Auditing Chinese Web-scale Corpora via Sampled BPE Token Statistics

⚠Caution: this paper may include offensive and upsetting content.

Qingjie Zhang[1], Ziqi Tang[1], Jie Zhang[2], Gelei Deng[3], Jinfeng Li[4], Yuefeng Chen[4], Yitong Yang[4], Hui Xue[4], Tianwei Zhang[3], and Han Qiu[1*]

[1]Tsinghua University [2]SiliconProspect AI [3]Nanyang Technological University [4]Alibaba Group

Emails: {qj-zhang24@mails., qiuhan@}tsinghua.edu.cn [*]Corresponding author

Abstract

Chinese web pollution has surfaced in LLMs, motivating audits of upstream Chinese corpora. However, auditing such corpora faces three challenges: (1) their web-scale size makes full scan costly; (2) prior analyses are often too coarse to expose token-level pollution; (3) Chinese web pollution is implicit and rapidly changing. We propose SAMPLED-BPE, a lightweight token-level auditing pipeline that samples a small subset and trains BPE tokenizer to surface polluted tokens. Experiments show that SAMPLED-BPE preserves usable estimates while substantially reducing runtime and memory: a 148.4× speedup and a 35.8× memory reduction induce only 4.25% relative error for pollution categories. We apply the pipeline to 11 open Chinese corpora and 6 Chinese Common Crawl snapshots from 2021 to 2026. The audit reveals widespread but uneven pollution across open corpora, as well as highly polluted and temporally shifting Chinese web content. We further release a hierarchical Chinese web token dataset[1] with 630k+ token records, each with web context, category, and explanation fields, organized as 92k+ trees to support review and tracing of pollution.

1 Introduction

Chinese web pollution has surfaced in LLMs, including spam- and pornography-related Chinese tokens in ChatGPT's vocabulary (Yang, 2024) and Chinese gambling content in Codex outputs (OpenAI Developer Community, 2026). Together with recent work linking polluted Chinese tokens to corpus pollution (Zhang et al., 2025), these observations motivate auditing upstream Chinese corpora to understand where such pollution originates, how prevalent it is, and how it may propagate into downstream datasets.

However, auditing upstream Chinese corpora remains difficult. *First*, mainstream Chinese corpora are often at web scale, consisting of hundreds of gigabytes to terabytes of text (Yuan et al., 2021; Chen et al., 2023; He et al., 2023; Oepen et al., 2025). Fully scanning such corpora is costly (Penedo et al., 2023, 2024; Abbas et al., 2023), e.g., 99 days for indexing 83TB data on a 128-core CPU node (Xu et al., 2025). *Second*, existing corpus analyses often remain coarse-grained such as language distribution, source domains, or quality scores at document-level (Dodge et al., 2021; Kreutzer et al., 2022; Soldaini et al., 2024; Hagar and Bandy, 2025). These views describe broad corpus composition, but cannot reveal fine-grained pollution at token-level. *Third*, Chinese web pollution is often implicit and rapidly changing (MacAvaney et al., 2019; Jiang et al., 2021; Zhang et al., 2025). A predefined keyword list can miss newly emerging polluted tokens and quickly become stale. Together, these challenges call for ***a lightweight, token-level, periodic auditing pipeline***.

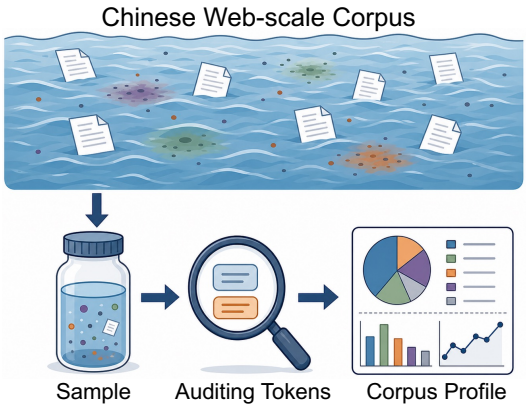


Figure 1: Intuition from marine pollution monitoring: sample a subset, train a tokenizer, map token-level pollution signals, and estimate a corpus profile.

Inspired by marine pollution monitoring (illustrated in Figure 1), which estimates overall pollution from sampled observations (Karydis and Kitsiou, 2013), we propose SAMPLED-BPE, a simple yet effective pipeline for auditing web-scale Chinese corpora when full inspection is too costly.

[1]https://github.com/qingjiesjtu/SampledBPE

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 30352–30379
October 24-29, 2026 ©2026 Association for Computational Linguistics

Specifically, we sample a small subset, train a byte-pair encoding (BPE) tokenizer (Sennrich et al., 2016) to surface and count high-frequency tokens, map these tokens to content categories using Internet search evidence, and aggregate token statistics into corpus-level profiles. This design avoids relying on a fixed keyword list, keeps the audit at token level, and makes repeated auditing feasible as Chinese web pollution evolves.

Experiments show that SAMPLED-BPE preserves usable token-level and category-level estimates while substantially reducing runtime and memory. Using this pipeline, we find that pollution is widespread but uneven across open Chinese corpora. We further show that Chinese Common Crawl (Common Crawl, 2011) is highly polluted and shifts over time, indicating the need for periodic upstream auditing. Finally, we release a hierarchical Chinese web token dataset to make these token-level findings reviewable.

This work makes three contributions:

- **Method.** We propose and validate SAMPLED-BPE, a lightweight pipeline for token-level auditing of web-scale Chinese corpora. *It reduces token-level auditing cost from months to hours while preserving usable estimates.*
- **Results.** We systematically audit *11 open Chinese corpora and 6 upstream Chinese Common Crawl snapshots from 2021 to 2026*, revealing widespread and shifting pollution. For example, 68.72% of 2026 Chinese Common Crawl snapshot is Adult Content.
- **Dataset.** We release a *hierarchical Chinese web token dataset* containing 630,684 token records, each with web context, category, and explanation fields. The dataset organizes tokens into 92,972 trees, supporting review and tracing of Internet pollution.

2 Background

In this section, we first summarize the corpora audited in this paper, then discuss related work on corpus auditing and Chinese web pollution.

2.1 Open Chinese Corpora and Landscape

LLM corpora span from MB-scale documents, to GB-scale tuning data, TB-scale pretraining corpora, and PB-scale upstream web archives (Wang et al., 2018; Longpre et al., 2023; Brown et al., 2020; Common Crawl, 2011). In this work, we use *web-scale* to refer to corpora in the TB range (Li et al., 2025; Roziewski and Stokowiec, 2016). Figure 2 places the Chinese corpora audited in this work within this scale landscape.

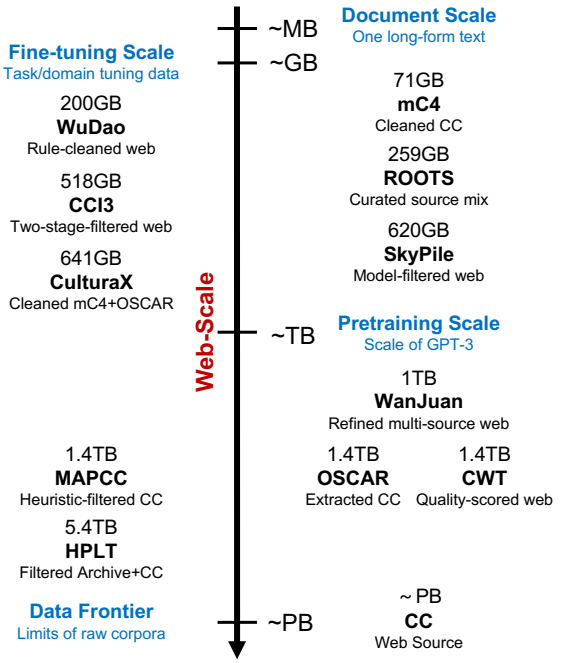


Figure 2: Scale landscape of Chinese web corpora audited in this work.

Mainstream open Chinese corpora include Chinese portions of broad multilingual corpora, such as OSCAR (Abadji et al., 2022), mC4 (Xue et al., 2021), HPLT (Oepen et al., 2025), CulturaX (Nguyen et al., 2024), and ROOTS (Laurençon et al., 2022); and Chinese-specific releases, such as CWT (Chen et al., 2023), WanJuan (He et al., 2023), MAPCC (Du et al., 2024b), SkyPile (Wei et al., 2023), CCI3 (Wang et al., 2024), and WuDao (Yuan et al., 2021). Detailed descriptions are provided in Appendix A.

2.2 Related Work

Corpus Curation and Auditing. Web-scale corpora are usually curated from web text through source selection, language identification, quality filtering, safety or blocklist filtering, and deduplication (Penedo et al., 2024; Soldaini et al., 2024; Weber et al., 2024; Penedo et al., 2023). Prior documentation and auditing work further improves transparency and reveals residual risks such as unexpected sources, low-quality text, duplication, contamination, and filtering side effects (Lee et al., 2022; Kreutzer et al., 2022; Dodge et al., 2021; Gebru et al., 2021). These studies mainly examine source-level, document-level, quality-level, or deduplication-level properties. In contrast, *our work audits corpora at token-level.*

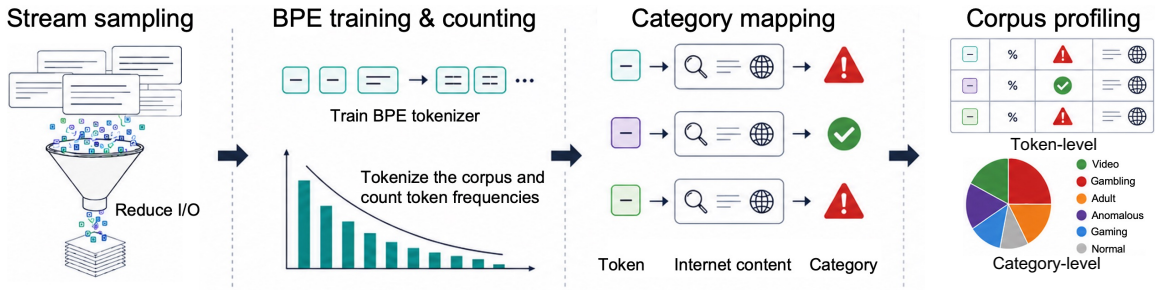


Figure 3: Conceptual overview of the SAMPLED-BPE auditing pipeline. It samples web-scale corpora, trains and counts BPE tokens, maps tokens to content categories with Internet evidence, and aggregates token statistics into corpus pollution profiles.

Chinese Web Pollution and Polluted Tokens. The research community has reported multiple cases where Chinese web pollution surfaces in LLMs. For example, it appears in GPT vocabularies: after GPT-4o was released, researchers reported that many of its longest Chinese tokens were not ordinary words, but long phrases associated with spam, pornography, gambling, or scams (Yang, 2024). Moreover, it can even appear in outputs for otherwise normal tasks: users reported Codex outputs that unexpectedly inserted Chinese gambling or lottery-like strings during coding contexts, including around `apply_patch` outputs and malformed tool-call text (OpenAI Developer Community, 2026). Recent work studies this issue through polluted tokens in LLM vocabularies, connecting their presence to possible training corpora pollution (Zhang et al., 2025). However, these observations inspect downstream models rather than upstream corpora.

3 Sampled-BPE

To enable low-cost and accurate pollution auditing of web-scale Chinese corpora, SAMPLED-BPE does not aim to reconstruct the full vocabulary. Instead, it approximates the token-level and category-level statistics of corpora.

3.1 Auditing Pipeline

Figure 3 gives a conceptual overview of the four-stage auditing pipeline.

Step 1: streaming sampling. Given a target corpus, we sample documents in a single sequential pass. Compared with random sampling which requires enumerating or shuffling document identifiers, streaming sampling makes the sampling cost nearly linear in the input scan. This system choice reduces I/O and avoids full materialization of large corpus, while yielding audit accuracy comparable to random sampling (see detailed comparison in Appendix D).

Step 2: BPE training & counting. We train a BPE tokenizer on the selected corpus and use the same tokenizer to count token frequencies (Sennrich et al., 2016). We use BPE for three reasons. *First*, because BPE is a compression method that repeatedly merges frequent adjacent units, its learned vocabulary surfaces recurrent lexical patterns that characterize the corpus. *Second*, high-frequency polluted tokens can be exposed without relying on a fixed keyword list, allowing new pollution terms to be discovered (MacAvaney et al., 2019; Jiang et al., 2021; Zhang et al., 2025). *Third*, BPE is efficient for web-scale auditing: training is driven by local adjacent-pair frequency statistics under a fixed merge budget, and counting only requires a linear pass with the learned tokenizer rather than expensive document-level semantic analysis.

Step 3: category mapping. Polluted tokens are often subtle, abbreviated, or context-dependent: the token alone may not reveal its category. Therefore, inspired by prior work on Chinese polluted-token detection (Zhang et al., 2025), we map tokens, with Internet search results as contextual evidence, into one of six content categories: *Normal Content, Adult Content, Online Gambling, Online Gaming, Online Video, and Anomalous*. To automate this step, we use GLM-4-32B, an open-source Chinese LLM with strong Chinese comprehension ability, as the base category classifier. The model is fine-tuned to predict a category from each token and its corresponding Internet search results, using a train/test split constructed from expert annotations of GPT Chinese vocabularies in Zhang et al. (2025). It achieves 97.32% classification accuracy on the test split after fine-tuning. Search-evidence settings and category-mapping robustness analyses are provided in Appendix B.

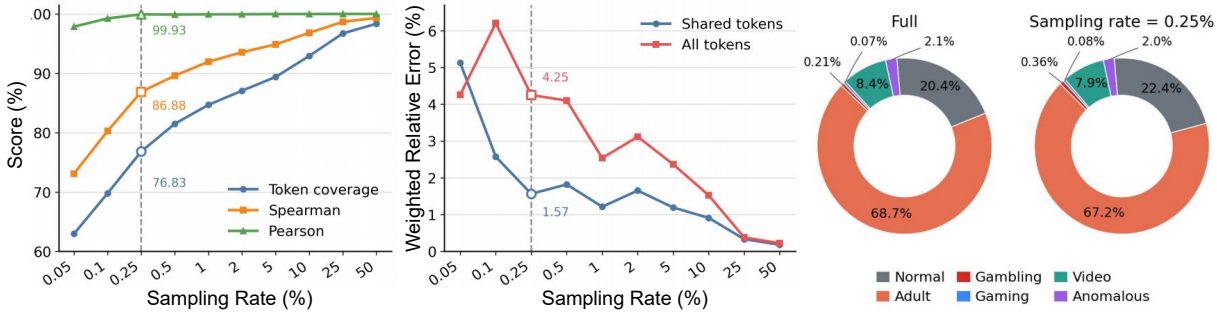


Figure 4: Sampling preserves auditing accuracy. **Left**: token-level coverage and token-ratio correlations, with the 0.25% sampling rate annotated. **Middle**: category-level weighted relative error, with the 0.25% sampling rate annotated. **Right**: category composition of 0.25% sampling rate compared with `Full`.

Step 4: corpus profiling. We combine token frequencies with category assignments to build a corpus profile. At the token level, the profile records each token's ratio, predicted category, Internet evidence, and classification rationale. This enables later analyses of shared polluted tokens, corpus-specific polluted terms, and high-frequency lexical evolution. At the category level, we aggregate token ratios into category prevalence and total pollution ratios, which are the quantities reported in later corpus-level comparisons and temporal analyses.

We refer to the unsampled corpus as `Full`. In the following experiments, each sampled corpus is audited with the same pipeline and compared against `Full` to evaluate whether sampling preserves the token-level and category-level statistics needed for pollution analysis.

3.2 Sampling Preserves Auditing Accuracy

We first ask whether SAMPLED-BPE preserves the token-level and category-level statistics needed to audit corpus pollution, that is, whether sampling gives an acceptable approximation to `Full`.

For each sampling rate, we compare the sampled corpus with `Full`. Let V_{Full} denote the token set extracted from `Full`, and let V_{sample} denote the token set extracted from a sampled corpus. At the token level, we measure token coverage as $|V_{\text{Full}} \cap V_{\text{sample}}|/|V_{\text{Full}}|$, and compute Spearman and Pearson correlations between token ratios over shared tokens $V_{\text{Full}} \cap V_{\text{sample}}$. At the category level, after mapping tokens into content categories, we compute weighted relative error both over shared tokens and over all tokens.

Figure 4 (Left) shows that sampled corpora still provide an acceptable token-level approximation to `Full` at low sampling rates. For example, at the 0.25% sampling rate, token coverage remains: the sampled corpus recovers 76.83% of `Full` tokens. More importantly, the ratios of the covered tokens closely match `Full`, with Spearman reaching 86.88% and Pearson reaching 99.93%. This suggests that ***sampling preserves an acceptable approximation of token-level statistics***.

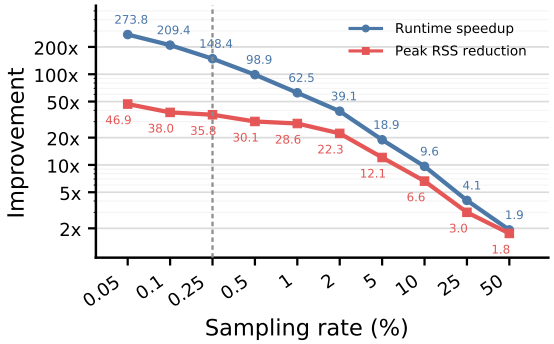


Figure 5: Runtime and memory improvement across sampling rates for SAMPLED-BPE.

Figure 4 (Middle) shows that the same trend appears after aggregating tokens into content categories. Even when token coverage is imperfect, the weighted relative error remains around 5% at the lowest sampling rate, for both shared tokens and all tokens. Figure 4 (Right) further illustrates this at the 0.25% sampling rate, where the category composition remains close to the `Full` distribution. This shows that ***sampling also preserves category-level statistics for pollution auditing***.

3.3 Sampling Reduces Auditing Cost

Having shown that sampling preserves the statistics needed for pollution auditing, we next examine its runtime speedup and memory reduction.

We measure end-to-end auditing cost as BPE training time plus token-counting time, and we report peak RSS as the memory footprint. All runs are conducted on the same machine with 96 CPU cores and 768 GB RAM.

Figure 5 shows that smaller samples sharply reduce both runtime and memory. At the 0.25% sam-

	OSCAR	mC4	HPLT	CulturaX	CWT	ROOTS	WanJuan	MAPCC	SkyPile	CCI3	WuDao
Pollution	**83.38**	**25.41**	**4.95**	**3.36**	**2.35**	**2.02**	**0.75**	**0.69**	**0.61**	**0.55**	**0.50**
Adult	76.10	0.31	1.95	0.15	0.05	0.23	0.04	0.10	0.06	0.16	0.10
Gambling	0.56	23.17	1.61	2.45	0.22	0.00	0.15	0.08	0.04	0.00	0.01
Gaming	0.16	0.46	0.25	0.16	0.29	0.00	0.17	0.09	0.05	0.01	0.06
Video	5.02	0.32	0.72	0.18	0.06	0.00	0.05	0.04	0.04	0.00	0.01
Anomalous	1.53	1.15	0.43	0.42	1.73	1.79	0.33	0.38	0.42	0.37	0.31

Table 1: Pollution ratios (%) for the Chinese portions of 11 representative open corpora, sorted by total pollution.

pling rate, the auditing pipeline achieves a 148.4× runtime speedup. The memory reduction is also substantial, with peak RSS reduced by 35.8×. Extremely small samples provide even larger savings, but Figure 4 shows that their token coverage and category estimates are less accurate. Larger samples improve accuracy, but their computational advantage decreases.

Overall, sampling makes web-scale corpus auditing ***substantially cheaper while retaining acceptable accuracy for pollution analysis***. The time cost of auditing 1TB corpus can be reduced from months to hours.

4 Pollution in Open Chinese Corpora

This section audits pollution in 11 open Chinese corpora or the Chinese portions of multilingual corpora as stated in Section 2.1. Following (Zhang et al., 2025), percentages are computed over tokens containing at least three Chinese characters over six categories. Release provenance and upstream-processing details are provided in Appendix A.

4.1 Pollution Is Widespread but Uneven

Table 1 shows that each corpus contains polluted tokens, however, the pollution ratio varies sharply, ranging from 0.50% in the cleanest corpora to 83.38%. OSCAR is the highest-pollution corpus, and mC4 is also heavily polluted at 25.41%. By contrast, WanJuan, MAPCC, SkyPile, CCI3, and WuDao all remain below 1% total pollution. This shows that ***Chinese pollution is widespread across open corpora***, but is highly uneven across corpus families and construction pipelines.

The lower pollution ratios in several corpora suggest that cleaning and curation substantially reduce pollution. Broad multilingual web pipelines such as OSCAR, mC4, HPLT, and CulturaX have the highest pollution ratios in Table 1. In contrast, corpora such as WanJuan, MAPCC, SkyPile, CCI3, and WuDao, which emphasize Chinese-specific collection, trusted sources, or quality filtering, are much cleaner. At the same time, even the lowest-pollution corpora retain measurable polluted tokens, indicating that ***cleaning helps, but does not eliminate pollution***. To support this claim, Appendix E gives two more focused comparisons between cleaned and uncleaned corpus variants.

The category rows in Table 1 show that pollution is not monolithic. OSCAR is dominated by Adult Content, which accounts for 76.10%. mC4 has a different profile: Online Gambling alone reaches 23.17%. HPLT contains both Adult Content and Online Gambling at non-trivial levels. Cleaner corpora exhibit a different residual profile: their remaining pollution is usually not dominated by a single category; instead, the largest residual component is often Anomalous, reflecting rare, peculiar, or contextually irrelevant phrases that survive quality filtering. These contrasts show that ***the composition of pollution varies across corpora***. We provide word-cloud views to show these corpus differences in Appendix I.

4.2 Pollution Is Shared, Yet Mostly Unique

Since open Chinese corpora differ sharply in both total pollution and category composition, we further investigate whether the same polluted tokens recur across corpora or whether each corpus has its own lexical artifacts.

To compare polluted tokens across the 11 representative corpus in Table 1, we define a directional coverage metric:

$$\text{Coverage}(A \rightarrow B) = \frac{|P_A \cap P_B|}{|P_A|}.$$

Here P_A and P_B denote the polluted tokens in corpora A and B.

Figure 6 shows that pairwise overlap depends on the types of corpora being compared. Cleaner or more curated Chinese corpora tend to overlap substantially with one another: MAPCC and SkyPile cover more than half of each other's polluted tokens (similar for WanJuan, MAPCC, SkyPile, CCI3, and WuDao). This suggests that ***cleaning and curation strategies still miss a shared set of residual polluted tokens***, revealing a common limitation of

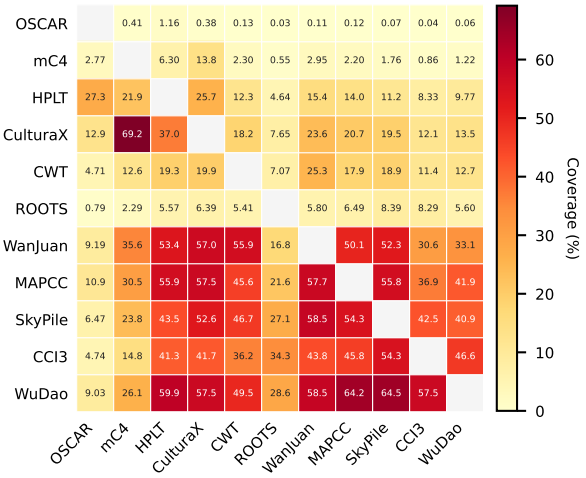


Figure 6: Pairwise coverage of polluted tokens across corpora. Each cell (omitting the diagonal) reports the percentage of polluted tokens in the row corpus that also appear in the column corpus.

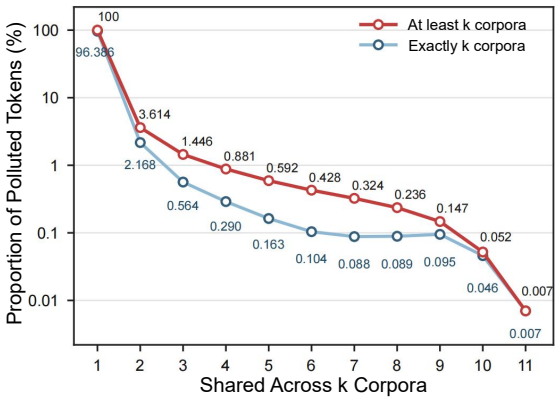


Figure 7: Proportion of polluted tokens shared across corpora. Blue line show the proportion appearing in exactly k corpora, and the red line shows the cumulative proportion appearing in at least k corpora.

existing filtering pipelines. By contrast, broader web corpora show more asymmetric containment. For example, 69.2% of CulturaX polluted tokens appear in mC4, but only 13.8% of mC4 polluted tokens appear in CulturaX. This suggests that ***some broad web corpora subsume the polluted tokens of other corpora while also introducing a much larger additional tail***.

To measure sharing beyond pairwise views, Figure 7 aggregates over all polluted tokens by counting how many corpora each token appears in. Across the 11 corpora, we observe 106,671 polluted tokens. Among them, 102,816 (96.386%) appear in only one corpus. Only 1,542 (1.446%) appear in at least three. This shows that ***most polluted tokens are corpus-specific, with a small shared core recurring across multiple datasets***.

Category	Example	Explanation
Adult	老司机 (old driver)	*Someone who knows or shares adult resources.*
Gambling	威尼斯人 (Venetian)	*One of the world's largest casino resorts.*
Gaming	王者荣耀 (Honor of Kings)	*A popular multiplayer online game.*
Video	在线观看 (watch online)	*The act of watching video content on the web.*
Anomalous	自己的小 (one's own little)	*Semantically incomplete, lacking a clear referent.*

Figure 8: Representative examples from the shared core of polluted tokens. The shared core contains tokens that appear in at least 8 of the 11 corpora in Table 1.

Figure 8 lists representative examples of the small shared core. 252 polluted tokens meet this criterion, yet they span all five polluted categories. These tokens are shared because they reflect long-lived lexical anchors. For example, "老司机 (old driver)" is a persistent euphemism in adult-resource pages, referring someone who knows or shares adult content resources; "威尼斯人 (Venetian)" is repeatedly reused as one of the world's largest casino resorts. In contrast, Appendix H lists corpus-specific polluted tokens.

Together, these results show that ***polluted tokens form a small shared core with much larger corpus-specific tails.***

5 Evolution of Chinese Web Content

The previous section shows that polluted tokens appear, to varying degrees, in mainstream open Chinese corpora. Since many of these corpora are directly or indirectly derived from Common Crawl (Common Crawl, 2011; Xue et al., 2021; Abadji et al., 2022; Nguyen et al., 2024; Penedo et al., 2023), we next ask what the upstream source looks like: ***How polluted is Chinese Common Crawl itself, and how does this pollution evolve over time?***

Common Crawl is updated monthly, with each release containing web text at terabyte scale. For this temporal case study, we audit six Chinese Common Crawl snapshots from 2021 to 2026. For each snapshot, we run the same SAMPLED-BPE auditing pipeline with 1% sampling rate, which yields lower than 3% relative error according to Figure 4. Snapshot-selection and extraction details are provided in Appendix A.

5.1 Pollution Is High and Shifting

Table 2 shows that Chinese Common Crawl contains strikingly high pollution, with Adult Content as the dominant source in several years. In 2026, to-

	2021	2022	2023	2024	2025	2026
Pollution	**35.35**	**61.87**	**62.90**	**34.83**	**38.01**	**79.52**
Adult	24.73	54.53	55.86	28.01	30.53	68.72
Gambling	5.31	1.05	0.54	1.48	0.74	0.21
Gaming	0.40	0.16	0.23	0.33	0.22	0.07
Video	3.93	4.75	4.89	3.86	5.05	8.39
Anomalous	0.98	1.38	1.37	1.16	1.47	2.12

Table 2: Temporal pollution ratios (%) in Chinese Common Crawl snapshots. Chinese web content is highly polluted, and its pollution profile shifts over time

tal pollution even reaches 79.52%, with Adult Content alone accounting for 68.72%. This indicates that pollution in the upstream Chinese web source is not a marginal tail phenomenon, but becomes comparable to or even exceeds normal content.

Besides, different pollution categories evolve in different directions. Adult Content is both high and volatile, dropping from 55.86% in 2023 to 28.01% in 2024 before rising to 68.72% in 2026. Online Gambling follows the opposite direction, decreasing from 5.31% in 2021 to 0.21% in 2026, a trend that is consistent with intensified enforcement against cross-border gambling activities targeting Chinese users[2]. Online Video gradually increases from 3.93% to 8.39%, while Anomalous rises modestly and Online Gaming remains consistently small. These trends show that ***Chinese web content is highly polluted, and its pollution profile shifts over time***. Appendix word clouds in Appendix I further visualize the changing surface vocabulary across Common Crawl snapshots.

5.2 Some Tokens Turn Over, Others Persist

We further examine the evolution of tokens. For each category, we compute Jaccard overlap of tokens for adjacent years and for the five-year span, which captures how tokens turn over and persist.

Figure 9 shows that tokens in Normal Content, Online Gaming, and Online Video are replaced less frequently over time. Their adjacent-year overlap is relatively high, and their five-year overlap also remains non-trivial. This means these categories have less year-to-year token turnover and more long-lived tokens, suggesting that *ordinary web content, online gaming, and online video are expressed through more stable forms on the web*.

By contrast, Adult Content, Online Gambling, and Anomalous show lower adjacent-year overlap and much weaker five-year persistence. The low adjacent-year overlap indicates frequent short-term

[2]https://english.scio.gov.cn/pressroom/2025-02/07/content_117699840.html

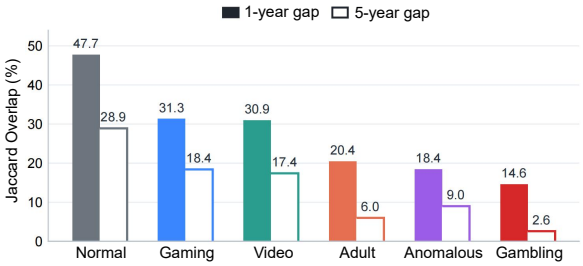


Figure 9: Jaccard overlap of tokens across years. Filled bars show average adjacent-year overlap, and hollow bars show the five-year (2021–2026) overlap.

Category	Short-lived tokens	Persistent tokens
Normal	世界杯 World Cup	联系我们 contact us
Adult	水蜜桃无码国产精品 peach uncensored domestic premium	亚洲精品 Asian premium content
Gambling	乐橙体育 Lecheng Sports	娱乐城 casino
Gaming	国际完美世界下载 International Perfect World download	王者荣耀 Honor of Kings
Video	盗墓笔记第二季 The Lost Tomb Season 2	在线观看 watch online
Anomalous	古风君子以泽 ancient-style Junzi Yize	大香蕉 big banana

Figure 10: Representative tokens illustrating turnover and persistence in Chinese Common Crawl snapshots.

token turnover, while the low five-year overlap indicates that few tokens persist over longer horizons. This may suggest that *Adult Content and Online Gambling sites continually reappear on the Chinese web and change their surface forms even after being repeatedly banned or blocked, while Anomalous reflects the fast-changing vocabulary of subcultural or irregular web expressions* (see examples in Appendix G).

Figure 10 provides representative tokens to interpret the contrast between stable and fast-changing categories. In the more stable categories such as Normal Content, Online Gaming, and Online Video, persistent tokens are reusable templates or stable entry points, such as "联系我们 (contact us)" and "王者荣耀 (Honor of Kings)". Their short-lived tokens tend to be concrete events, titles, or platform-specific phrases rather than broad changes in the category's surface form.

In the faster-changing categories, short-lived tokens more often include adult-site phrases, gambling platforms, or subcultural expressions, such as "水蜜桃无码国产精品 (peach uncensored domestic premium)", "乐橙体育 (Lecheng Sports)", and "古风君子以泽 (ancient-style Junzi Yize)". Their persistent tokens mostly capture recurring templates, such as "亚洲精品 (Asian premium content)" and "娱乐城 (casino)".

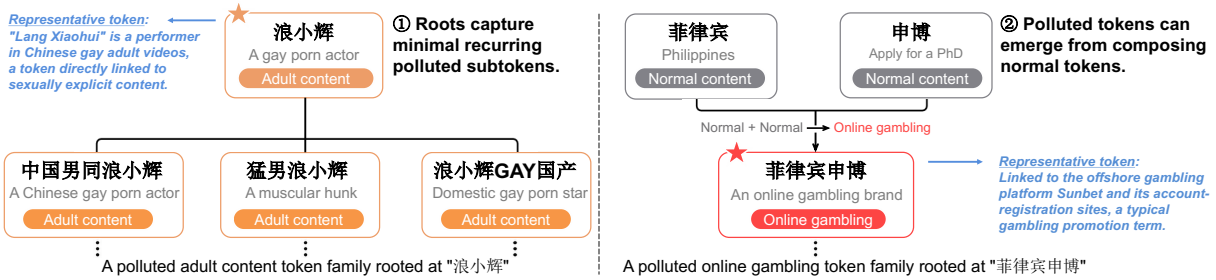


Figure 11: Examples of token trees in the hierarchical Chinese web token dataset. **Left**: roots identify minimal recurring subtokens of longer surface variants. **Right**: two normal subtokens can emerge pollution from combination.

Thus, Common Crawl's Chinese pollution exhibits both category-level temporal drift and token-level lexical renewal. Static keyword lists or one-time filtering strategies can therefore quickly become stale, while **SAMPLED-BPE** ***supports periodic auditing of upstream web corpora***.

6 Hierarchical Chinese Web Tokens

Building on our token-level audits of Chinese open corpora and Common Crawl, we release a hierarchical Chinese web token dataset with 630,684 tokens, so that these token-level findings can be reviewed, reused, and extended.

6.1 Tokens Are Organized as Trees

Chinese web tokens contain many substring relations. A short token may appear as a subtoken inside many longer tokens, and these longer tokens often form related surface variants. Instead of releasing a flat token list, we organize tokens as a forest (shown in Figure 11), consisting of 92,972 trees. ***Each tree is rooted at the shortest subtoken in the family, and its descendants are longer tokens that contain the root***.

Each token node contains Internet search evidence and a category label. Each tree further includes representative tokens and explanations for their categories. The dataset is therefore not only a collection of token strings, but also an auditable resource that records the contextual evidence and classification rationale.

6.2 Hierarchy Traces Pollution

We give two ways to trace pollution leveraging this hierarchical structure.

First, *when all tokens in a tree are from the same pollution family, the root identifies a minimal recurring subtoken*. This lets us summarize a set of surface variants with a shorter representative token, rather than treating every longer variant as an unrelated keyword. Figure 11 (Left) shows an example of "浪小辉 (A gay porn actor)" which forms a large pollution family containing long variants such as "浪小辉GAY国产".

Second, *trees with tokens from different categories reveal pollution that emerges from token composition*. A subtoken may be normal in isolation, but become polluted when combined with another normal token. Figure 11 (Right) shows an example that "菲律宾 (Philippines)" and "申博 (Apply for a PhD)" can appear as normal tokens separately, while "菲律宾申博 (An online gambling brand)" is a gambling-related polluted token. Such cases are difficult to capture because the risk emerges from the composition rather than from either subtoken alone.

Overall, the hierarchy makes the dataset more informative than a flat list of polluted tokens. ***It traces polluted token families back to minimal recurring subtoken and exposes hidden compositional cases.*** This provides evidence for future review, cleaning, and periodic auditing.

7 Conclusion

Motivated by Chinese web pollution surfacing in ChatGPT's vocabularies and Codex's outputs, this work presents SAMPLED-BPE, a lightweight token-level auditing pipeline that trains BPE tokenizers on sampled corpora to estimate Chinese corpora pollution profiles. Experiments show that SAMPLED-BPE substantially reduces runtime and memory while preserving usable estimates.

We audit 11 open Chinese corpora and 6 Chinese Common Crawl snapshots. Results show that pollution is widespread but uneven across open Chinese corpora, and upstream Chinese web content is highly polluted and temporally shifting. A hierarchical Chinese web token dataset with 630k+ token records, each with web context, category, and explanation fields, is released for reviewing and tracing Chinese web pollution.

Limitations

Pollution in other languages. This work focuses on polluted tokens in Chinese web-scale corpora and does not claim that the measured ratios transfer to other languages. Appendix C reports a diagnostic Chinese–English boundary pilot, suggesting that English auditing likely requires phrase- or sentence-level modeling and language-specific evidence rather than direct transfer of the Chinese token-level setup.

Taxonomy scope. Our six categories follow prior work on polluted Chinese tokens (Zhang et al., 2025) and are not intended as a universal data-safety taxonomy. Toxicity and political bias are usually context-dependent at the sentence level, while personally identifiable information is often compositional at the record level. Auditing these risks therefore requires complementary sentence- or document-level methods.

Readability to non-native Chinese readers. This paper necessarily includes many Chinese tokens, some of which are offensive, euphemistic, abbreviated, or context-dependent. We provide English glosses for representative examples when possible, but many tokens cannot be translated literally without losing their web-specific usage or pollution cues. Our goal is not to study Chinese linguistic expression itself, but to use Chinese web-scale corpora as a concrete case for studying upstream corpora pollution. We therefore treat translations as readability aids rather than complete semantic equivalents, and future work with broader multilingual expertise can improve these explanations.

Ethics Statement

ACL Ethics Policy is respected in this work. This work studies pollution in open Chinese corpora and upstream Chinese Common Crawl snapshots. We use publicly available corpora and web evidence for research purposes, and we respect the terms, conditions, and copyright requirements of the corresponding data sources. Because the paper analyzes web pollution, it necessarily includes examples related to adult content, gambling, spam, anomalous expressions, and other potentially offensive or upsetting material. These examples are presented only for documenting and auditing corpus pollution, and should be used with caution in future research.

We adhere to the Association for Computational Linguistics (ACL) guidelines on responsible NLP research[3], with particular attention to transparency, research-use framing, and responsible handling of harmful content.

Use of AI Assistants

The authors used AI assistants for language polishing, LaTeX editing, and phrasing suggestions during paper preparation. All substantive claims, experimental results, analyses, citations, and final text were reviewed and verified by the authors.

Acknowledgements

This work was supported by Alibaba Group through Alibaba Innovative Research Program.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355.

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.

Allen Institute for AI. 2024. C4: Colossal clean crawled corpus. https://huggingface.co/datasets/allenai/c4. Hugging Face dataset card.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jianghao Chen, Pu Jian, Tengxiao Xi, Dongyi Yi, Qianlong Du, Chenglin Ding, Guibo Zhu, Chengqing Zong, Jinqiao Wang, and Jiajun Zhang. 2023. Chinesewebtext: Large-scale high-quality chinese web text extracted with effective evaluation model. *arXiv preprint arXiv:2311.01149*.

Common Crawl. 2011. Common crawl. https://commoncrawl.org/.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 1286–1305.

[3] https://aclrollingreview.org/responsibleNLPresearch/

Xinrun Du, Zhouliang Yu, Songyang Gao, Ding Pan, Yuyang Cheng, Ziyang Ma, Ruibin Yuan, Xingwei Qu, Jiaheng Liu, Tianyu Zheng, Xinchen Luo, Guorui Zhou, Wenhu Chen, and Ge Zhang. 2024a. Chinese tiny llm: Pretraining a chinese-centric large language model. *Preprint*, arXiv:2404.04167.

Xinrun Du, Zhouliang Yu, Songyang Gao, Ding Pan, Yuyang Cheng, Ziyang Ma, Ruibin Yuan, Xingwei Qu, Jiaheng Liu, Tianyu Zheng, et al. 2024b. Chinese tiny llm: Pretraining a chinese-centric large language model. *arXiv preprint arXiv:2404.04167*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Nick Hagar and Jack Bandy. 2025. Practical datasets for analyzing llm corpora derived from common crawl. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 2454–2464.

Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. 2023. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models. *arXiv preprint arXiv:2308.10755*.

High Performance Language Technologies Project. 2025. HPLT monolingual datasets 3.0. https://hplt-project.org/datasets/v3.0. Dataset catalogue.

Menghan Jiang, Xiang Ying Shen, Kathleen Ahrens, and Chu-Ren Huang. 2021. Neologisms are epidemic: Modeling the life cycle of neologisms in china 2008-2016. *PloS one*, 16(2):e0245984.

Michael Karydis and Dimitra Kitsiou. 2013. Marine water quality monitoring: A review. *Marine pollution bulletin*, 77(1-2):23–36.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan Van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.

Jeffrey Li, Mohammadreza Armandpour, Seyed Iman Mirzadeh, Sachin Mehta, Vaishaal Shankar, Raviteja Vemulapalli, Samy Bengio, Oncel Tuzel, Mehrdad Farajtabar, Hadi Pouransari, et al. 2025. Tic-lm: A web-scale benchmark for time-continual llm pretraining. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32231–32273.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International conference on machine learning*, pages 22631–22648. PMLR.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, et al. 2025. Hplt 3.0: Very large-scale multilingual resources for llm and mt. mono-and bi-lingual data, multilingual evaluation, and pre-trained models. *arXiv preprint arXiv:2511.01066*.

OpenAI Developer Community. 2026. Chinese gambling characters in Codex CLI message and code output? OpenAI Developer Community. Forum thread, first posted January 27, 2026.

OSCAR Project. 2023. OSCAR-2301-HPC. https://huggingface.co/datasets/oscar-corpus/oscar-2301-hpc. Hugging Face dataset card.

PaddleNLP Contributors. 2025. WuDaoCorpus2.0 base corpus. https://paddlenlp.readthedocs.io/en/latest/llm/tools/preprocess/docs/WuDaoCorpusBase.html. Documentation page.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Sasha Luccioni, Yacine Jernite, and Anna Rogers. 2023. The roots search tool: Data transparency for llms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 304–314.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Szymon Roziewski and Wojciech Stokowiec. 2016. Languagecrawl: A generic tool for building language models upon common-crawl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2789–2793.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1715–1725.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355.

Liangdong Wang, Bo-Wen Zhang, Chengwei Wu, Hanyu Zhao, Xiaofeng Shi, Shuhao Gu, Jijie Li, Quanyue Ma, TengFei Pan, and Guang Liu. 2024. Cci3. 0-hq: a large-scale chinese dataset of high quality designed for pre-training large language models. *arXiv preprint arXiv:2410.18505*.

Maurice Weber, Daniel Y Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, et al. 2024. Redpajama: an open dataset for training large language models. *Advances in neural information processing systems*, 37:116462–116492.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, et al. 2023. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Hao Xu, Jiacheng Liu, Yejin Choi, Noah A Smith, and Hannaneh Hajishirzi. 2025. Infini-gram mini: Exact n-gram search at the internet scale with fm-index. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24955–24980.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 483–498.

Zeyi Yang. 2024. GPT-4o's Chinese token-training data is polluted by spam and porn websites. MIT Technology Review. Published May 17, 2024.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. 2021. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *Ai Open*, 2:65–68.

Qingjie Zhang, Di Wang, Haoting Qian, Liu Yan, Tianwei Zhang, Ke Xu, Qi Li, Minlie Huang, Hewu Li, and Han Qiu. 2025. Speculating llms' chinese training data pollution from their tokens. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26124–26144.

A Details of Open Chinese Corpora

Following Section 2.1, this section summarizes Common Crawl and the 11 representative open Chinese corpora audited in our experiments. For each corpus, we describe its source, public scale, audited Chinese portion when available, and relation to LLM training when the release is explicitly tied to a model. These details provide context for the corpus landscape in Figure 2 and the pollution ratios in Table 1.

Table 3 summarizes the audited releases and their upstream processing.

CC (Common Crawl Foundation, founded in 2007; data collected since 2008): Common Crawl is an open web-crawl archive maintained by a nonprofit organization. It has collected public web data since 2008 and, according to the official overview, now contains more than 10 PiB of archived data, with crawls published approximately once a month and each crawl typically containing more than two billion web pages. Common Crawl provides raw HTTP responses in WARC files, extracted metadata in WAT files, and extracted plaintext in WET files, making it the upstream source for many web-derived training corpora. In our WET-file accounting, each monthly crawl is approximately 6 TB compressed and 20 TB decompressed. Because Common Crawl does not publish official Chinese-only sizes, we estimate the language distribution by reading the WET header `WARC-Identified-Content-Language` and counting the first tag as the record's primary language; English accounts for 45.14% (about 9 TB decompressed) and Chinese for 5.19% (about 1 TB decompressed), as shown in Figure 12. We audit Chinese WET-derived snapshots from 2021–2026 by selecting the February crawl in each year and sampling 1% of the Chinese content, about 10 GB per crawl. Filtered or processed Common Crawl data has been used directly or indirectly in major models and corpora, including T5 through C4, GPT-3 through filtered Common Crawl, BLOOM through OSCAR/ROOTS, and Falcon through RefinedWeb (Common Crawl, 2011; Raffel et al., 2020; Workshop et al., 2022; Penedo et al., 2023).

OSCAR (OSCAR project, Inria/University of Mannheim/DFKI, 2023): OSCAR-2301-HPC is a document-oriented multilingual web corpus derived from Common Crawl WET files with the Ungoliant pipeline. The public OSCAR-2301-HPC table reports a Chinese portion of 138.48M documents, 44.38B space-separated words, and 1.4TB of content. OSCAR is intended as raw multilingual pretraining data; earlier OSCAR releases were also a major source in ROOTS, the corpus used to train BLOOM (Abadji et al., 2022; OSCAR Project, 2023; Laurençon et al., 2022; Workshop et al., 2022).

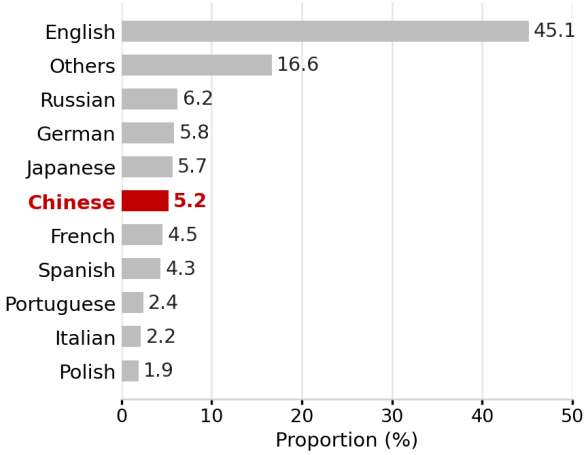


Figure 12: Language distribution in our Common Crawl WET accounting.

mC4 (Google Research, 2021): Multilingual C4 is the multilingual variant of C4, built by cleaning Common Crawl text and extending the C4 filtering recipe to many languages. The public C4 release lists mC4 as a 9.7TB multilingual collection, and the Chinese files audited in this work total 71.42GB. mC4 was introduced as the pretraining corpus for mT5, a multilingual text-to-text Transformer covering 101 languages (Xue et al., 2021; Allen Institute for AI, 2024).

HPLT (European HPLT initiative, 2025): HPLT 3.0 is a very large multilingual web corpus built from Internet Archive and Common Crawl data from 2012–2024. HPLT 3.0 applies HTML extraction, language identification, deduplication, quality scoring, metadata annotation, and filtering; its Simplified Chinese Mandarin portion, `cmn_Hans`, is reported as 5.40TB, 2.21B documents, 2.97T tokens, and 4.14T characters. The release is designed as a broad LLM and MT resource and is accompanied by HPLT-trained reference models and multilingual evaluations (Oepen et al., 2025; High Performance Language Technologies Project, 2025).

CulturaX (UONLP and collaborators, 2024): A cleaned multilingual dataset constructed by combining mC4 and OSCAR releases, followed by language identification, URL filtering, metric-based

Corpus	Audited release/subset	Upstream processing status	Our audit protocol
HPLT	HPLT 3.0 `cmn_Hans`	Extraction, LID, deduplication, quality scoring, metadata annotation, filtering	Audit the public Chinese portion
Common Crawl	February crawls 2021–2026; Chinese WET-derived records	Raw WET text; no single corpus-level cleaned release	Same Chinese extraction + 1% sampling protocol across years; supports temporal comparison
OSCAR	OSCAR-2301-HPC Chinese portion	Derived from Common Crawl WET via Ungoliant	Audit the public Chinese portion
mC4	mC4 Chinese files	C4-style multilingual cleaning/filtering	Audit the public Chinese files
CulturaX	CulturaX Chinese partition	mC4/OSCAR combination; LID, URL filtering, metric cleaning, document refinement, MinHash deduplication	Audit the public Chinese files
CWT	ChineseWebText full release; cleaner subset used in appendix comparison	EvalWeb deduplication, LID, rule filtering, BERT quality scoring; cleaner subset uses $>$ 90 quality score	Audit the public full/cleaner variants
ROOTS	ROOTS Chinese language group / audited Chinese slice	Curated multilingual corpus from OSCAR and documented sources	Audit the Chinese slice, with official denominator noted
WanJuan	Intern-WanJuan 1.0 text portion	Fine-grained cleaning, deduplication, safety-oriented filtering	Audit the public text release
MAPCC	MAP-CC `zh-cc` portion	Chinese-centric corpus including Common Crawl and curated sources	Audit the `zh-cc` portion
SkyPile	SkyPile-150B public plain-text portion	Filtering, deduplication, sensitive-data filtering, low-quality filtering	Audit the public plain-text portion
CCI3	CCI3-HQ public files	Trusted Chinese Internet data; two-stage hybrid filtering	Audit the public high-quality subset
WuDao	WuDaoCorpus2.0 Base public subset	High-density page extraction; rule-based cleaning	Audit the public base subset

Table 3: Corpus provenance and upstream processing for audited corpora.

cleaning, document refinement, and MinHash deduplication. The full corpus contains 6.3T tokens in 167 languages; the Chinese partition contains 218.62M documents and 227.06B tokens, and the public Chinese files used for our audit total 641.02GB. CulturaX is released as reusable multilingual LLM training data rather than as a dataset paired with one flagship model (Nguyen et al., 2024).

CWT (CASIA-LM, 2023): ChineseWebText is a Chinese web corpus extracted with the EvalWeb toolchain from noisy Common Crawl data. EvalWeb combines deduplication, language identification, rule-based filtering, and a BERT-based quality evaluation model, and assigns a quality score to each text. The official release contains 1.42TB of scored Chinese web text and a cleaner 600GB subset with quality scores above 90%, intended for Chinese LLM pretraining and threshold-based data selection (Chen et al., 2023).

ROOTS (BigScience, 2022): The Responsible Open-science Open-collaboration Text Sources corpus is a 1.6TB multilingual collection spanning 59 languages, assembled from OSCAR and manually documented sources. The ROOTS Search Tool reports 259.01GB for the Chinese language group used as our official Chinese-size denominator. ROOTS was the training corpus for the 176B-parameter BLOOM multilingual language model (Laurençon et al., 2022; Piktus et al., 2023; Workshop et al., 2022).

WanJuan (Shanghai AI Laboratory/OpenDataLab, 2023): The text portion of Intern-WanJuan 1.0 is a cleaned pretraining corpus with more than 500M documents and more than 1TB of data. It integrates web pages, encyclopedias, books, patents, textbooks, and exam questions, converting heterogeneous HTML, text, PDF, and EPUB sources into a unified JSONL format after fine-grained cleaning, deduplication, and safety-oriented filtering. WanJuan was used in the training of InternLM (He et al., 2023).

MAPCC (M-A-P/CT-LLM team, 2024): MAP-CC is a Chinese-centric pretraining corpus released with the Chinese Tiny LLM project. MAP-CC contains about 800B Chinese tokens and is organized into components such as `zh-cc` from Chinese Common Crawl, encyclopedic data, papers, books, and other sources; the `zh-cc` portion audited here has a public size of 1.408TB. MAP-CC serves as the main Chinese pretraining data for CT-LLM, a 2B Chinese-centric language model trained together with English and code data (Du et al., 2024b,a).

SkyPile (Skywork, 2023): SkyPile-150B is a large-scale Chinese web dataset for pretraining language models. The public portion contains about 233M unique web pages, around 150B tokens, and 620GB of plain text. Its construction applies filtering, deduplication, sensitive-data filtering, and low-quality filtering with tools such as fastText and BERT. The release is associated with the Skywork bilingual foundation-model effort (Wei et al., 2023).

CCI3 (BAAI, 2024): CCI3-HQ is a high-quality subset of Chinese Corpora Internet 3.0 designed for LLM pretraining. The release is described as

an approximately 500GB Chinese corpus, and the public files used for our audit total 518GB. It is produced from trusted Chinese Internet data using a two-stage hybrid filtering pipeline. The technical report evaluates the dataset by training a 0.5B model from scratch on 100B tokens and comparing against CCI3.0, SkyPile, and WanJuanV1 (Wang et al., 2024).

WuDao (Beijing Academy of Artificial Intelligence, 2021): WuDaoCorpus2.0 Base is the public base subset of WuDaoCorpora, a large-scale Chinese corpus collected for pretraining. The full WuDaoCorpora is described as a 3TB Chinese corpus, while the publicly released WuDaoCorpus2.0 Base contains 200GB of plain text. It is derived from large-scale Chinese web pages and other sources, with text extracted from high-density pages and cleaned through a rule-based pipeline. It is commonly used as a Chinese pretraining baseline and is associated with the WuDao large-model line (Yuan et al., 2021; PaddleNLP Contributors, 2025).

B Robustness of Category Mapping

We test whether category mapping depends on a particular search configuration or classifier backbone. All variants are evaluated on the same expert-annotated held-out split: Table 4 compares search backends and regions, while Table 5 compares classifier backbones.

Search setting	Held-out accuracy (%)
Google / China region	97.32
Baidu API	96.30
Google / USA region	96.56
Google / Japan region	96.81
Google / Germany region	96.68
Google / Brazil region	96.30

Table 4: Robustness across search backends and regions.

LLM backbone	Held-out accuracy (%)
GLM-4-32B	97.32
Qwen-2.5-72B	95.64

Table 5: Robustness across classifier backbones.

As Table 4 shows, the search variants remain within 1.02 percentage points of the Google/China setting. Table 5 further shows that Qwen-2.5-72B remains close to GLM-4-32B. Thus, mapping accuracy is not tied to one tested search configuration or backbone.

To measure how residual classification errors affect corpus-level estimates, we perturb token labels at the 2.68% held-out error rate and recompute total pollution for each Common Crawl snapshot. The resulting estimates are reported in Table 6.

Year	Original	Perturbed	Δ (pp)
2021	35.35	36.89	+1.54
2022	61.87	62.56	+0.69
2023	62.90	63.56	+0.66
2024	34.83	36.38	+1.55
2025	38.01	39.46	+1.45
2026	79.52	79.64	+0.12

Table 6: Impact of 2.68% label perturbation on total pollution ratios (%). Δ is the absolute change from Table 2 in percentage points.

As Table 6 shows, the largest change is 1.55 percentage points, leaving the corpus-level scale and temporal conclusions unchanged.

C Chinese–English Comparison

To examine whether the same token-level discovery setup transfers beyond Chinese, we compare the full vocabularies of 150K-token BPE models trained on Chinese and English Common Crawl snapshots. As Table 7 shows, Chinese BPE directly surfaces complete polluted phrases and templates, whereas English BPE mainly surfaces isolated words or domain fragments.

Chinese	**Count: 76,404**
Examples	"成人av在线" ("adult AV online"); "无码免费视频" ("uncensored free video"); "線上娛樂城" ("online casino")
BPE unit	Complete polluted phrases or templates
English	**Count: 413**
Examples	`porn`; `casino`; `NakedCams`
BPE unit	Isolated words or no-space domain fragments, rather than complete multi-word pollution units

Table 7: Diagnostic comparison of the units surfaced by Chinese and English BPE vocabularies. Counts follow the language-specific checks used in this pilot.

The smaller English count should not be interpreted as lower pollution prevalence. It instead marks a methodological boundary: English pollution is often expressed as multi-word phrases and therefore requires phrase- or sentence-level discovery rather than direct transfer of the Chinese token-level setup.

D Streaming vs. Random Sampling

Following Section 3.1, this section compares the prefix-style streaming sampler used in SAMPLED-BPE with an index-based random sampler on the same CC reference setting. The random sampler first materializes a shuffled index set and then reads the selected rows, whereas the streaming sampler reads a contiguous prefix and can stop once the requested sample size has been reached. We use the 0.10% sample size relative to the 30% CC reference as a representative setting. In Table 8, arrows in the Streaming row show relative changes from Random.

	Time Cost	Token Coverage	Category WRE
Random	1731.7s	70.19%	2.65%
Streaming	77.4s ▼22.4×	69.79% ▼0.57%	2.57% ▼2.69%

Table 8: Streaming and random sampling comparison on CC at 0.10% sampling rate.

Table 8 shows that streaming sampling substantially reduces sampling time while preserving audit quality. Compared with random sampling, streaming sampling is 22.4× faster, while token coverage changes only from 70.19% to 69.79%. The category weighted relative error is also comparable, decreasing slightly from 2.65% to 2.57%. These results support our use of streaming sampling as a low-cost system choice for web-scale auditing.

E Cleaning Reduces but Does Not Eliminate Pollution

Following Section 4, this section provides focused comparisons supporting the claim that cleaning and curation reduce, but do not eliminate, Chinese web pollution. We compare ChineseWebText full and cleaner variants, and two ROOTS Chinese variants with different source and curation composition. Values follow the same normalization as Table 1.

	ChineseWebText		ROOTS	
	Full	Cleaner	Uncorpus	Public
Pollution	**2.35**	**0.50**	**2.02**	**1.37**
Adult	0.05	0.05	0.23	0.19
Gambling	0.22	0.02	0.00	0.01
Gaming	0.29	0.05	0.00	0.01
Video	0.06	0.02	0.00	0.00
Anomalous	1.73	0.37	1.79	1.16

Table 9: Pollution ratios (%) for cleaned and less-cleaned Chinese corpus variants.

Table 9 shows that cleaning sharply lowers pollution in ChineseWebText: total pollution drops from 2.35% in CWT-Full to 0.50% in CWT-Cleaner. The reduction is especially clear for Anomalous content, which decreases from 1.73% to 0.37%, and Online Gambling, which decreases from 0.22% to 0.02%. ROOTS shows a similar but weaker pattern: ROOTS-Public has lower total pollution than ROOTS-Uncorpus (1.37% vs. 2.02%), but still retains measurable residual pollution, mainly from Anomalous and Adult Content. These comparisons support the main-text claim that cleaning helps, but does not remove polluted tokens entirely.

F Sampling Rates of Chinese Corpora

Following Section 4, this section reports the actual sampling rates used for auditing the 11 open Chinese corpora and the corresponding full-token weighted relative errors estimated from Figure 4. Sampling rates use official public corpus-size denominators where available. Weighted errors are obtained by log-linear interpolation over the full-token category weighted-error curve.

For ROOTS, the denominator is the complete Chinese ROOTS size, not only the audited `roots_zh_uncorpus` slice. For WanJuan, the official denominator is reported only as larger than 1TB, so the sampling rate is an upper bound and the error is interpolated at 0.53%.

Table 10 shows that we use small sampling rates for most corpora, ranging from 0.19% for HPLT to 2.03% for CCI3, with mC4 as an exception because the audited Chinese slice is much smaller. Despite these low sampling rates, the estimated full-token weighted relative errors remain small: all are below 4%, and most fall between 2% and 3%. This supports the result that SAMPLED-BPE can provide usable corpus pollution estimates without full-corpus auditing.

G Additional Token Evolution Analyses

Following Section 5, this section supplements the main all-token temporal analysis with high-frequency top-100 token views and a top-k sensitivity check. We focus on token turnover and persistence in Chinese Common Crawl snapshots, using two complementary signals: Jaccard overlap across time gaps and the average number of yearly entrants in the top-100 list.

Figure 13 shows a clear contrast between stable and fast-changing categories. Online Video has the highest adjacent-year overlap and the fewest yearly entrants, indicating that its high-frequency

Corpus	OSCAR	mC4	HPLT	CulturaX	CWT	ROOTS	WanJuan	MAPCC	SkyPile	CCI3	WuDao
Sampling rate	1.53%	14.00%	0.19%	0.94%	0.76%	0.46%	0.53%	0.36%	1.94%	2.03%	1.34%
Weighted relative error	2.34%	0.24%	3.63%	2.75%	2.93%	2.90%	3.02%	2.68%	2.06%	2.00%	2.46%

Table 10: Sampling rates and estimated full-token weighted relative errors for auditing the 11 open Chinese corpora.

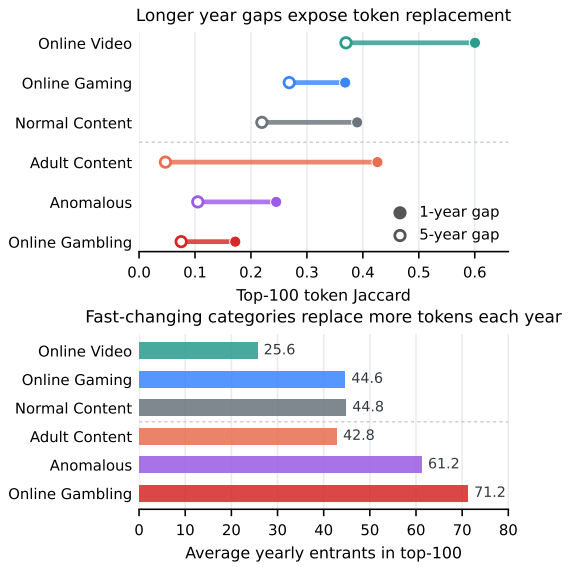


Figure 13: Additional top-100 token evolution signals across Chinese Common Crawl snapshots. **Top**: adjacent-year and five-year Jaccard overlap. **Bottom**: average yearly entrants in the top-100 lists.

token head is comparatively stable. Online Gambling shows the opposite pattern: it has the lowest overlap and the largest number of yearly entrants, indicating rapid replacement of high-frequency tokens. Anomalous also shows substantial turnover, with many new top-100 tokens entering each year.

Adult Content is more nuanced. Its adjacent-year top-token Jaccard remains moderate, but its five-year overlap is much lower, suggesting that this category combines short-term template stability with longer-term lexical replacement. This explains why Adult Content still shows substantial temporal drift even when recurring fragments such as "国产精品 (domestic premium content)", "精品视频 (premium videos)", and "精品国产 (domestic premium series)" remain visible across snapshots.

Sensitivity to the top-k threshold. We further repeat the lexical-evolution analysis with top-10, top-100, and full category vocabularies to test whether the conclusions depend on the truncation threshold. The main pattern is that smaller k emphasizes stability in the lexical head, whereas the full vocabulary exposes volatility in the long tail. This effect is especially strong for Online Video, whose adjacent-year Jaccard drops from 0.76 at top-10 to 0.60 at top-100 and then to 0.31 over the full vocabulary. Adult Content and Anomalous show similar declines, from 0.53 to 0.43 to 0.20 and from 0.49 to 0.25 to 0.18, respectively. In contrast, Normal Content remains relatively stable across thresholds, with adjacent-year Jaccard of 0.39 at top-10, 0.39 at top-100, and 0.48 over the full vocabulary. Online Gambling is unstable at every scale, with adjacent-year Jaccard staying low from 0.24 to 0.17 and 0.15. Overall, these ablations motivate the main-text all-token view when the goal is to expose long-tail lexical renewal, while the top-100 analysis keeps the high-frequency patterns interpretable.

Examples of short-lived and long-lived tokens.

- **Normal Content.** Short-lived examples include "影片名称 (film title)", "世界杯 (World Cup)", and "我欲封天 (I Shall Seal the Heavens)", which spike in a single year; long-lived examples include "中文字幕 (Chinese subtitles)", "联系我们 (contact us)", "有限公司 (limited company)", and "关于我们 (about us)", which persist across most or all snapshots. ***This category is comparatively stable overall.***
- **Adult Content.** Short-lived examples include "日本一级特黄大片 (Japanese explicit-content phrase)", "亚洲成人 (Asian adult)", "亚洲天堂 (Asian paradise)", and "一区二区在线观看 (Zone 1/2 watch online)"; long-lived examples include "国产精品 (domestic premium content)", "精品视频 (premium videos)", "精品国产 (domestic premium series)", and "久久久久久 (repeated jiu-jiu template)", reflecting a stable template-like core mixed with rapid surface turnover. ***The key pattern is stable templates plus changing surface realizations.***
- **Online Gambling.** Short-lived examples include "太阳城 (Sun City)", "乐橙体育 (Lecheng Sports)", "申博太阳城 (Shenbo Sun City)", and "云顶集团手机版 (Genting Group mobile version)"; long-lived examples include "开奖结果 (lottery results)", "双色球 (Double Color Ball)", "娱乐城 (casino)", "免费一级 (free first-level content)", and "澳门正版资料

(Macau official materials)", showing recurring gambling-related templates but rapid platform and brand replacement. ***This is the clearest case of rapid lexical turnover.***

- **Online Video.** Short-lived examples include "盗墓笔记第二季 (The Lost Tomb Season 2)", "完美世界txt下载 (Perfect World TXT download)", and "网址在线观看 (watch online at URL)"; long-lived examples include "在线观看 (watch online)", "在线播放 (stream online)", "在线视频 (online video)", "免费视频 (free video)", and "免费观看 (watch for free)", consistent with the high stability of generic video-access phrases. ***This is the most stable category.***
- **Anomalous.** Short-lived examples include "小说零 (novel zero)", "一二 (Zone 1/Zone 2)", "五月婷 (May Ting)", and "最高占成 (highest commission share)"; long-lived examples include "一区二区三区 (Zones 1/2/3)", "一区二区 (Zones 1/2)", "一二三区 (Zones 1/2/3)", "大香蕉 (big banana)", and "卡三卡 (card-three-card)", illustrating both recurring peculiar phrases and rapidly changing surface variants. ***This category combines contextually irrelevant recurring forms with strong year-to-year drift.***
- **Online Gaming.** Short-lived examples include "完美世界前传下载 (Perfect World prequel download)", "国际完美世界下载 (International Perfect World download)", and "小说改编的网页游戏 (web games adapted from novels)"; long-lived examples include "破解版 (cracked version)", "小游戏 (mini games)", "游戏平台 (game platform)", "王者荣耀 (Honor of Kings)", and "游戏下载 (game download)", indicating a more persistent gaming vocabulary than Gambling or Anomalous. ***Its stability is closer to mainstream content than to Gambling or Anomalous.***

H Corpus-Specific Polluted Tokens

Following Section 4.2, this section complements the shared-core analysis by showing representative polluted tokens that are unique to one audited corpus slice. These examples are grouped by predicted category.

Figure 14 shows that corpus-specific polluted tokens are not limited to a single pollution type. Different corpus slices contain distinct Adult Content, Online Gambling, Online Gaming, Online Video, and Anomalous tokens, reflecting source-specific and pipeline-specific lexical artifacts. This provides qualitative support that most polluted tokens form corpus-specific tails rather than a shared vocabulary across all corpora.

I Token Clouds

Following Section 4 and Section 5, this section provides qualitative token-cloud views of high-frequency Chinese tokens in the audited open corpora and Chinese Common Crawl snapshots. For each corpus, we read long Chinese tokens (more than 2 Chinese characters), use token-count statistics to rank tokens, and visualize the top-2,000 tokens. Font size is determined by count ranking. These token clouds visualize surface forms rather than category ratios, providing qualitative context for Table 1 and Table 2.

Figure 15 shows that the six representative broad-web corpus slices differ sharply in their high-frequency surface forms. OSCAR is visually dominated by video and adult-access templates, whereas mC4 contains many lottery- or betting-like phrases. HPLT, CulturaX, CWT, and ROOTS show more generic web, institutional, or template-like text. This qualitative contrast aligns with the main result that pollution is widespread but uneven across corpus families, and that different corpora exhibit different dominant pollution profiles.

Figure 16 shows a different pattern for curated Chinese corpora. Their top tokens are mostly generic content or web-template phrases, consistent with their much lower pollution ratios in Table 1. However, because token clouds only visualize the highest-count tokens, they should not be read as evidence that these corpora are pollution-free. Rather, they show that in cleaner corpora, residual polluted tokens are less visually dominant and often remain in a lower-frequency tail.

Figure 17 provides a temporal view of Common Crawl. The dominant high-frequency tokens change across yearly snapshots, with visible shifts among video-access phrases, site templates, and adult-content templates. This supports the temporal analysis in Section 5: Chinese Common Crawl pollution is not only high in aggregate, but also changes in its surface vocabulary over time.

	Adult	Gambling	Gaming	Video	Anomalous
OSCAR	久久精品国产 Jiujiu premium domestic 性欧美 European porn 禁网站 banned site	免费一级 free first-level 年买特马最准网站 most accurate site for buying special number picks 澳门特马今晚开奖 tonight's Macau special number draw	完美世界前传下载 Perfect World prequel download 国际完美世界下载 International Perfect World download 完美世界国际版下载 Perfect World International version download	盗墓笔记第二季 The Lost Tomb Season 2 完美世界txt下载 Perfect World TXT download 最近中文字幕 recent Chinese subtitles	小说零 zero novels 亚洲人成色 Asian skin tone 人禽杂交 human-animal hybridization
mC4	免费黄片视频在线观看 free online porn videos 一级A做爱片免费视频 free X-rated adult film videos 凯发直播 Kaifa live streaming	搏天堂 Bo heaven 凯发官网 Kaifa official website 凯发官网平台 Kaifa official platform	真人游戏下载手机版 mobile download for live-action games 天堂官方网站 Heaven official website 网络真人游戏 online live-action games	真人游戏在线 live-action games online 在线快 online fast 娱乐官网免费下载 free download from entertainment official site	鎏焯檬只 liu chao meng zhi 环亚官网 Huanya official website 计划大师苗军 Plan Master i Miaojun
HPLT	韩国女主播 Korean female streamer 人 person 撸一撸 masturbate	重彩 chongcai pc蛋蛋幸运 PC egg lucky 分分彩 minute-by-minute lottery	新开传奇私服 newly opened Legend private server 迷失传奇 Lost Legend 迷失传奇私服 Lost Legend private server	影音先 audio-video first 在播放 currently playing 在看 watching	圣王中文 Shengwang chinese 张布施 Zhang Bushi 第一王风 first Wang Feng
CulturaX	直播 live streaming 性增长 sexual growth 污小说 dirty novels	澳门百家 Macau hundreds 万达娱乐平台 Wanda entertainment platform 百家 hundreds	游戏游戏 game game 电竞下载 esports download 完美世界电竞app Perfect World Esports app	电竞比分直播 live esports scores 依法开展互联网视听节目服务 legally provide internet audiovisual program services 弘扬民族优秀文化传统的互联网视听节目 internet audiovisual programs promoting outstanding ethnic cultural traditions	好织梦 good Zhimeng 企和 qi he 建和 jian he
CWT	n大学生 n college students n生活中 n in daily life n前列腺 n prostate	体育官方app下载 official sports app download n爱游戏 n love gaming 开云体育app Kaiyun Sports app	n王者荣耀 n Honor of Kings n全民奇迹 n Quanmin Qiji n阴阳师 n Onmyoji	全场录像 full-game recording 全场录像回放 full-game replay n短视频 n short videos	n上一篇 n previous post n上一条 n previous item n下一条 n next item
ROOTS	性虐待和 sexual abuse and 性和透明度 sex and transparency 性旅游 sex tourism	捕鱼活动 fishing-game activities 可赔区 payout area	放置武器 placing weapons 格查尔 Quetzal 对消除 match-elimination	联合国网播 UN Web TV 和录像 and recordings 下载今天的 download today's	肯尼亚和 Kenya and 议员和 members of parliament and 国际贸易和 international trade and
WanJuan	性站点 adult sites 性好的 sex is okay 轮过后 after the rounds	好彩网帝 good lottery web emperor 澳门游戏网站 Macau gaming website 抽奖机会 lottery-draw chance	策略手游 strategy mobile games AG超玩会 AG Super Play 仙侠手游 fantasy cultivation mobile games	视频加载中 video loading 视频中心 video center 免费足球 free football	向天岚 Xiang Tianlan 空包网 empty-package site 牧玄瑜 Mu Xuanyu
MAPCC	用道具 use props 窥视卡 peeping card 原色地带 primary-color zone	六彩 six-color lottery 通吃岛证券网 take-all island securities site 六合采 Mark Six	暗夜精灵族 Night Elf race 网游之 Online Game: 泡泡堂 Crazy Arcade	下载点点通 download Diandiantong 影视速递 film and TV express 点击观看 click to watch	只看该作者大 view only this author big 浏览过此新闻的网友还阅读了以下新闻人 readers of this news also read the following news person 此主题相关图片如下 related images for this topic are as follows
SkyPile	性消费 sexual consumption 性和准确性 sex and accuracy A级以上 above grade A	雷火网址 Leihuo website 扑克牌游戏 poker-card games 排列三第 arrangement three number	创意小游戏 creative mini-games 外挂软件 cheat software 小游戏的 mini-game's	电子竞技游戏开播量 esports-game livestream starts 视频传输 video transmission Bob直播间 Bob livestream room	网通社从 netcom news agency from 损失以 loss by 日起任 ri qi ren
CCI3	性特征 sexual characteristics 性和创造性 sexuality and creativity 和学生的 with the student's	——	团大战 regiment battle 云十六 cloud sixteen	看动画片 watch cartoons	了学生的 le student's 了学生 le student 股易舒 Guyishu
WuDao	阳痿早泄 erectile dysfunction and premature ejaculation 男忄生问题 male problems 男忄生 male	多万字 many ten-thousand characters	游戏大小 game size 游戏加载 game loading 游戏加载完毕点击 click after game loading completes	视频接口 video interface	全村有 the whole village has 重梦境神话 heavy dream mythology 施绮 Shi Qi

Figure 14: Representative dataset-specific polluted tokens, with English glosses. Each cell lists high-ratio tokens that appear in only one of the 11 representative corpus slices, grouped by predicted category.

Figure 15: Token clouds for six representative broad-web corpus slices.

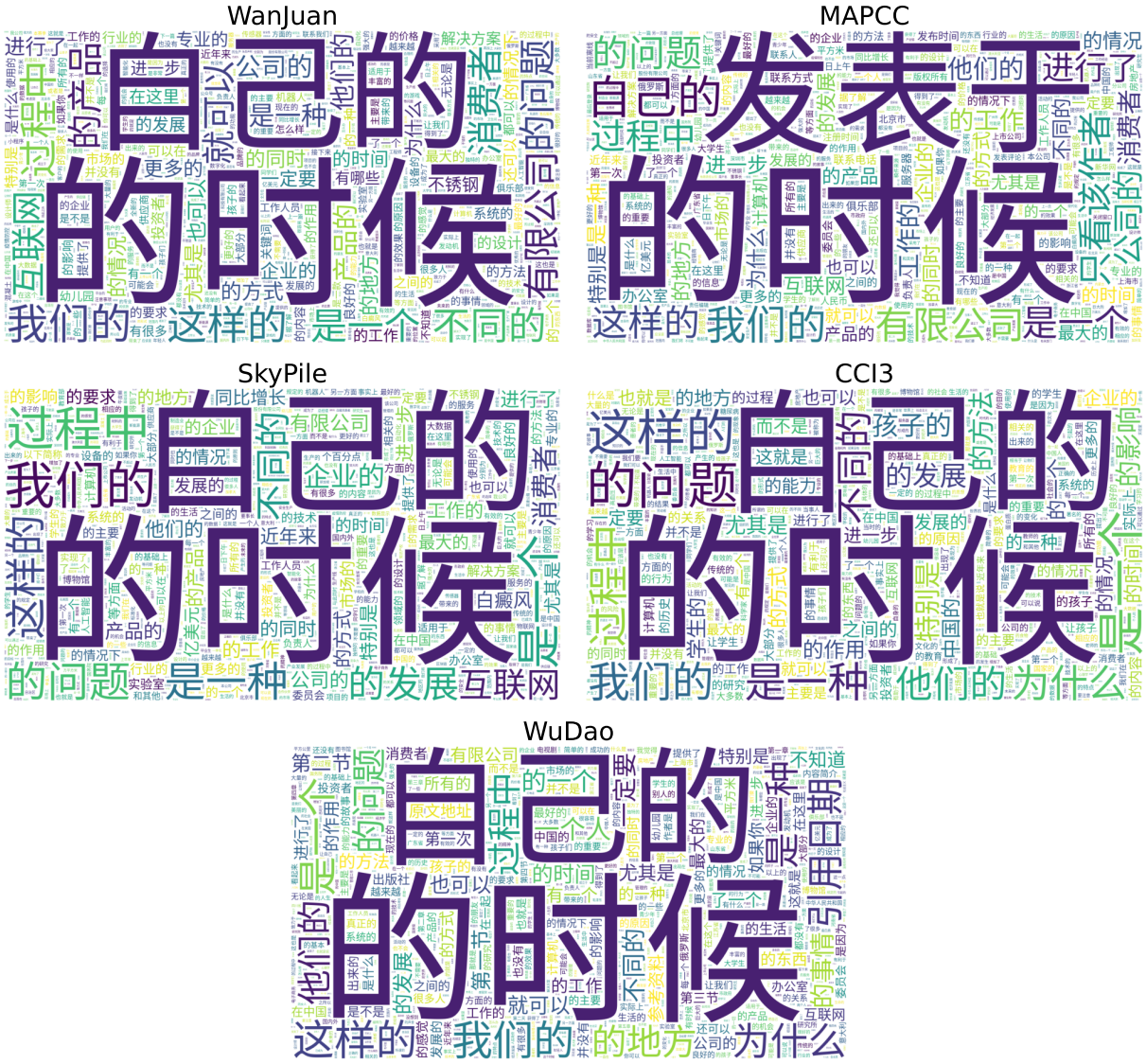


Figure 16: Token clouds for curated Chinese corpus slices.

Figure 17: Token clouds for Chinese Common Crawl snapshots from 2021 to 2026.

J Additional Token Tree and Search Evidence Examples

Following Section 6, this section provides additional token-tree examples and their corresponding search evidence. These examples expand the main-text discussion in two directions. First, they show ***how a tree can trace an entire polluted family back to a compact root***, making related surface variants reviewable as one group rather than as disconnected strings. Second, they show ***how pollution can arise compositionally***: a longer token can become polluted even when the shorter subtokens from which it is formed are normal in isolation. The paired search-evidence figures further show why category labels are assigned with contextual web evidence rather than from token strings alone.

Figure 18 illustrates the compositional case with "北京赛车". The shorter subtokens "北京 (Beijing)" and "赛车 (Car racing)" are normal when interpreted separately, but their combination refers to a PK10-style lottery betting term. The descendants then capture common gambling-search variants such as platform, strategy, purchase, and WeChat-group phrases. This example shows that a flat keyword list would either miss the composed polluted expression or over-block its normal components; the tree instead localizes the polluted meaning to the composed token family.

Figure 19 shows a second compositional case. Here "菲律宾 (Philippines)" and "申博 (Apply for a PhD)" can be normal tokens separately, but their combination is associated with an online gambling brand. Descendants such as agent, official-site, login, and international-casino variants reveal how one composed token expands into a broader promotional family. This case is important because it shows that pollution can emerge from ordinary-looking named entities and abbreviations, rather than from an obviously harmful root.

Figure 20 and Figure 21 illustrate the polluted-root pattern. In Figure 20, "大道香蕉" functions as a recurring adult-site keyword, and the descendants add modifiers such as duration, language, online viewing, and video type. In Figure 21, "浪小辉" refers to an adult-content entity, and the descendants add identity, regional, and content-type variants. In both cases, the root itself is already polluted, so the tree acts as a compact summary of many longer surface forms. This makes the dataset easier to review and helps future cleaning systems generalize beyond exact-string matching.

The corresponding search-evidence figures provide the contextual basis for these labels. Figure 22 and Figure 23 show that the gambling labels are supported by search results and webpages connected to lottery betting, gambling platforms, and betting-site promotion. Figure 24 and Figure 25 show that the adult-content labels are supported by adult-site results and explicit-content contexts. These examples illustrate a recurring difficulty in Chinese web-pollution auditing: many polluted tokens are short, euphemistic, or overloaded with benign meanings. Web context is therefore necessary for auditable category assignment, while the tree structure makes the resulting decisions traceable across related token variants.

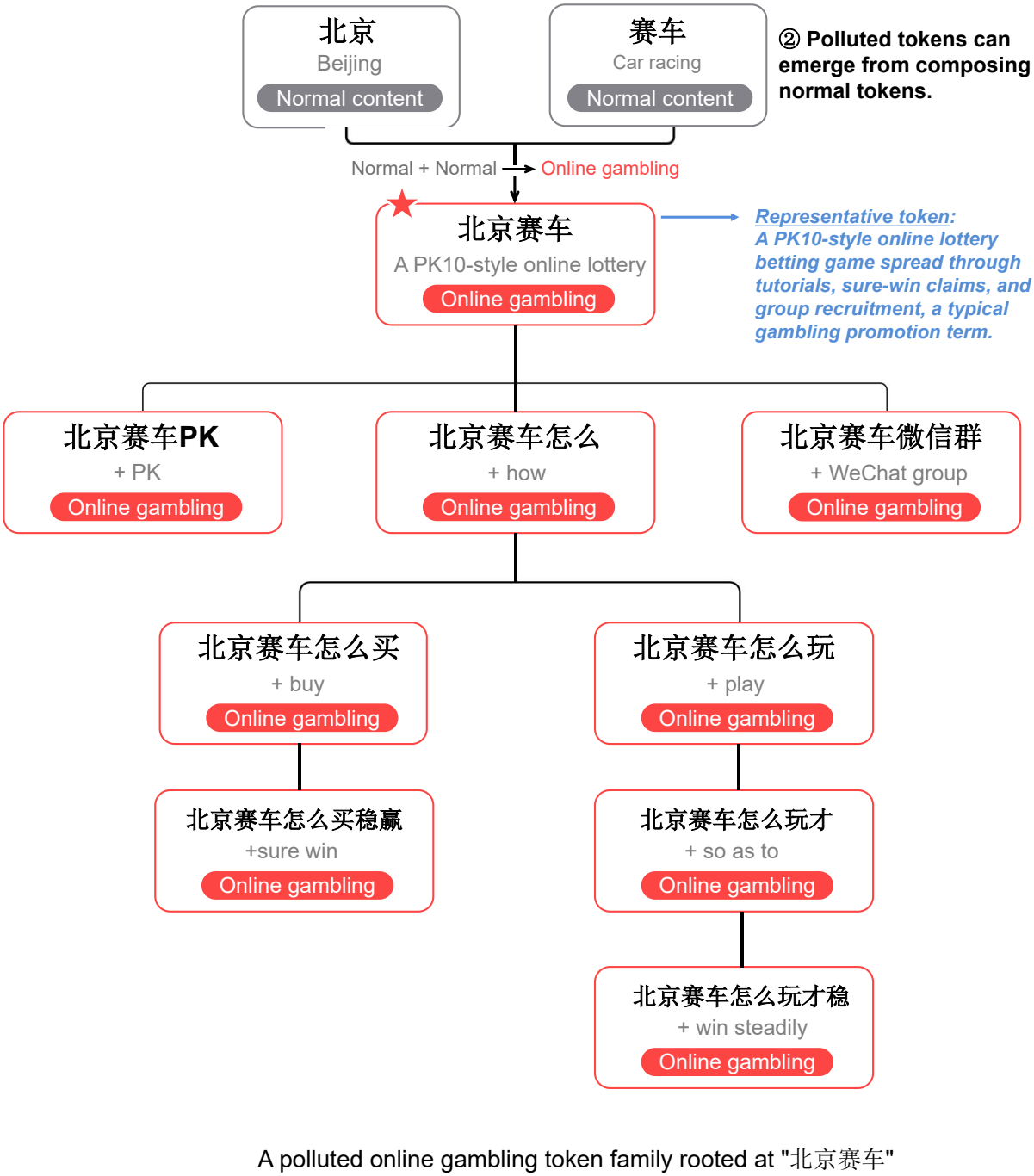
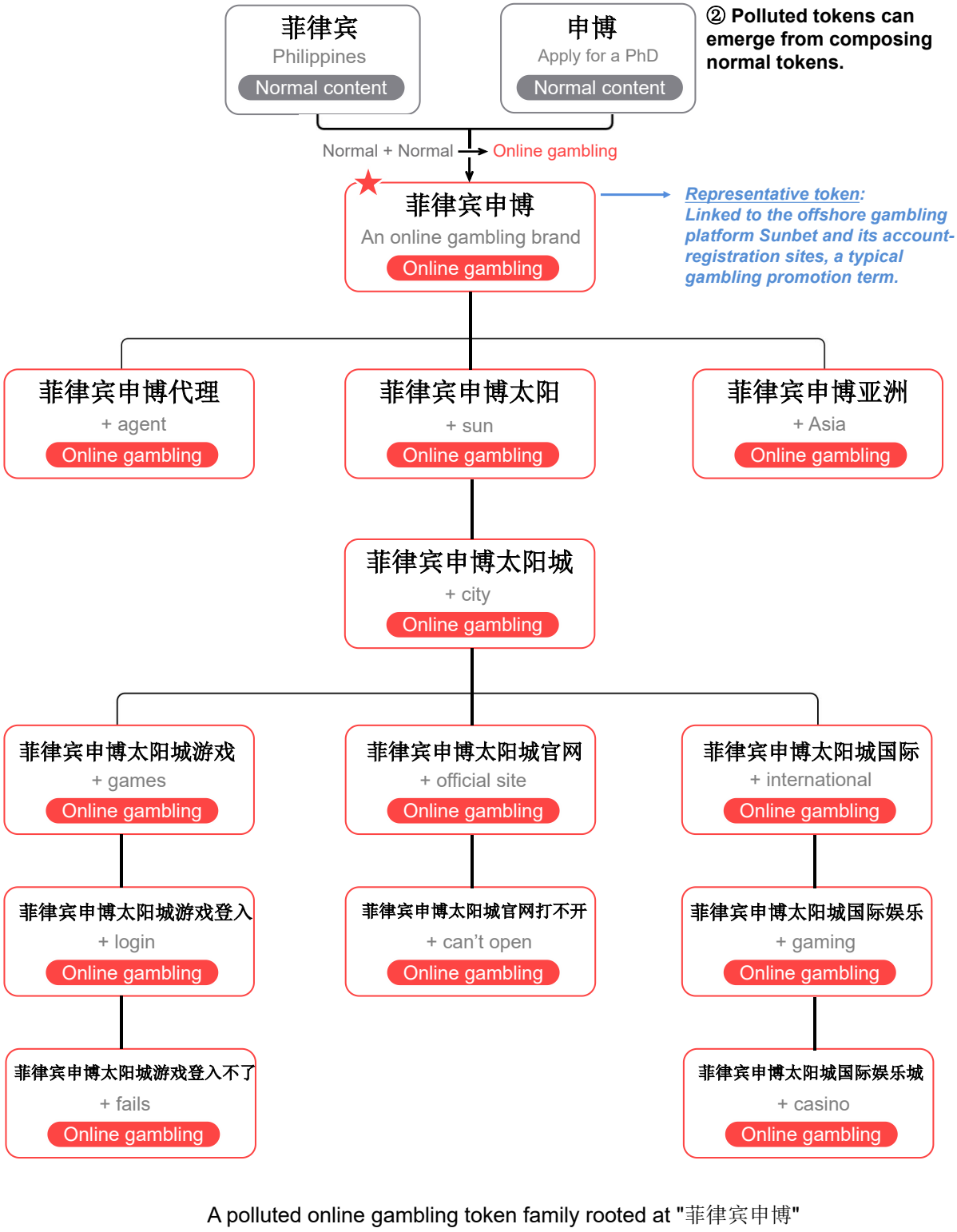


Figure 18: Token tree rooted at “北京赛车”. The tree shows how two individually normal subtokens, “北京” and “赛车”, compose into an Online Gambling token family.



A polluted online gambling token family rooted at "菲律宾申博"

Figure 19: Token tree rooted at "菲律宾申博". The tree shows how normal subtokens can compose into a family of Online Gambling tokens.

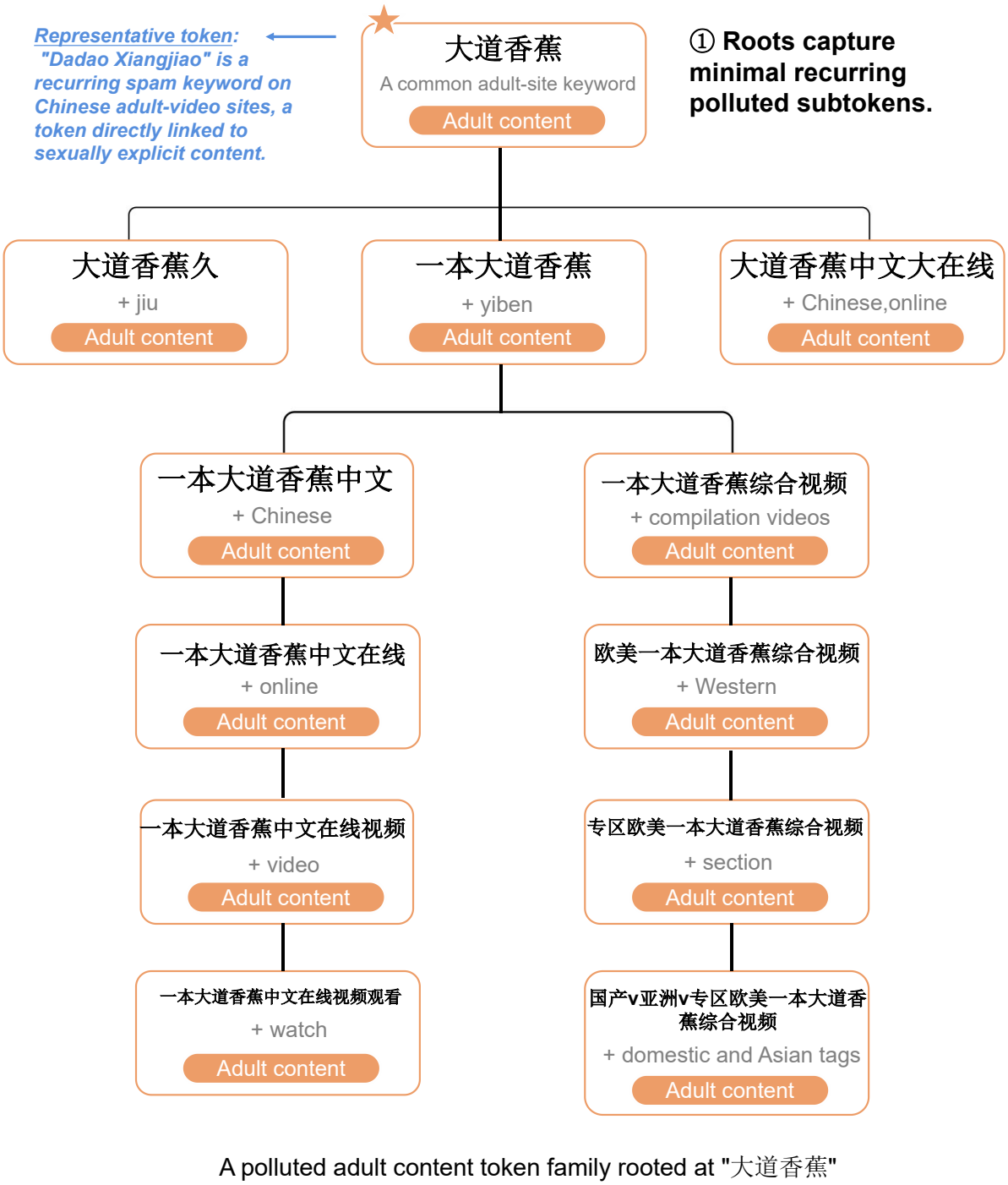


Figure 20: Token tree rooted at “大道香蕉”. The root captures a recurring Adult Content subtoken, and descendants show longer adult-site variants.

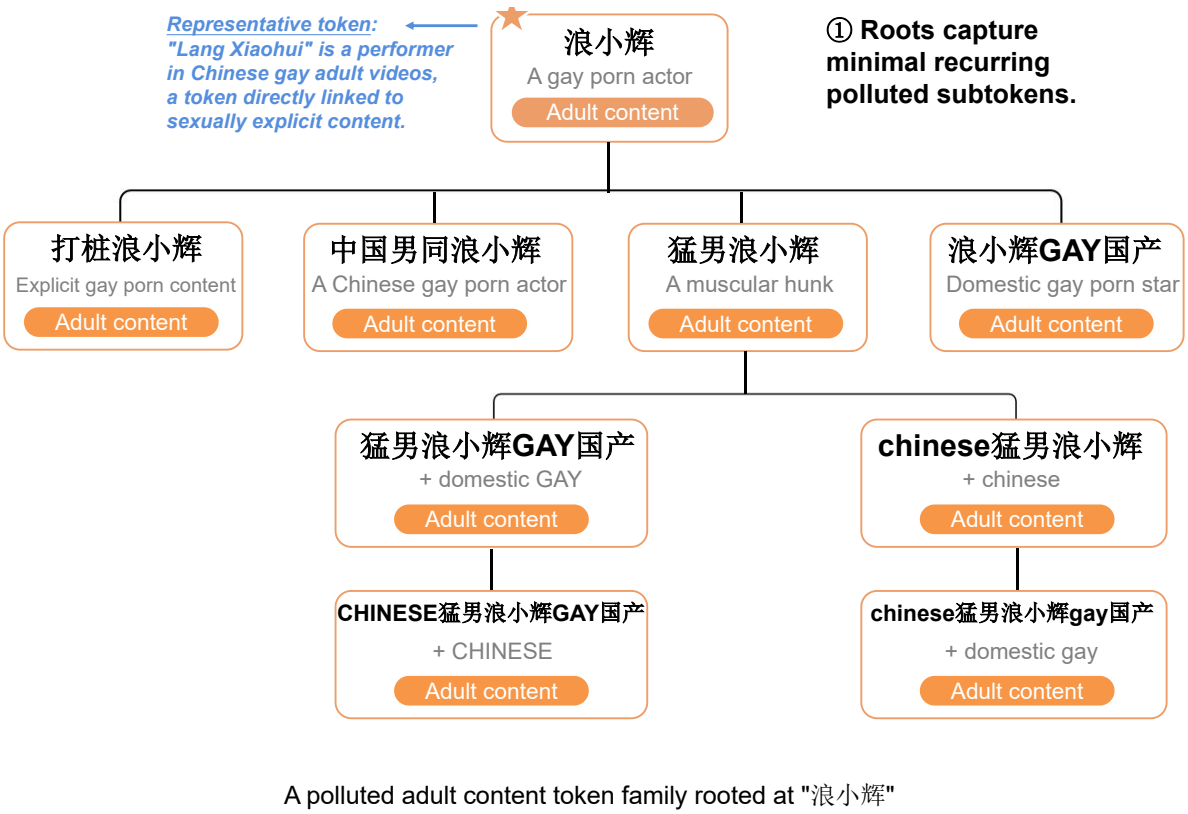


Figure 21: Token tree rooted at “浪小辉”. The root captures a recurring Adult Content entity, and descendants show longer surface variants.

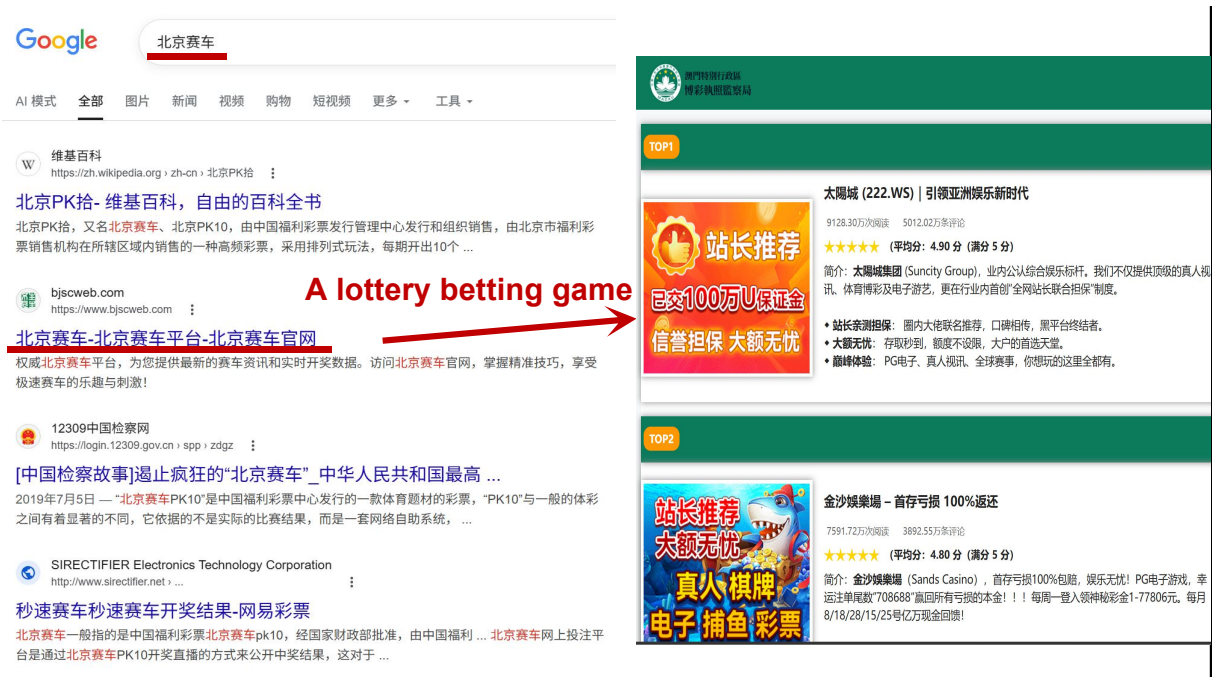


Figure 22: Search evidence for “北京赛车”, showing web context to identify the token family as Online Gambling.

Figure 23: Search evidence for “菲律宾申博”, showing web context to identify the token family as Gambling.

Figure 24: Search evidence for “大道香蕉”, showing web context to identify the token family as Adult Content.

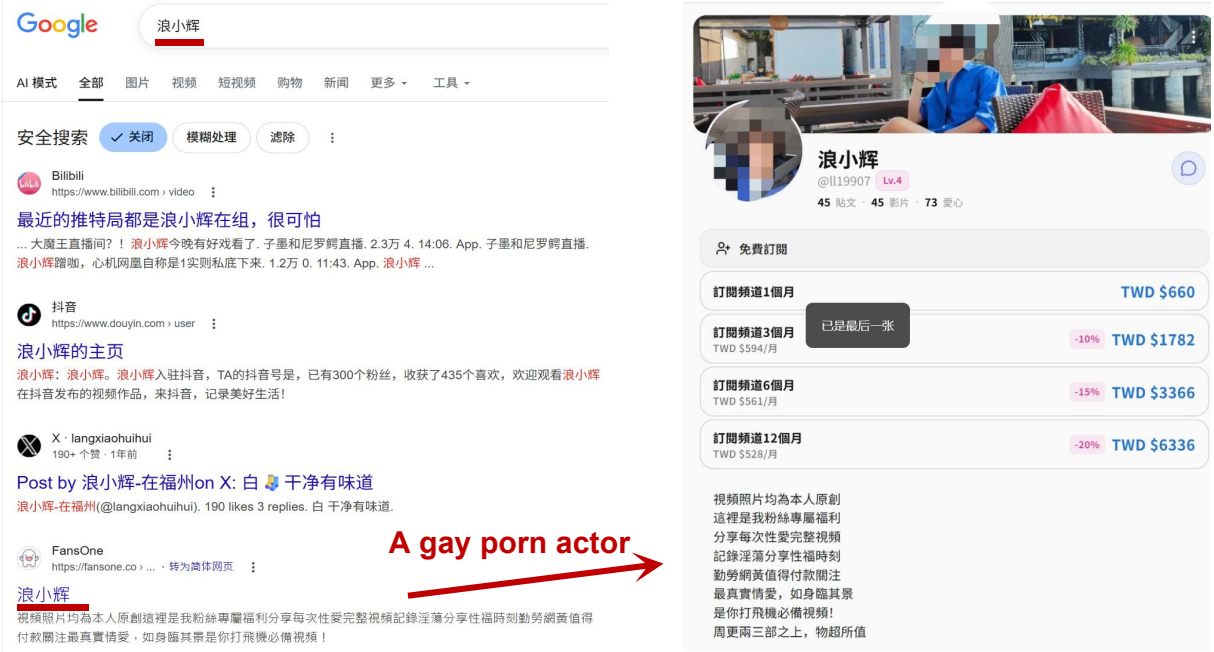


Figure 25: Search evidence for “浪小辉”, showing web context used to identify the token family as Adult Content.