Label Leakage in ASR-Derived Text Channels: A Hygiene Audit and Pre-Registration Protocol for Self-Typed Corpora

Michael Abdo
Independent Researcher · Correspondence: Michael@michaelabdo.com

Methods and protocol article.

Abstract

Researchers increasingly build supervised prediction sets from creator media: a person speaks on video, an automatic-speech-recognition (ASR) system transcribes them, and a human annotator who watched the same video assigns a label, which then serves as the target for the transcript-as-input. This common pipeline carries a silent, structural failure mode shared provenance. When the text channel and the label both descend from one recording, naive text to label prediction scores well for reasons unrelated to the construct: the model re-reads the cues the annotator read, or reads the construct's own jargon spoken aloud. The number stays tautological, not evidential. We quantify this contamination on a real corpus of 87,943 transcript clips spanning 370 human-labeled subjects, and we release a reusable hygiene-plus-pre-registration protocol. The audit decomposes four leakage and quality channels. Start with the dominant one: 78.3% of clips are not the subjects' speech but two annotator/host narrators talking about them, dropped wholesale by narrator person_id. Of the 13,060 remaining real-subject clips, 10.47% leak typology jargon, 2.72% carry host third-person commentary mis-attributed to the celebrity, 18.39% are within-person near-duplicates, and 14.89% are too-short backchannel; only 61.78% survive all filters. Requiring at least 20 clean clips per subject yields a leakage-screened allowlist of 189 subjects and 5,916 clip IDs, down from a raw ceiling of 243. 37 subjects carried at least 20% jargon and would have silently poisoned a raw benchmark. We frame this as a gate, not a result. It removes the crudest leakage but cannot remove the shared-provenance circularity itself, so a strong text to label score stays ambiguous until paired with a cross-modal or independent-text guard.

---

1. Introduction

1.1 The shared-provenance circularity problem

A growing class of datasets predicts attributes of a person, such as personality, affect, political stance, health risk, or deception, from the transcript of that person speaking. The appeal is obvious: ASR is cheap, video is abundant, and a human annotator can watch the clip and assign a ground-truth label in seconds. The pipeline looks like supervised learning. Text in, label out, held-out accuracy reported.

The hidden problem is that the input and the target are not independent. They descend from the same recording. Concretely, three distinct shortcuts can produce a high score with zero construct validity:

1. **Vocabulary leakage.** If the construct has its own jargon and the subject (or anyone in the clip) speaks that jargon, the transcript literally contains the label. A model that reads "I'm a double-activated blast" and predicts "blast" has learned a string match, not a behavior.

2. **Annotator-cue leakage.** Even with no jargon, the annotator assigned the label *from this recording*. Any surface cue the annotator used - topic, register, phrasing - is present in the transcript for a model to re-read. The model is not predicting the construct; it is reverse-engineering the annotator.
3. **Attribution leakage.** Speaker diarization is imperfect. When a host or narrator discusses the subject in the third person and the diarizer mis-attributes that speech to the subject, the "subject's transcript" is contaminated with someone else *describing* the subject - frequently including the label itself.

None of this is exotic. These three shortcuts represent the default condition for any transcript-derived prediction set where labels come from the same source media. And the risk lies precisely in their silence: the pipeline executes successfully with high accuracy, while the standard train-test split offers no clue that the provenance link was never broken. The held-out clips belonging to a subject share provenance with that subject's label, just as the training clips do.

1.2 Why this is the failure mode that sinks fields, not just papers

Two cautionary cases motivate treating this as a first-class methodological hazard rather than a nuisance.

Take the distressed personality, or Type-D construct. It gained traction after accumulating hundreds of positive associations with cardiac outcomes (Denollet, 2005 [1]). Methodologists later argued that much of this signal stemmed from how the construct was operationalized and measured rather than reflecting an independent trait. Because the measurement and the outcome were not cleanly separable, published commentary and reanalysis contended that the reported associations were spurious. These scholars attributed the findings to flawed modeling and shared measurement (de Voogd, Sanderman, & Coyne, 2012 [2]; Coyne & de Voogd, 2012 [3]). The lesson lands hard: a construct can amass a long positive track record while the positivity is an artifact of shared measurement provenance.

Physiognomy ML. The "predicting traits from faces" literature - e.g., sexual orientation (Wang & Kosinski, 2018 [4]) or criminality (Wu & Zhang, 2016 [5]) from photographs - drew sustained scientific and ethical criticism for reporting "accuracy" that, on inspection, reflected demographic and presentation confounds - grooming, pose, image source - rather than any inner trait (Agüera y Arcas, Mitchell, & Todorov, 2017 [6]). The reputational landmine there is identical in shape. A model scored well by reading a *correlate of the labeling process*, not the construct.

Both fields illustrate the same trap our corpus sits in: the apparatus reads the confound, and the confound co-occurs with the label by construction. A transcript-derived typology benchmark that does not audit provenance is one regex away from being the Type-D / physiognomy story told in text.

1.3 Our corpus and contribution

Our setting is a personality-typing project called the Objective Personality System, or OPS. Human annotators watch interview and monologue videos of public figures and assign a fine-grained type. The videos are diarized into per-speaker clips and transcribed with ASR. A downstream experiment, E1, proposes to feed about 20 of a subject's transcript clips to a large language model, then asks the model to predict each binary type dimension, or coin, with no identity given, using majority-voting across clips. That experiment is the program's load-bearing internal question. Can a stock model

type people from text well enough that scaling becomes a budget problem, not a research problem? It is exactly the kind of text-to-label benchmark that shared provenance can quietly invalidate.

This paper serves as the essential gate that must precede such a benchmark. We contribute:

1. A **quantified contamination audit** of a real 87,943-clip corpus, decomposed into four leakage/quality channels with per-subject detail across 370 subjects.
2. A **reusable, fully specified hygiene protocol** - host-person removal, a tiered jargon regex, third-person host-commentary detection, exact-plus-near deduplication, and a minimum-length monologue filter - each with thresholds and rationale, reproducible from queries and CPU alone (no GPU, no paid model).
3. The **recovered clean subset** (189 subjects, 5,916 clip IDs) and the per-subject contamination table, released as machine-readable allowlists.
4. A **pre-registration template** (instantiated in §2.8) that turns the cleaning step into a declared, frozen gate, and an explicit statement of the **residual circularity the cleaning cannot remove**, with the cross-modal / independent-text guards required before any external claim.

A benchmark number does not yet exist. This paper does not produce one either. What it generates instead is the hygiene and honesty that any such figure requires to be interpretable.

---

2. Methods / Protocol

All steps run from SQL queries and CPU-only Python with no GPU no paid LLM and no network model calls so the audit is cheap to reproduce and re-run. The corpus is a PostgreSQL clips table keyed by clip_id with a person_id foreign key to a gold-typed persons table and a normalized ASR text field. The protocol is a cascade of five filters a clip is CLEAN if and only if it survives all of them. We define each filter its threshold and the design reason.

2.0 Notation and the CLEAN definition

For every scanned clip featuring a real subject, we calculate four boolean flags: jargon, host, dup, and short, and define

```
CLEAN  =  NOT (jargon OR host OR dup OR short)
```

Flags can overlap. A clip might be both a duplicate and too short, for instance. So the channel percentages in Section 3 sum to more than `1−{=latex} CLEAN`, and we report them separately rather than as a disjoint partition. The choice is deliberate: separate channels show a downstream user which hygiene step is doing the work for their corpus. A single "% dirty" number would hide that.

2.1 Filter 0 - Host / narrator removal (the dominant contaminant)

Most of the corpus consists of narrator speech rather than subject speech. Two annotator-narrator identities, the individuals who produce the typing content and narrate over the subject footage, are correctly attributed to their own person_ids and together account for 68,861 of 87,943 clips (78.3%). This breaks down to one narrator at 38,947 clips (44.29%) and another at 29,914 clips (34.02%). These are not the subjects at all. They are the typers talking. A further 6,016 clips (6.84%) carry a NULL person_id and are unattributable.

We drop all clips whose person_id matches known host or narrator identities. We also exclude clips with NULL attributions from the subject-level analysis. That leaves 13,060 real-subject gold clips, which represent 14.85% of the table, and this group serves as the population for every subsequent filter.

Why wholesale, not per-clip. Because the host identities are *correctly* attributed to their own `person_id`s, the whole block can be dropped cleanly by `person_id`: their clips are narration about subjects, dense with the label vocabulary, and offer no subject behavior to predict from. Per-clip salvage is not worth the contamination risk - a single surviving host clip in a subject's bag can inject the answer. Dropping the host persons wholesale is the highest-leverage hygiene action in the entire protocol, removing more than three quarters of the table. (The genuine diarizer *mis-attribution* problem - host commentary attributed to a celebrity - is the separate, much smaller Filter 2 channel of §2.3.)

2.2 Filter 1 - Typology-jargon leakage (tiered regex)

The construct boasts a rich and recognizable vocabulary. When it appears in a clip, the transcript contains a near-direct read of the label. We detect it with a two-tier, collision-guarded regex.

- **Tier A - unambiguous construct phrases.** Phrases that essentially never occur outside the typing discourse: `double activated` / `single activated`, `savior`, `demon <X>`, the animal lexicon `blast` / `play` / `consume` / `sleep` (in their typological sense), `quadra`, `four sides`, explicit type strings such as `FF-Fi/Se` or `MF-Ni`, and function-pair codes such as `Ni/Fi`.
- **Tier B - construct vocabulary.** Single terms that signal the discourse but are slightly more ambiguous: `observer`, `decider`, `sensory`, `(intro/extra)verted feeling/thinking`, `MBTI`, and the 16-type four-letter codes.
- **Function codes.** The two-letter function tokens `Se Si Ne Ni Te Ti Fe Fi Oe Oi De Di`, matched **case-sensitively** and behind a **common-word collision filter** so that ordinary English does not trigger them: "Santa Fe", "female", "December", "seven", and similar are explicitly excluded. This filter is the difference between a usable detector and one that flags half the corpus.

Rule. Flag jargon = TRUE if any tier matches. Across the real-subject clips, 1,368 (10.47%) leak jargon. The distribution is heavy-tailed: 37 subjects carry $\geq$20% jargon, and the worst offenders exceed 60% (one subject's clips are 63.2% jargon). These subjects would silently dominate any naive benchmark - the model would "predict" their type by reading it.

2.3 Filter 2 - Host third-person commentary mis-attributed to the subject

Unlike Filter 0, some clips that are correctly attributed, or attributed to a celebrity by the diarizer, actually contain narration about the subject in the third person rather than the subject's own speech. These instances carry the labeling logic explicitly.

Apply the rule of flagging host as true for third-person typing-narration patterns. These include phrases like "you're going to see…", "original bias", "going to be a single decider", "are you a decider or observer", "very good, ", and similar coaching or narration constructions. This process flags 355 clips, which accounts for 2.72% of the real-subject population. Small in aggregate. But it concentrates on exactly the subjects whose footage was most heavily narrated, and in these cases the flag co-occurs with high jargon rates.

2.4 Filter 3 - Within-person deduplication (exact + near)

When ASR processes re-uploaded, re-cut, or overlapping source videos, it generates large blocks of repeated transcript within a single subject. These duplicates inflate the apparent sample size and allow a single phrasing to dominate a majority vote.

Rule. Within each person_id, we apply two checks. First, for exact duplicates, we flag any repeats of the normalized transcript text and keep only the first occurrence. Second, for near duplicates, we flag repeats of the first 60 characters of the normalized text, but only for clips that are at least 40 characters long, again keeping just the first occurrence.

This identifies 2,402 clips, which accounts for 18.39 percent, as within-person duplicates. We deliberately limit deduplication to within each person because two different subjects uttering the same generic sentence constitute legitimately distinct training instances, whereas a single subject saying it twice does not.

2.5 Filter 4 - Minimum-length monologue filter

Backchannel ("yeah", "right", "for sure") and other sub-sentence fragments carry no behavioral signal and pollute a per-clip vote with noise.

Rule. Set short to TRUE when the normalized transcript length is less than or equal to 50 characters. This action flags 1,945 clips, which accounts for 14.89% of the total. The threshold is intentionally conservative. It targets non-monologue backchannel rather than merely brief-but-substantive utterances.

2.6 Assembling the allowlist

Once we finish computing CLEAN, we aggregate the data to the subject level and then apply a usability gate:

Rule. Keep a subject in the runnable allowlist only if they have ≥ 20 clean clips. This yields 189 subjects and 5,916 clean clip IDs.

The floor of 20 clips comes from the downstream majority-vote design, which uses roughly 20 clips per subject, rather than from the audit itself. A different downstream task would establish a different floor against the same per-subject clean counts, which we release in full.

2.7 Outputs and enforcement

The protocol generates two machine-readable artifacts:

- `e8_clean_subset.csv` - one row per allowlisted subject: `person_id, name, type_full, n_clean_clips` (189 rows).
- `e8_clean_clips.csv` - the exact clean-clip allowlist: `person_id, clip_id` (5,916 rows).

Enforcement is built into the downstream harness. With the allowlist present, it uses only allowlisted clips. With the allowlist absent it refuses to run by default, requiring an explicit `--allow-dirty` override that prints a loud "NOT PUBLISHABLE" warning. Making the clean state the default and the dirty state a deliberate, flagged act is how the gate stays a gate.

2.8 The pre-registration template

The cleaning step only becomes a declared, frozen gate if the choices above are committed before any downstream score is seen. We therefore instantiate the gate as a pre-registration template, which is a short set of fields that a user fills in and timestamps before running the benchmark. Every field below is a commitment to a choice already specified in this protocol, and none asserts a result. The template is the in-paper deliverable named in the title and §1.3, and is released as such (Section 7).

```
PRE-REGISTRATION TEMPLATE - text`\ensuremath{\rightarrow}`{=latex}label hygiene gate

1. Frozen allowlist.      Identifier (path or DOI) and content hash/version of
                          the clean-clip allowlist (here: e8_clean_clips.csv,
                          5,916 clip IDs across 189 subjects) and the subject
                          allowlist (e8_clean_subset.csv). The hash freezes the
                          exact clips before any score is computed.
2. Declared filters.      The filter thresholds applied, frozen as declared
                          (here, per §2.1-§2.6): host/narrator person_ids
                          dropped wholesale; jargon = Tier-A OR Tier-B OR
                          case-sensitive collision-guarded function code; host
                          3rd-person commentary patterns; within-person exact +
                          60-char-prefix near-duplicate (len `\ensuremath{\geq}`{=latex} 40); s
                          chars; subject usability floor = `\ensuremath{\geq}`{=latex} 20 clear
3. Primary task + stop.   The single primary downstream task, its metric, and a
                          stopping rule fixed in advance (here, per §2.6: ~20
                          clips/subject, majority-vote per binary coin), so the
                          analysis cannot be re-cut until it passes.
4. Required guard.        The cross-modal or independent-text guard (§4.2) that
                          will accompany the cleaned-text leg before any external
                          claim. A cleaned-text-only score is declared, in
                          advance, to answer feasibility only.
5. Skeptical commitment.  An explicit pre-commitment to the skeptical
                          interpretation (§4.1): the result is reported as a
                          confound-aware feasibility gate, not a construct-
                          validity claim, regardless of how the score turns out.
```

Filling and timestamping these five fields before the benchmark is what converts "we cleaned the data" into "we declared, in advance, the gate the data must pass." The template carries no number that this audit has not already produced; it freezes the decisions, not the outcome.

---

3. Results

3.1 Whole-corpus contamination

Metric	Count	% of table
Total clips	87,943	100%
Host clips dropped (two narrator identities)	68,861	78.3%

Metric	Count	% of table
- narrator A	38,947	44.29%
- narrator B	29,914	34.02%
NULL `person_id` (unattributable)	6,016	6.84%
Real-subject gold clips scanned	**13,060**	**14.85%**

The scanned population of 13,060 is derived directly as the exact sum of `n_clips` across the 370 per-subject rows rather than by subtraction. Approximately 6 clips fall outside the gold-person join because they lack a usable `person_id` link and are thus excluded. So a naive subtraction of the host and NULL rows (87,943 − 68,861 − 6,016 = 13,066) over-counts by 6. All real-subject percentages in §3.2-§3.4 use 13,060 as the denominator.

The most critical finding appears in the first row of the breakdown: more than three quarters of an 87,943-clip "subject" corpus is not the subjects at all. Any benchmark built on raw clips would have spent most of its compute reading two narrators describe other people.

3.2 Contamination within the real-subject clips

Flag	Clips	% of scanned
OPS-jargon (leaks typology)	1,368	10.47%
Host 3rd-person commentary (mis-attributed)	355	2.72%
Within-person duplicate (exact + near)	2,402	18.39%
Too short ($\leq$ 50 chars, non-monologue)	1,945	14.89%
CLEAN (passes all filters)	**8,068**	**61.78%**

The flags overlap, and the definition of CLEAN is NOT(jargon OR host OR dup OR short).

Roughly 38% of even the real-subject clips fail at least one hygiene filter. Duplication is the largest single channel. Jargon is the most dangerous channel, since it leaks the answer directly, and host-commentary is the smallest but most concentrated.

3.3 The recovered clean subset

Stage	Subjects	Clips
Raw ceiling ($\geq$20 clips before cleaning)	243	-
Clean clips, all subjects	368 (have $\geq$1 clean clip)	8,068
Runnable allowlist ($\geq$20 clean clips)	**189**	**5,916**

Of the 370 gold persons scanned (those with ≥1 raw clip), 2 - Robin Williams and Idina Menzel - had every attributed clip flagged as jargon, leaving them with 0 clean clips; the remaining 368 retain ≥1 clean clip. That a subject can be cleaned down to zero is the gate working as designed.

Cleaning removes 54 subjects from the runnable pool, reducing the count from 243 to 189. These are subjects who appeared viable based on raw counts but, after host, jargon, duplicate, and short removal, no longer clear the threshold of 20 usable clips. The allowlist contains 5,916 clips. That figure is the subset of the 8,068 total clean clips that belong to the 189 subjects meeting the floor. Clean clips belonging to subjects below the floor are correctly excluded from the runnable set.

3.4 The heavy tail: subjects who would have poisoned the benchmark

37 subjects carry ≥20% jargon. The worst, by jargon rate among subjects with ≥10 clips:

Subject	n_clips	% jargon	n_clean
Simone Biles	19	63.2	6
Tamera Mowry-Housley	58	62.1	10
Ricky Gervais	34	58.8	12
Jared Padalecki	27	55.6	11
Amy Winehouse	20	50.0	7
Richard Feynman	28	42.9	14
David Lynch	38	42.1	18
Meghan Trainor	69	40.6	30
Tia Mowry	58	39.7	31
Minnie Driver	144	39.6	47

These are not edge cases to be hand-waved. They are the subjects most exposed to vocabulary leakage where a model could score well by reading the label rather than the behavior. For several the high jargon clips are not even the subject speaking. They are the host narrating the subject type mis attributed by the diarizer. Cleaning takes Tamera Mowry Housley from 58 raw clips with 62.1% jargon to 10 clean clips dropping her below the usability floor entirely. That is the protocol working as designed. It would rather lose a subject than admit a poisoned one.

---

4. Discussion

4.1 This is a gate, not a result

We want to be unambiguous about what this paper is and is not. It does not report that anything can be predicted from text. It reports the contamination structure of a transcript-derived corpus and a procedure for removing the removable part of it. The value is methodological: a downstream benchmark run on the cleaned allowlist is interpretable, whereas the same benchmark run on raw clips is not. Treating cleaning as a declared, frozen gate rather than a discretionary preprocessing choice buried in a repo is the contribution.

This approach is also the safer one. As the Type-D and physiognomy cases demonstrate (Section 1.2; [2,3], [4-6]), reputational harm in this area stems from positive claims that are later revealed to be confound-reads. A hygiene-and-honesty paper secures its credibility regardless of the final benchmark outcome because it pre-commits to the skeptical interpretation.

4.2 The residual circularity that cleaning cannot remove

The primary caveat is that our protocol removes input-layer leakage but not shared-label provenance. Imagine a scenario where every jargon clip, every duplicate, every host clip, and every backchannel has been removed, leaving only a pure subject monologue. The label on that subject was still assigned by a human watching the same video. Any surface regularity the annotator used to assign the label, such as topic, register, or the way the subject frames a story, remains in the cleaned transcript. A strong text to label score on the cleaned set is therefore still open to the circularity attack. The model may be re-reading the annotator's cues, not measuring the construct.

Regex falls short here. No amount of string filtering can sever a provenance link that exists within the semantics. To resolve it, you must break the shared-provenance link itself, using one of:

- **A cross-modal guard.** Predict from one modality (e.g., voice or face), label from another (text). If a coin's signal survives across modalities that do not share an input channel, the result is no longer a single-channel tautology. The cleaned text leg becomes one arm of a convergence design rather than a standalone claim.
- **An independent-text guard.** Predict from text the annotator never saw - writing scraped from an independent source - against the same labels. This controls *input-layer* leakage cleanly. It does **not** by itself remove *shared-label* leakage (the label still came from the original video), so even here the residual must be framed, not assumed away.

The honest hierarchy is therefore: the cleaned benchmark answers an internal feasibility question (is the cold-start soft can a model do this at all on clean data) and de-risks downstream engineering spend; it is not an externally publishable construct-validity claim until paired with one of the guards above. We state this so that no reader, and no future use of the allowlist, mistakes a feasibility number for a validity result.

4.3 A generalizable recipe

The failure mode is not specific to personality typing. Any team building a prediction set where the text is a transcript of the same media the labels came from inherits the same four channels. The transferable recipe:

1. **Audit attribution first.** Run the host/narrator and NULL-attribution check before anything else. In our corpus this single step removed 78.3% of the data. Mis-attribution is the cheapest huge win and the easiest to overlook.
2. **Build a construct-vocabulary detector, tiered and collision-guarded.** If your construct has jargon, your subjects (or their interviewers) will say it. Tier unambiguous phrases separately from ambiguous tokens, and *always* add a common-word collision filter - an unguarded code detector flags ordinary language and destroys the corpus.
3. **Detect third-person commentary, not just vocabulary.** Narration *about* the subject leaks the label even when no jargon token matches.
4. **Deduplicate within entity, exact and near.** Re-cuts and re-uploads create blocks of repeats that inflate n and let one phrasing win a vote.
5. **Filter to monologue length.** Backchannel is noise; set a conservative character floor.
6. **Report channels separately, then gate on a usability floor.** Disjoint "% dirty" hides which step matters; per-channel reporting tells the next user where their risk lives.
7. **Freeze the allowlist and make clean the default.** Emit explicit clip-ID allowlists; make the downstream tool refuse dirty data unless overridden with a loud warning.

8. **State the residual.** Cleaning is necessary, not sufficient. Name the shared-provenance circularity and the guard you will pair with it *before* you report any score.

---

5. Limitations

Regex under-cleans, and this represents a floor rather than a ceiling. Each filter serves as a lower bound on contamination. Because the jargon detector matches surface forms, paraphrases of the construct, implicit type-talk, and jargon mangled by ASR pass through. The host-commentary patterns are hand-curated and will miss novel narration phrasings. Near-duplicate detection on a 60-character prefix misses re-orderings and mid-clip repeats. So the reported 10.47% jargon and 18.39% duplicate rates are minimums, meaning the true contamination is higher. The protocol therefore guarantees that surviving clips are cleaner, not that they are clean. A residual-leakage estimate, such as a manual audit of a random sample of clean clips or an embedding-based paraphrase detector, would tighten the bound and is the natural next methodological step.

Diarization happens upstream and is far from perfect. Because host-person removal relies on accurate speaker identities, errors in the diarizer can have cascading effects. If the system splits or merges speakers incorrectly, some host speech might survive under a subject's ID, though Filter 2 catches some of it. At the same time, some subject speech could be lost entirely. We treat the diarization as a given, meaning its error rate sets the limit on how clean the results can be.

The usability floor depends on the corpus and the task. We set the gate at 20 or more clean clips based on a specific downstream majority-vote design. A different task calls for a different floor. By releasing per-subject clean counts, we allow the floor to be re-chosen without re-running the audit.

Single-corpus quantification. These percentages apply specifically to this OPS corpus and its production pipeline, which relies on heavily narrated source videos and two dominant hosts. The channels generalize. The magnitudes do not. The 78.3% host fraction is an artifact of this corpus's narrated-content structure and should not be read as typical.

The cleaning does not validate the construct. We make no assumptions about the validity of the underlying typology. Instead, we clean the text channel for whatever the labels mean. Construct validity is a separate question. That issue is addressed by separate work in the program, which is pre-registered and null-tolerant.

---

6. Ethics and Data Provenance

We tackle this directly. The paper highlights the ethical track record of physiognomy literature (Section 1.2, [4]-[6]) as a cautionary tale. Taking that record seriously demands that we apply the same level of scrutiny to the artifacts we release.

The released labels are third-party annotator opinions, not validated traits. The type strings attached to named public figures are subjective annotations produced by the OPS typing community about public figures. These strings are propagated through the per-subject audit table in `e8_leakage_audit.md` (e.g., `Simone Biles, FF-Se/Fi-CP/S(B)`) and, for allowlisted subjects, in `e8_clean_subset.csv`. They represent *opinions* about each figure's personality type. They are not measured, consented to, or scientifically validated facts. This manuscript makes no claim that these labels are correct. Its entire argument is that even a strong text→label score would not

establish their validity (Section 4.2). The labels should be read and cited as annotator opinions about how a public figure presents. Nothing more.

These subjects are public figures, and no private or identifying media is redistributed. Each named individual spoke in publicly available video. The released artifacts contain only derived metadata, such as clip identifiers, per-channel contamination flags, and counts, along with the annotator type label and the subject's already-public name. We do not redistribute any transcript text, audio, video, image, or other potentially identifying or private media in any released artifact.

Source-corpus provenance and usage constraints. The OPS typing project assembled the underlying corpus from publicly posted interview and monologue videos via third-party scraping of celebrity media. We did not collect this material under an explicit research-consent protocol, and we treat its provenance as a known limitation rather than a clean chain of custody. For this reason the corpus itself is **not** redistributed (Section 7 below). Only derived, non-reconstructive metadata is released. Any holder reproducing the audit must supply their own copy of the source media and observe the usage constraints, including platform terms and copyright, attaching to it. We make no representation that bulk re-scraping of the source platforms is permissible. We encourage downstream users to obtain media through channels consistent with the original platforms' terms.

Why does this matter for this paper specifically? A benchmark that attaches fine-grained personality labels to named individuals carries the same reputational hazard as the physiognomy work we cite. It can be read as asserting facts about people that it has not established. We therefore frame the labels as opinions. We release no identifying media. We name the scrape's provenance limits. We gate any predictive claim behind the cross-modal / independent-text guards of Section 4.2. The honest contribution is a hygiene-and-provenance discipline, not a validated trait attribution.

---

7. Data and Code Availability

You can reproduce the protocol using SQL queries and CPU-only Python, without needing a GPU, a paid model, or any network model call. The audit artifacts released with this paper:

- `e8_leakage_audit.md` - the full audit: whole-corpus contamination table, within-real-subject breakdown, detection-method specification, and the complete per-subject contamination table (370 subjects: `person_id, name, type_full, n_clips, pct_jargon, pct_dup, pct_short, pct_host, n_clean`).
- `e8_clean_subset.csv` - the allowlisted subjects (189 rows: `person_id, name, type_full, n_clean_clips`).
- `e8_clean_clips.csv` - the frozen clean-clip allowlist (5,916 rows: `person_id, clip_id`).
- The **pre-registration template** (§2.8) - the five-field gate-declaration instrument, released in-paper and reproducible verbatim from §2.8; it freezes the allowlist hash, the declared filter thresholds, the primary downstream task and stopping rule, the required cross-modal / independent-text guard, and the skeptical-interpretation commitment.

Access. We release the pre-registration template within the paper (§2.8) and ensure it is reproducible verbatim from this manuscript. You can obtain the three machine-readable artifacts (`e8_leakage_audit.md`, `e8_clean_subset.csv`, `e8_clean_clips.csv`) from the author upon request. These files will also be deposited as a single versioned archive at a persistent identifier once the work is published.

The downstream benchmark harness consumes the allowlist and refuses to run on unscreened clips without an explicit override. This process is documented in the companion `e1_typer` protocol, which includes the predict-direction prompt and the pre-registered falsifier. That harness is released alongside the same archive. The underlying corpus consists of proprietary third-party-scraped celebrity media and is not redistributed. The released artifacts are derived metadata, specifically clip identifiers, contamination flags, and counts. These are sufficient to reproduce the audit logic and to reconstruct the allowlist against a holder's own copy of the corpus. Tier-A/Tier-B regex definitions and the common-word collision filter are specified in full in Section 2.2 and the audit document. This enables reimplementation on any transcript corpus.

Competing Interests

The author developed, and holds a continuing interest in, the proprietary OPS-style typing system that produced the trait-axis labels and assembled the source corpus audited here, which represents a potential commercial stake. This paper makes no validity claim for those labels; its contribution is a hygiene-and-provenance discipline that, if anything, raises the bar against over-reading them. We disclose the interest explicitly. No other competing interests are declared.

Funding

This work received no external funding and was self-funded by the author. No funder had any role in the study design, analysis, interpretation, decision to publish, or preparation of the manuscript.

References

[1] Denollet, J. (2005). DS14: Standard assessment of negative affectivity, social inhibition, and Type D personality. *Psychosomatic Medicine*, 67(1), 89-97. https://doi.org/10.1097/01.psy.0000149256.81953.49

[2] de Voogd, J. N., Sanderman, R., & Coyne, J. C. (2012). A meta-analysis of spurious associations between Type D personality and cardiovascular disease endpoints [Letter]. *Annals of Behavioral Medicine*, 44(1), 136-137. https://doi.org/10.1007/s12160-012-9356-7

[3] Coyne, J. C., & de Voogd, J. N. (2012). Flawed meta-analysis of a flawed literature: Commentary on Versteeg et al. [Commentary]. *European Journal of Preventive Cardiology*, 19(6), 1381-1382. https://doi.org/10.1177/2047487312437716

[4] Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114(2), 246-257. https://doi.org/10.1037/pspa0000098

[5] Wu, X., & Zhang, X. (2016). Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135*. https://arxiv.org/abs/1611.04135

[6] Agüera y Arcas, B., Mitchell, M., & Todorov, A. (2017). Physiognomy's new clothes. *Medium*. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a