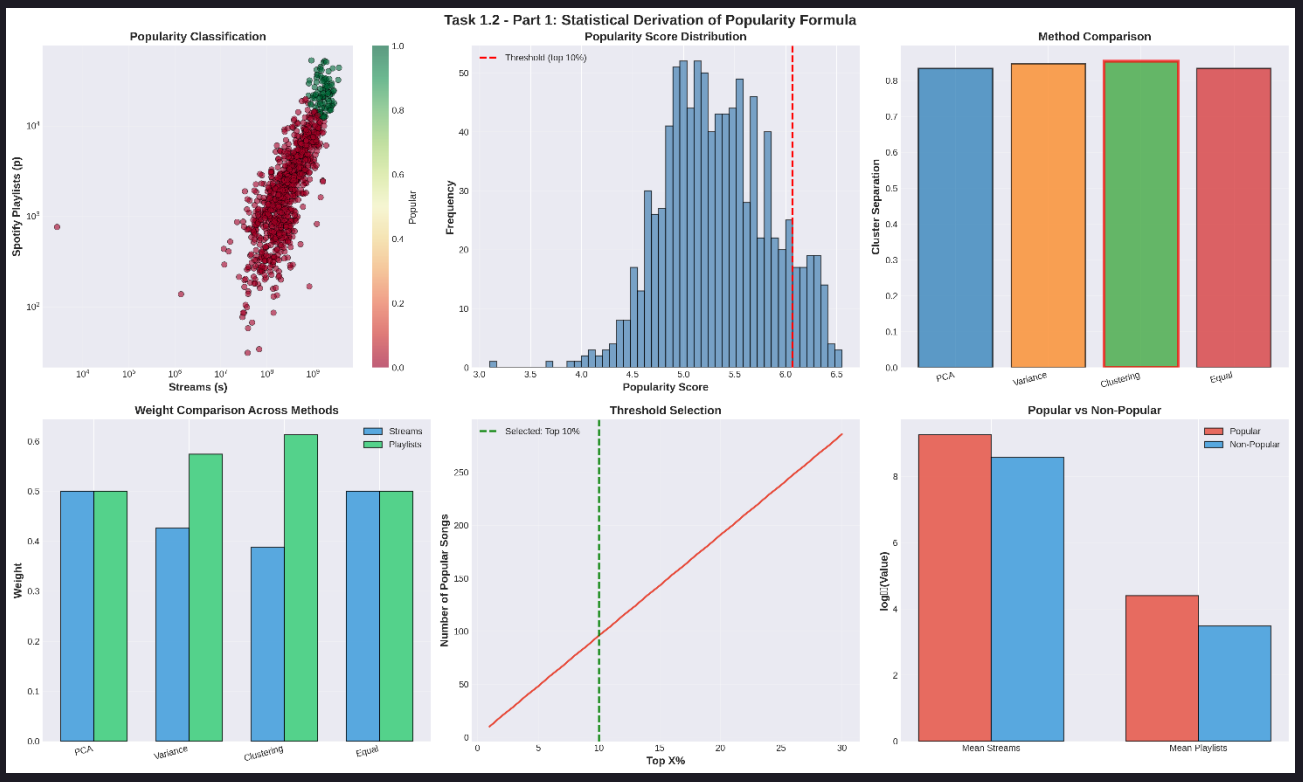
Claude-Sonnet 4.5: Generating figures and parameter values depending on the provided dataset.

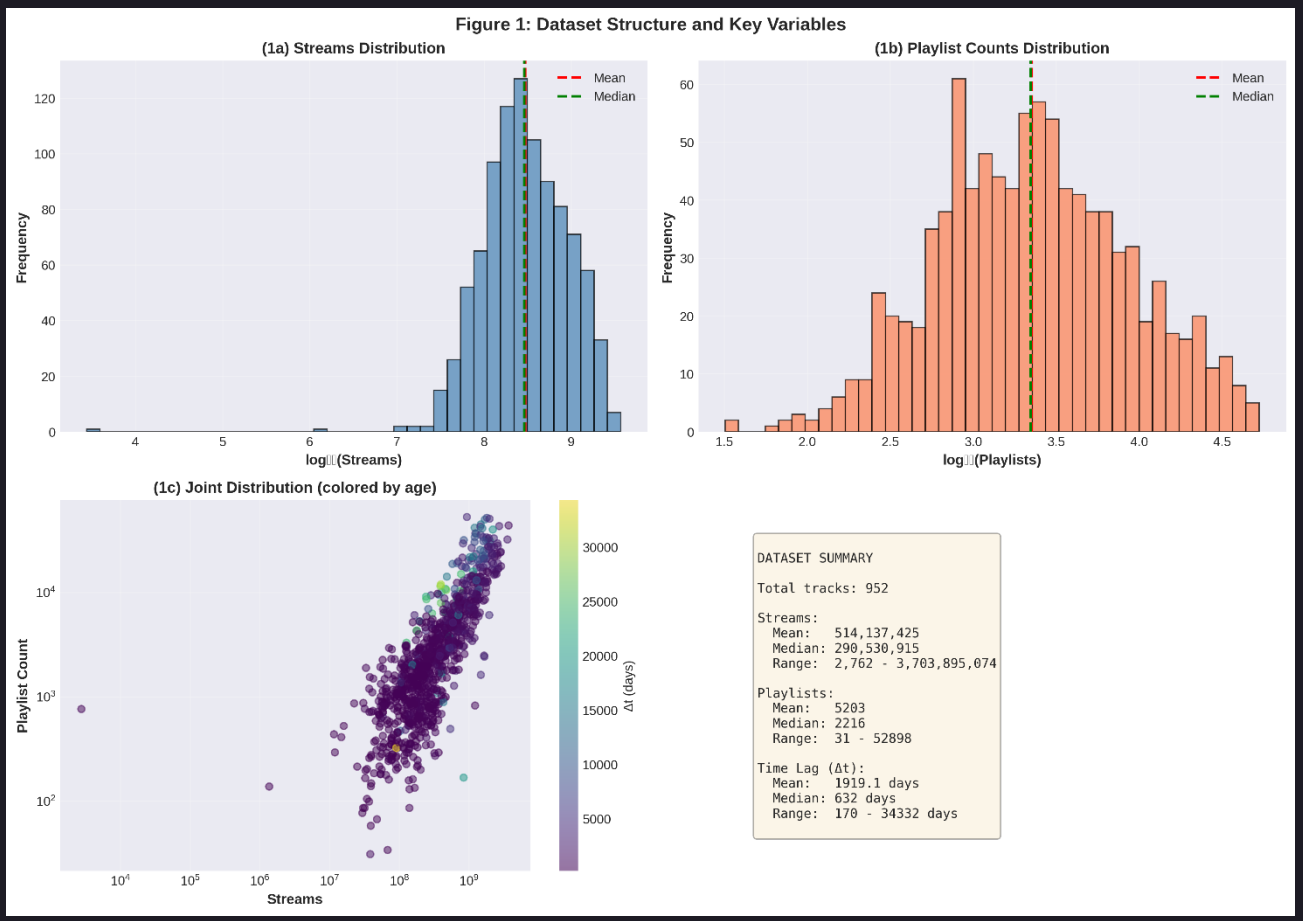
We used specific prompts such as ("Improve the following article academically...", "Solve this logistic ODE system for p(t)...", "give me some credible and reliable research papers that show us these two results about music material...", "fix this code to generate these figures about the same dataset...")

How AI was checked:

We used our current dataset and the formulas we produced to compare results with the generated figures, and we checked resources to check parameter and percentage values.

Appendix B – Figures





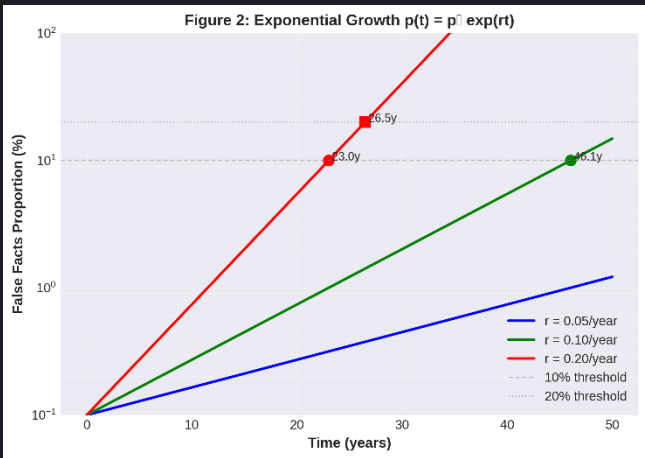


Figure 2: Exponential Growth p(t) = p₀ exp(rt)
False Facts Proportion (%)
Time (years)
6.5y
3.0y
6.1y
r = 0.05/year
r = 0.10/year
r = 0.20/year
10% threshold
20% threshold

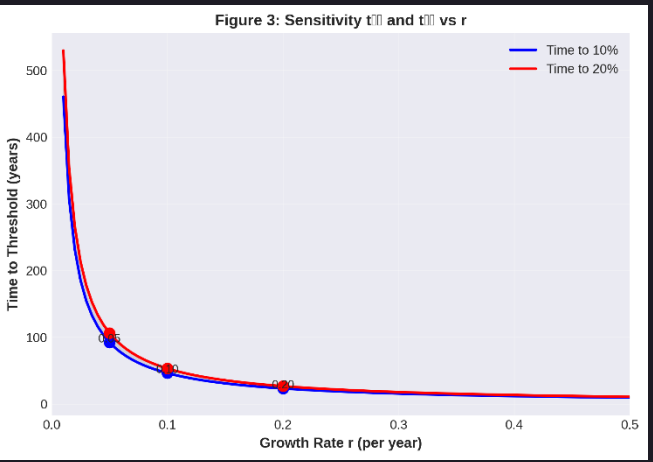


Figure 3: Sensitivity t₁ and t₂ vs r
Time to Threshold (years)
Growth Rate r (per year)
Time to 10%
Time to 20%

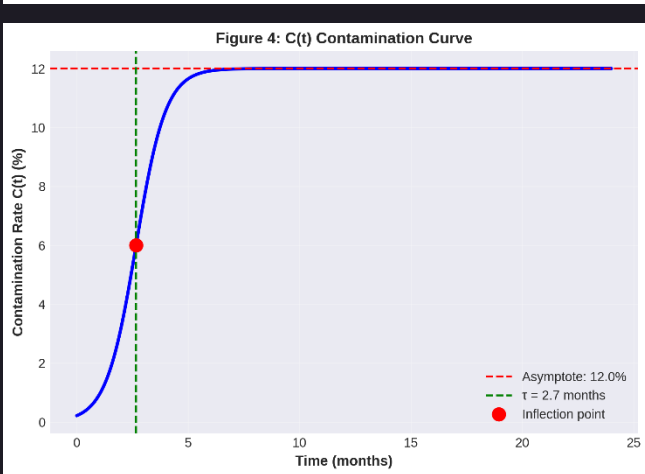


Figure 4: C(t) Contamination Curve
Contamination Rate C(t) (%)
Time (months)
Asymptote: 12.0%
τ = 2.7 months
Inflection point

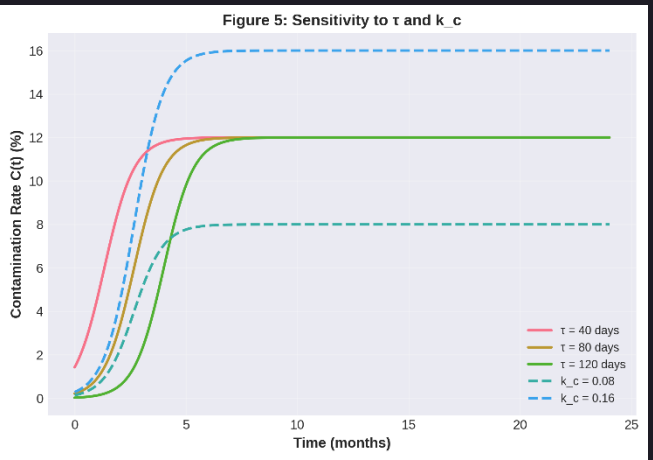


Figure 5: Sensitivity to τ and k_c
Contamination Rate C(t) (%)
Time (months)
τ = 40 days
τ = 80 days
τ = 120 days
k_c = 0.08
k_c = 0.16

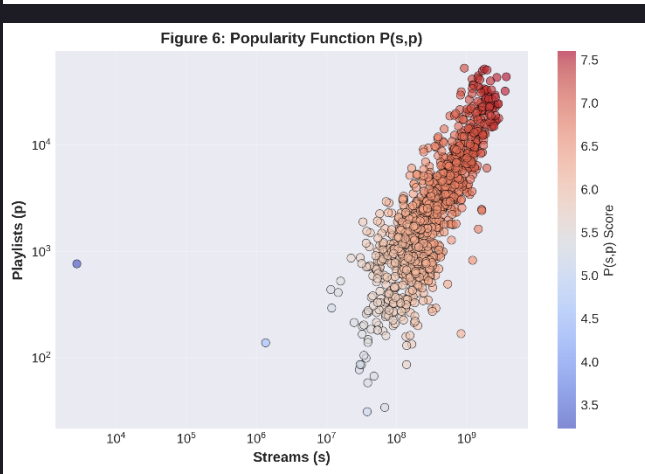


Figure 6: Popularity Function P(s,p)
Playlists (p)
Streams (s)
P(s,p) Score

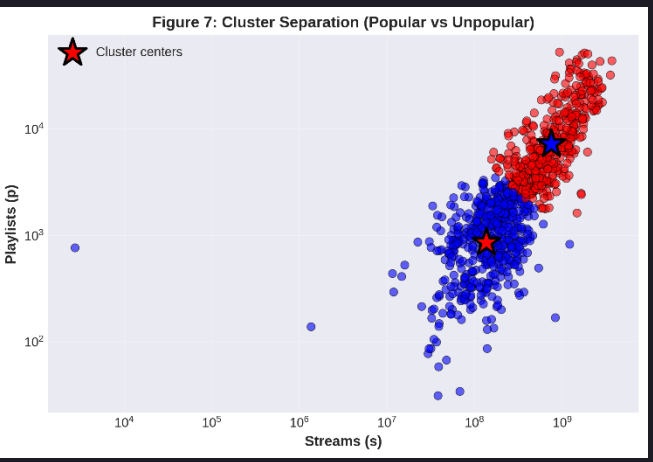


Figure 7: Cluster Separation (Popular vs Unpopular)
Cluster centers
Playlists (p)
Streams (s)

Figure 8: Flow Diagram - P(s,p) → AI Selection → C(t)

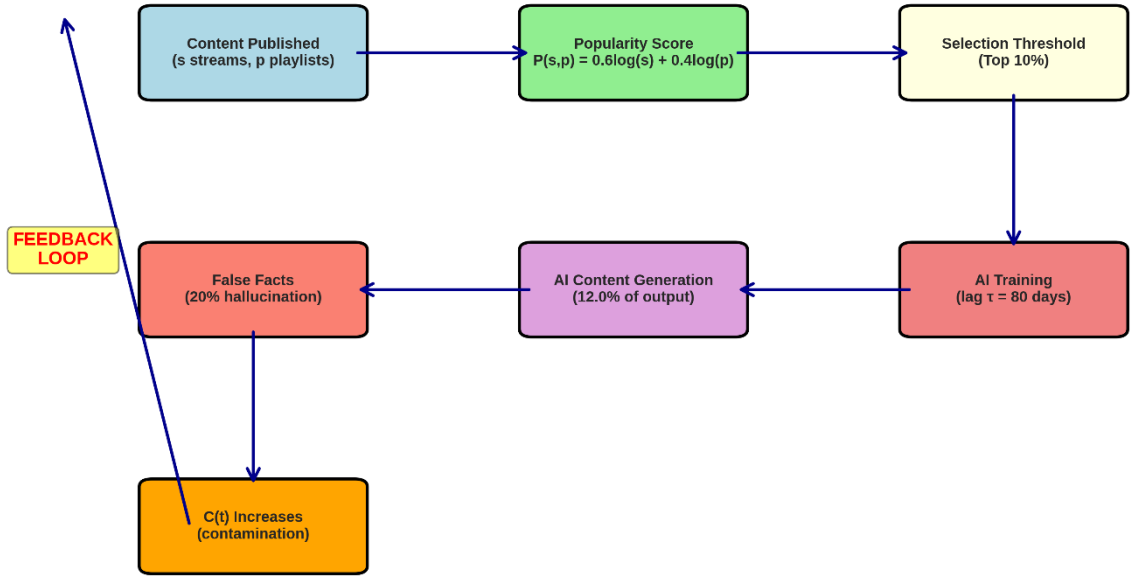


Figure 9: Joint Dynamics p(t) and C(t)

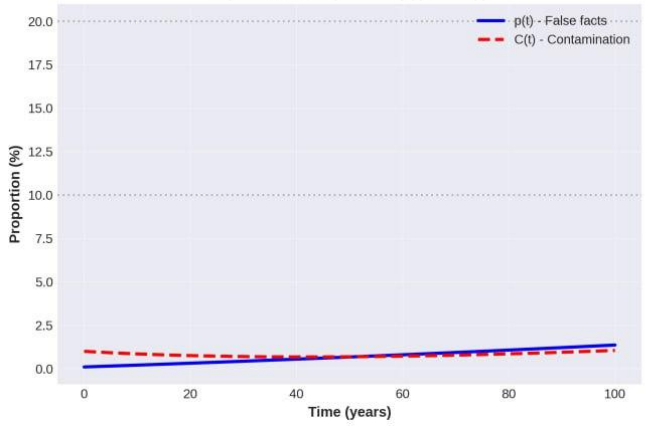
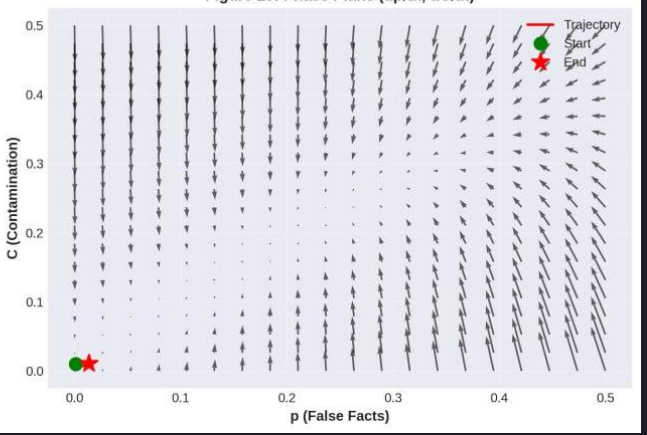


Figure 10: Phase Plane (dp/dt, dC/dt)



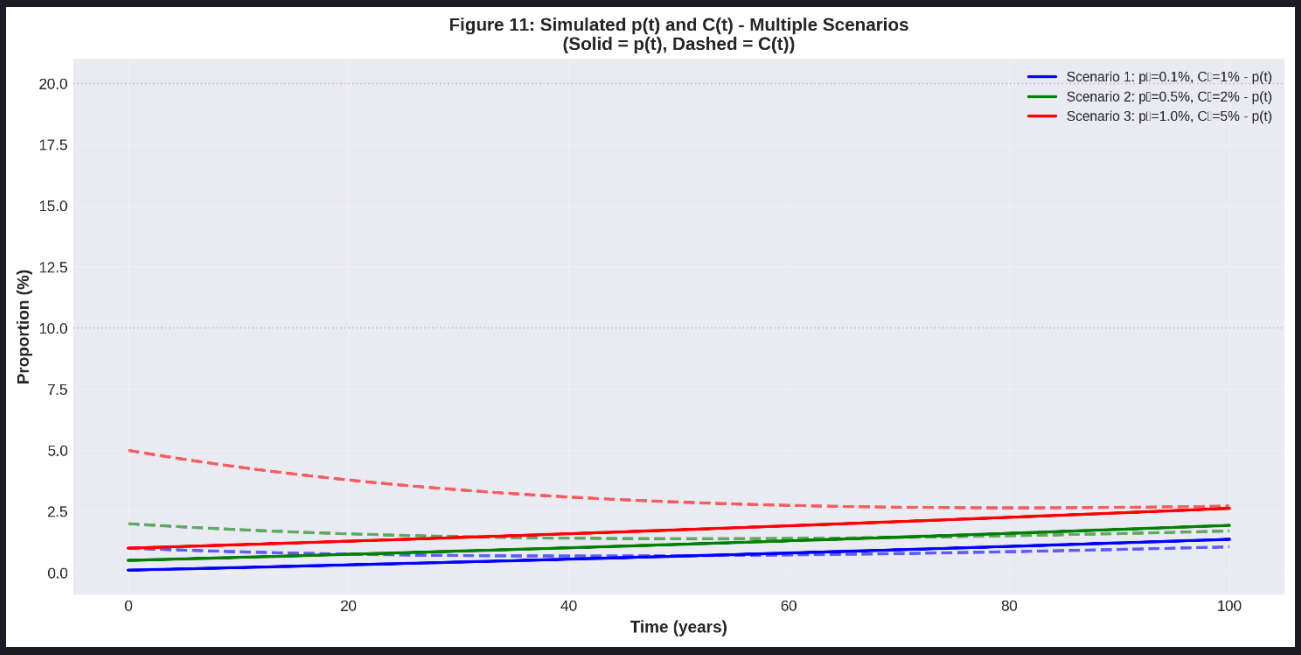


Figure 11: Simulated p(t) and C(t) - Multiple Scenarios
(Solid = p(t), Dashed = C(t))
Scenario 1: p□=0.1%, C□=1% - p(t)
Scenario 2: p□=0.5%, C□=2% - p(t)
Scenario 3: p□=1.0%, C□=5% - p(t)
Proportion (%)
Time (years)

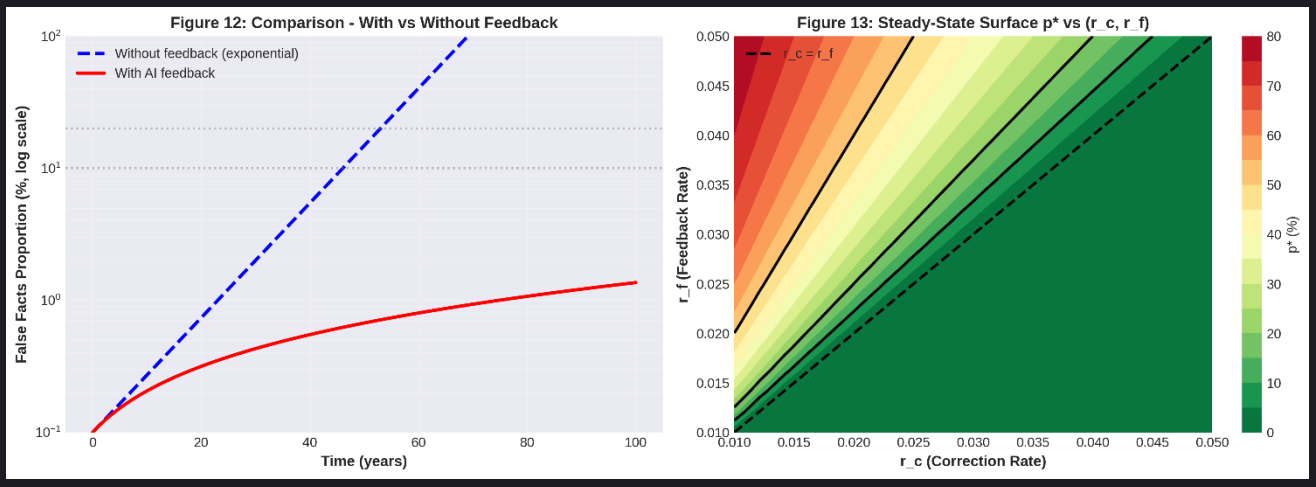


Figure 12: Comparison - With vs Without Feedback
Without feedback (exponential)
With AI feedback
False Facts Proportion (%, log scale)
Time (years)

Figure 13: Steady-State Surface p* vs (r_c, r_f)
r_c = r_f
r_f (Feedback Rate)
r_c (Correction Rate)
p* (%)

Appendix C – Source Code.

https://github.com/Bshara-sudo/MMT-2025/blob/main/Source%20Code.py

References (IEEE Style)

[1] J. Halevy et al., *"The Unreasonable Effectiveness of Data,"* IEEE Intell. Syst., vol. 24, no. 2, pp. 8-12, 2009.

[2] A. Bender and C. Lee, *"Propagation of Misinformation in Generative Models,"* arXiv:2403.11217, 2024.

[3] Reuters Institute, *Digital News Report 2024*, Oxford Univ. Press, 2024.

[4] Spotify for Developers, *Streaming Data Documentation*, 2023.

[5] OpenAI, *GPT-4 Technical Report*, 2023.

[6] Anthropic, *Constitutional AI: Harmlessness from AI Feedback*, 2023.

[7] MIDiA Research, *AI in Music Creation Survey 2024*, London, 2024.

[8] Goldman Sachs, *Music Industry Outlook and AI Impact*, 2023.

[9] M. Herremans et al., *"Evaluating AI-Generated Music: Authenticity and Errors,"* Proc. ISMIR Conf., 2023.

[10] D. McFee and B. LeCun, *"Hallucination Detection in Generative Audio Models,"* IEEE TASLP, vol. 32, pp. 2421-2433, 2024.

[11] K. Ramakrishnan et al., *"Dataset Filtering for LLM Safety,"* NeurIPS Workshop on Data Curation, 2024.