One Construct, Many Names: Synonyms Yield Divergent Judgements

Alexander Spangher **Daniel E. Ho** **Sanmi Koyejo** **Diyi Yang**
Stanford University
alexspan@stanford.edu

Abstract

LLM judges are often guided by prompts, constitutions and rubrics that verbalize normative criteria. Criteria can be coded in different dialects, formats and registers based on social group (Biber, 1988). Prior work tests LLM judges on paraphrased prompts and rubrics, but not on the different names communities use for the same criterion. We present the first large-scale analysis of evaluative lexicons across 11 creative domains (e.g., journalism, science, comedy, etc.): 53,413 criteria mined to saturation from online rubrics, professional guidance and community rules, drawn from texts spanning from ∼350 BCE (in translation) to the present.[1] We present three findings: (1) *Register* (i.e. formality of wording) differs by source: the rules communities write for themselves are worded less formally than criteria from institutional sources. (2) LLMs asked to name criteria speak at or above that institutional register, in every model tested; i.e. when one LLM writes a prompt for another, as in agentic pipelines, the criteria more often arrive in the formal register. (3) *Naming choice* (i.e. which of a construct's surface forms a rubric uses) changes what a judge measures: swapping only the name between register variants disrupts a judge's rankings while paraphrasing the *entire surrounding prompt* with the name held fixed does not. The register an agent chooses when it instructs another agent can therefore measurably change the agent's behavior. We release lexicon, source corpora, and silver labels as a dataset for studying coded evaluative knowledge across tasks.

1 Introduction

Outputs of agentic AI are increasingly evaluated and improved by other AI agents, or judges, following rubric (i.e. written criteria for how to identify good and bad work). As agents become more integrated into social and creative industries, which criteria are chosen and how they are expressed, thus, is becoming increasingly important. Increasingly, agents write prompts and rubrics to communicate with each other, as in in prompt optimization and multi-agent systems.

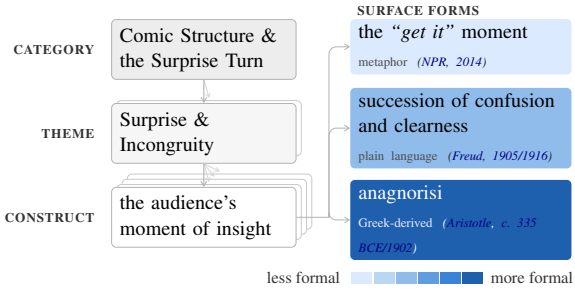


Figure 1: We collect 53k articulations of evaluative criteria (shown above as surface forms) from ∼350 BCE to the present and across 11 tasks (e.g. humor, shown here; journalism, math) and cluster them hierarchically. Shown is an evaluative category, one of its themes, one construct of that theme, and the three attested surface forms that lexicalize it. Shading of surface forms represents the formality of the phrase (lighter is less formal, darker is more formal).

Articulated judgment criteria are ubiquitous across creative domains, and philosophical, psychological and linguistic studies have explored how they are constituted, developed and agreed-upon (Ehrlich and Posner, 1974; Kaplow, 1992; Corson, 1985). However, while large-scale computational analyses of normative criteria have been performed (Fiesler et al., 2018; Chandrasekharan et al., 2018; Weld et al., 2022; Lloyd et al., 2025), they have largely focused on a *governance* perspective (i.e. which behaviors are banned and how rules are enforced). We find a comprehensive and cross-domain analysis of *quality*-focused norms lacking: (1) evaluative criteria have not been analyzed as a linguistic system, with measurement of

[1] Much older writing was never digitized; our saturation claim is over the criteria *reachable* online, not over everything humans have ever written down.

different phrasings for constructs and their social and institutional registers (2) the how these registers impact on LLM judging. Our study differs in three ways: the names we test are naturally discovered, we measure their sociolinguistic properties, and we swap one name for another while holding the construct and the rest of the prompt fixed.

In this work, we ask three questions. **RQ1**: How are quality criteria named across domains, and does the naming of a requirement vary with the class of source that states it? **RQ2**: Which names do LLMs choose when asked to communicate about evaluative constructs, as when one agent writes a rubric for another? **RQ3**: Does using a construct's synonym change what an LLM judge measures, and how well the judge agrees with the community's own judgments?

We conduct a broad-based analysis of *quality*-focused norms across 11 creative domains: Academic writing, creative writing, humor, math, etc. We start by building a comprehensive survey of existing norms: we build a novel web-search technique, inspired by simulated annealing, to discover webpages (e.g. from guidebooks, how-to articles) containing normative criteria and develop a *saturation* measure to mine until p(unseen criteria) drops below a threshold. Next, we hierarchically cluster these norms and analyze their *codability* (Corson, 1985), or how often the same norm gets encoded in different phrasings (Table 1), finding significant codability differences across field. A restatement of a known construct reuses an established name 64% of the time in software code, for instance, but only 17% in humor and 19% in math. We then measure two properties of each phrasing: (1) its *register*, the formality of the wording (Biber, 1988), and (2) its *metaphoricity*, whether the phrasing is figurative (*punch up*) or literal (*citation accuracy*).[2] Criteria from institutional sources (e.g. professional organizations, style guides) are worded more formally than the rules communities write for themselves (e.g. moderators writing rules for forums) in each of the four domains with institutional sources[3].

Finally, we study surface form in the context of LLM judgement. First, we do an *articulation/re-articulation* experiment: we ask an LLM to name a construct from its definition alone with the human name withheld (articulation), then re-score the same items under each phrasing of that construct in turn (re-articulation), holding everything else in the prompt fixed (§4, Fig. 3). We find that LLMs often names criteria in the *institutional* register and that this causally impacts the labels it applies: labels shift by half when a norm's name is swapped between register variants, irrespective of other structural changes to the prompt and inter-rater variance. This shift, we find, comes from the different requirements a construct's names state. An LLM writing a prompt for another LLM will therefore, by default, state criteria in the formal register, and the receiving judge's labels shift measurably as a result. Finally, we derive *silver labels* from comment threads, that we collect and validate, and we find no clear winner across register: a construct's more formal phrasings agree with the community's own judgments about as often as its less formal ones. In sum, our contributions are:

- A dataset of 53,413 evaluative criteria across 11 creative domains, mined to a measured notion of saturation and released with its source corpora, hierarchy, and silver criterion labels;
- An analysis of criteria as a linguistic system: a form→construct→theme→category hierarchy, with validated register, metaphoricity, and specificity instruments (Appendix F);
- Evidence that LLMs default to specific institutional speaking styles and that the *name* of a criterion steers judges, so that the register in which agents instruct each other is a measurable determinant of their behavior.

Taken together, this serves as the broadest known investigation into register and class in construct naming; our findings that LLMs default to institutional naming, and this changes their annotation style, suggests that larger agentic systems, where LLMs design prompts for each other, might recursively take us farther from the communities we might hope to serve.

2 Dataset Creation: Metric Gathering

We define a *criterion* as: a stated requirement for judging the quality of work in a domain, verbatim, as written in a source document: a name (e.g. *economy of language*), a description of the requirement, and any guidance the source gives for

[2] A criterion name is *figurative* when it describes the quality by borrowing from another domain (*a joke needs a strong **button***) and literal when it names the quality directly (*the ending should be decisive*).

[3] Register and metaphoricity are uncorrelated.

Domain	Form	Construct	Theme	Category
Software Code	8,118	3,848	42	9
Creative Writing	2,850	948	57	10
Grant Proposals	1,647	408	45	6
Humor	3,363	1,282	50	9
Legal Arguments	2,262	657	45	10
Math	3,337	820	40	7
Journalism	1,598	719	45	8
Regulatory Comments	2,034	555	39	7
Patents	2,397	888	52	8
Academic Articles	1,859	318	45	7
Press Releases	3,403	606	43	7
Total	32,868	11,049	503	88

Table 1: Criteria at each hierarchy level per domain. Higher levels are small (39–57 themes, 6–10 categories) (Appendix B.)

applying it. Our goal in this section will be to describe how we mined normative criteria discussed by different human communities.

Search until saturation. We collect criteria with an agentic web search (Appendix A.1, Fig. 8): per domain, LLM search agents issue seeded queries, fetch and archive pages, and an LLM parser extracts each stated criterion (name, description, guidance, subfield, audience) with a verbatim evidence quote (validation in Appendix A.3). A central question is when to stop: "the same criterion" can mean the same string, the same underlying construct, or the same broad theme. We therefore model the growing inventory as a Chinese Restaurant Franchise (CRF; Teh et al., 2006). The Chinese Restaurant Process is a model of sampling from an inventory whose size is not known in advance: each new draw is a diner who either joins a table already occupied (i.e. a node we have already found) or starts a new one; the chance of a new table is set by how many exist already. A *franchise* chains several such restaurants together: every table opened in restaurant A is itself a diner in the B (and diners in B either sits at an existing table or a new one). This allows us to model our hierarchy: each name draw enters as a customer of a surface-form restaurant; each *new* surface form is in turn a customer of the construct restaurant; each new construct a customer of the theme. The CRF is a descriptive model of how the inventory grows used to estimate remaining constructs; it is not the procedure that builds the hierarchy. The hierarchy is *built* from the search and clustering pipeline described below. CTF shows that novelty is a *conditional* probability,

$$p_{\text{unseen}} = p_{\text{form}}\, p_{\text{construct|form}}\, p_{\text{theme|construct}} \cdots ,$$

where p_{form} is the chance a draw is a never-seen phrasing and $p_{\text{construct|form}}$ the chance a new phrasing expresses a never-seen construct; each factor is estimated nonparametrically by Good–Turing, f_1/N, where f_1 is the number of types seen exactly once and N is the total number of draws, in other words, the share of draws whose type has been seen exactly once (Good, 1953) (we extend this parametrically by a Pitman–Yor fit, see Appendix B.1). *Saturation* has different meanings at each level: the surface level never saturates (i.e. new phrasings keep arriving at a high rate). The construct level settles onto a low, steady discovery rate. The theme and category levels quickly exhaust; theme-level draws are rare events and $p_{\text{theme|construct}}$ drops quickly early in the collection process at every new domain (Fig. 11). Collection proceeds in waves; early waves sample many different sources, later waves deepen within known ones. We treat a domain as saturated when: (1) construct-level novelty falls below λ=.3 and (2) *stays* there under a *re-heating probe* (i.e. a wave from a previously unsampled stratum). λ bounds $p_{\text{construct|form}}$, the probability that a new surface form states a new construct; 9 of 11 domains are below .3 (software code .36 and patents .31 are not). A low novelty rate can reflect a narrow sampler rather than a covered population; so, we validated our stopping mechanism on a frame where truth is known (Appendix B.3).

The corpus. The search visited 41,132 URLs and yielded 53,413 criterion statements and 32,868 distinct surface forms, drawn from formal guidelines, professional standards, textbooks, research articles, wikis, blogs, and community rule pages; some domains source mostly from institutional documents (e.g. patents, which draw on the patent examination manual) and creative domains from informal ones (Appendix A.1). Source dates span antiquity to the present, which lets us separate perennial norms from recently formalized ones (Fig. 18; Appendix H.5). For each webpage, we extract the criterion itself (name, description, application guidance, and a verbatim evidence quote) and extract or infer metadata: its page type, the community it speaks for, intended audience, and date (examples in Appendix A.2).

Hierarchy creation. We organize the criteria into four levels (Fig. 1). A **surface form** is a distinct phrasing as attested (e.g. *get it moment* and *anagnorisi* are different surface forms). A **construct** groups surface forms that state the *same normative requirement*: the thing a rater would check, in the sense of a psychometric construct (Cronbach and Meehl, 1955) or a qualitative *code* (Braun and Clarke, 2006). A **theme** groups constructs sharing an evaluative concern (e.g. Surprise & Incongruity), the analogue of a Braun–Clarke *theme*; a **category** groups themes into a domain's top-level evaluative dimensions, as in axial coding (Corbin and Strauss, 1990). The number of items at each level shrinks: many phrasings can state one requirement; many requirements can serve one concern (Table 1). Constructing the hierarchy is a large multi-step clustering process rather than a single embedding clustering. Construct identity is a *judgment about normative content*, not surface similarity, and embedding clustering merges both too much (collapsing distinct requirements that share words) and too little (splitting one requirement across paraphrases). Appendix C shows our pairwise method.

Collection Pipeline validation. We validate both steps of our pipeline. First, extraction: against a known-denominator gold, or subreddit pages with enumerable rule lists), the production pass recovers .69 [.62, .75] of the quality rules on a page, every extraction matches a real rule, and two judge families rate .74–.79 of extractions as faithful to the rule's wording; an all-norms pass recovers .995 (Appendix A.3). Second, clustering: we validated our approach against six hierarchies coded by hand[4] by asking how well our hierarchy-construction predicts existing links. Grouping surface forms into constructs recovers human links at AUC .55–.77; constructs into themes at .73–.96; themes into categories at .69–1.00. Themes and categories are where the pipeline agrees with human coders most closely. For the clustering algorithm, the validation design, baselines, and the full per-corpus table, see Appendix D. Finally, a corpus coded at two levels (Alami and Krancher, 2022), our passes settle on 39 clusters against 35 human codes and 11 against 14 human themes; the granularity matches within roughly 20% at both levels (Appendix D).

Behavioral dataset. To measure whether wording changes what a judge *measures*, we assess creative output with real, human-authored community judgments. For six of our domains we collect the community's own commentary on individual pieces of work (e.g. from reviews, critique threads, feedback comments). From these comments, we parse stated *norm applications* and map them onto the construct invoked (e.g., "this joke ran too long"). To do this, we use the same grouping mechanisms described above that we used to create the hierarchy itself. We collect $>$ 1 million norms mentioned in comments and map 64,820 of them onto 665 constructs; we view this as a weakly supervised method for generating silver labels for social constructs for which gold-annotated data may not exist. Because the criterion-to-construct mapping is the same operation validated in Appendix D, there is potential for circularity: i.e. the judge family that grouped surface forms into constructs also decides which construct a comment invokes, so any bias in how it reads the construct would distort the hierarchy and the labels in the same direction, increasing agreement. We controlled for this in two ways: results are computed at the leaf level, where the matcher has no grouping freedom[5]. And, A judge from a different family (GPT-5.6) read 300 seated matches: the comment invokes the assigned construct in 77% of them and at least partly in .92% Appendix E details the per-domain text sources, parsing, and label construction.

3 Measures and Measure Validation

We score and validate the following linguistic instruments, each with an LLM-judge:

- *Register* asks how elevated the name's wording is as a social signal: does it come from casual everyday speech or from technical-institutional language? It is annotated following the lexical-bar tradition (Corson, 1985; Biber, 1988), which combines: (1) formality judged on a 1–7 scale, validated against crowd lexical-formality gold (ρ=.85;

[4]The corpora we use: occupational work activities (National Center for O*NET Development, 2025), discourse relations (Webber et al., 2019), code-review feedback (Turzo et al., 2023), and three qualitative codebooks

[5]Grouping is a decision the judge controls, so we compared each family-level effect against a null built by reshuffling the same constructs into 500 random families with the observed size distribution; in four of five domains, those family-level groupings were dropped.

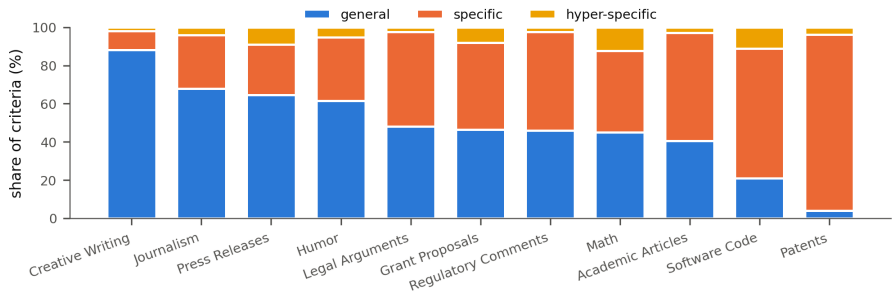


Figure 2: Norm specificity across eleven domains: the share of criteria that are *general* (travel across contexts: "the work should have a good style"), *specific* (document-particular rules: "the study design should be applicable to the research question"), and *hyper-specific* (concrete surface constraints: "units of measurement abbreviations should be in lowercase"). Pool: the census plus the newly collected lay-critic and community-rule criteria.

Pavlick and Nenkova, 2015); (2) etymological stratum (Germanic/Latinate/Greek), validated against the Etymological Wordnet (86–90%; de Melo, 2014); and (3) nominalization is validated against NOMLEX (98%; Macleod et al., 1998).

- *Metaphoricity* asks: is the name's use figurative borrowing or literal description? It is measured via the MIP/MIPVU procedure (Pragglejaz Group, 2007): Our LLM-judge scores 89% on type-level *idiom* (Haagsma et al., 2020). *Idiom* tests detection of figurative expressions; on construct names and descriptions, human annotators agree with the judge at κ=.49 (37 names; Appendix F.1).

- *Specificity* asks how broad a criterion's normative scope is with respect to the situations its verdict applies to: {general, specific, hyper-specific}. This label reflects how concrete the checked feature is, for e.g. *units of measurement abbreviations should be in lowercase* is labeled hyper-specific, although it applies across many documents. We validate specificity *behaviorally*, in part: rules the LLM judges as general are *also* adopted across more communities in a 100K-subreddit corpus (Lloyd et al., 2025) than rules judged specific (median 14.5 against 5 subreddits when 450 rules are relabeled by GPT-5.6; 10 against 5 under Claude), but rules judged hyper-specific (median 8) are not adopted less widely than specific ones[6]. Human annotators and GPT-5.6 and Claude agree with the deployed labels at κ=.49–.57.

- *Polarity* asks whether the phrasing frames the criterion as a quality to achieve or a fault to avoid and is classified in {positive standard, prohibition}. We validate against LexDeMod, Sancheti et al., 2022 at F_1=.82 [7].

Every judging batch contains a sample of items with a fixed correct answer, drawn from each instrument's validation gold or synthetically constructed and is discarded if these fail to be annotated; for more detail, see Appendices F and F.1.

4 Impacts of Register on Prompt Behaviors

We now turn from the lexicon itself to analyzing consequences for LLM judges: we perform three experiments to ask what register LLMs themselves speak (rearticulation), whether the name alone steers judgment (name swapping), and which phrasing best recovers the community's own judgments. Full details are in Appendix G.

Finding #1: LLMs speak the official register. First, we run a *rearticulation* experiment: each model family sees a construct's definition *with its name withheld* and is asked to name it fresh (Figure 3). Our experiment shows: first, we find that LLMs use greater uniformity compared with humans. Across model families, the modal LLM-rearticulated name nearly doubles, from .11 in

[6] Adoption breadth can also reflect copying and boilerplate shared between communities, so we read breadth as *convergent* evidence of generality rather than a measure. Specificity is not correlated with descriptive-detail ($\rho = -.29$), which asks how much concrete, particular description a phrasing contains; *detail* is also assessed via a judge, validated at ρ=+.64 against (Ko et al., 2019), and survives length control unchanged (Appendix F.1).

[7] Sancheti et al., 2022 provides obligation and prohibition clauses in lease agreements.

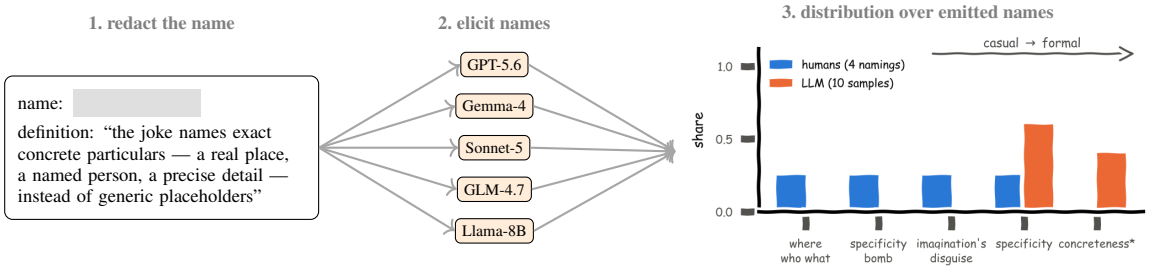


Figure 3: **The rearticulation experiment**. Each construct's definition is shown with the name withheld; each LLM family names the construct from the definition alone, and the emitted names form a distribution over terms. Right: real distributions for the construct shown — its four *human* namings are spread across the register range (*where who what*, *specificity bomb*, etc.), while ten independent GPT-5.6 samples all land on the most formal end.

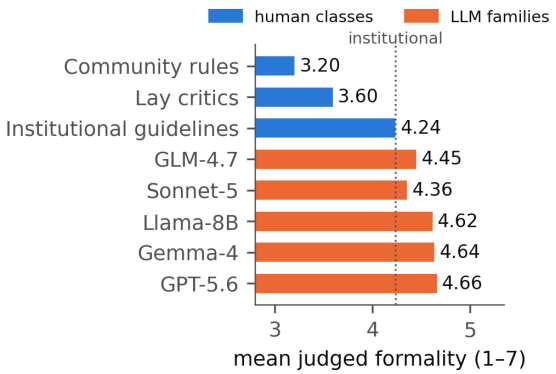


Figure 4: Mean judged formality of criterion names by speaker class. Every LLM family names at or above the *official* register; none approaches the register of the communities whose standards they are asked to apply.

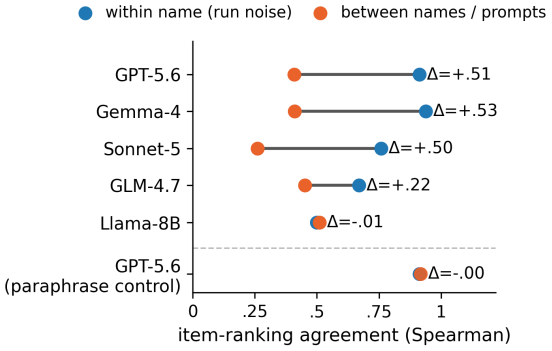


Figure 5: Swapping only a criterion's name collapses agreement (orange) below run-to-run reliability (blue).

the human inventories to .19 in LLM samples ($p \leq 1.8 \times 10^{-3}$), and mean per-construct name entropy falls from 2.66 bits (human) to 2.51. Secondly, we observe that the terms LLMs tend to converge on are the more formal terms. Every LLM family rearticulates concepts at the formality of the human *institutional* register, and **no LLM family approaches the plainness of the communities whose standards they are asked to apply** (Fig. 4). Where human {institutions, communities, individuals} write criterion names at mean judged formality {4.24, 3.20, 3.60} on the 1–7 scale, the five LLM families emit names at 4.36–4.66; every LLM family is above the human institutional mean. For instance, asked to rename the humor construct whose community names include *where who what* and *specificity bomb*, GPT-5.6 answers *specificity* or *concrete specificity* in ten of ten samples (Fig. 3).

Finding #2: Naming choices have robustness consequences for LLM behavior. In our next experiment, we prompt an LLM to label elements with identical prompts, with everything but the name of the construct swapped. A judge scores a fixed item set while only the criterion's name changes among that construct's variants. We find that using different names for the same construct in prompts lowers the agreement between the judge's item rankings across all models, as shown in Figure 5: GPT-5.6's rankings agree less across the two names (.41) than across its own repeated runs under one name (.91), $\Delta = +.51$; Gemma-4 loses +.53, Claude +.50, and GLM-4.7 +.22 ($p < 10^{-3}$). Only Llama seems to have very little effect ($\Delta = -.01$, $p = .64$), but its within-name agreement (.50; chance is 0) is also low, in fact, the reliability for swapped names is *not* lower for Llama (.51) than for the other models (.26–.45).

We test a control condition: paraphrasing the entire surrounding prompt with the name held fixed; we find this produces no disruption at all. Interestingly, we observe nearly no change in reliability from the within-name condition across all model families; the LLMs do seem robust to prompt perturbations *besides* those pertaining to

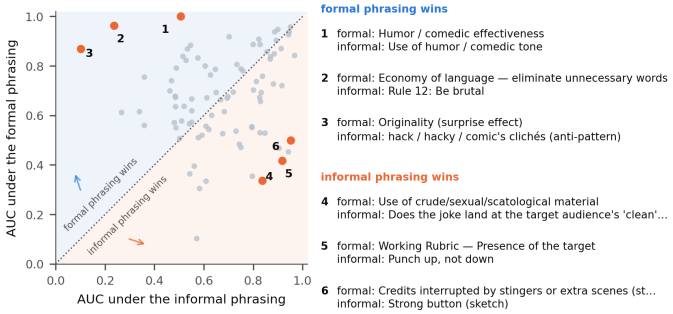


Figure 6: Agreement with community silver labels under different phrasings. Each point is one humor construct, scored under its most informal (x-axis) and most formal (y-axis) phrasing. The key lists both phrasings of the numbered points. Points scatter symmetrically about the diagonal: register does not predict the more valid instrument.

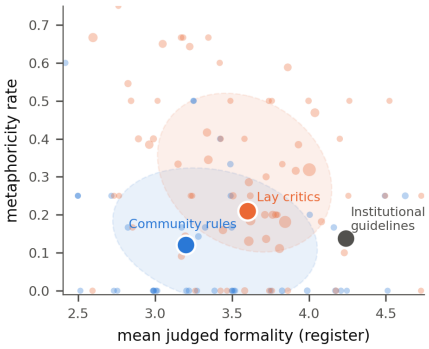


Figure 7: Register and Metaphoricity. x-axis, plainer↔formal: rated 1–7 (1 casual everyday speech, 4 neutral professional, 7 technical-institutional). y-axis, all literal↔all figurative: labeled figurative or literal. Large points are class means. Community rules are lowest in register and metaphoricity. Formality is highest for institutional sources.

the construct name. Judges condition on the lexical identity of the criterion, not it's paragraph.

Two human annotators labeled 97 of these name pairs blind (Appendix G); on pairs they judge to name the same construct, the swap disrupts rankings by Δ=.51–.60. A construct's names move a judge's rankings by about half the agreement scale, mostly because they differ in the requirement they ask the judge to check. Swapping names disrupts rankings more than paraphrasing in ten domains out of 11 for at least one judge family (Table 16).

Finding #3: The *correct* register of a phrase is difficult to predict a-priori Upon observing the large effects of these naming differences, a natural first question to ask is whether certain registers are reliably better than others; what should system designers default to? To answer this question, for each construct we compare the AUC under the *most* informal version with the AUC under the *most* formal version. As shown in Figure 6, we do not observe any meaningful differences. Scoring each construct under its most formal and most informal phrasing results in even wins; a single phrasing choice can affect silver label scores by as much as .73 AUC (Fig. 6; Appendix G.5).

A natural follow-up is whether matching the community's own register helps: is the phrasing closer to the community's dialect the better instrument? Surprisingly, it is not. When scoring 85 constructs under their least and most formal phrasings, (Appendix G.5), register distance does not predict which phrasing scores higher (ρ=.01); informal version only wins 45 to 39.

LLM consistency itself shows a register impact. Every family is more reliable within-runs under an institutional-register than under a more vernacular register (e.g. GPT-5.6 .935 vs. .892), but the asymmetry is small. LLM judges may be marginally steadier in the register they themselves prefer to speak. Because silver labels were drawn from comments, positive statements may have been overemphasized; however, we believe this is still a useful indicator of norm transmission in human communities. We address the limitation identified in Section 2 about this dataset: because the labels are weakly supervised (an LLM links human comments to our hierarchy), there may be biases in how well it links: LLMs overuse a recognizable lexical dialect (Juzek and Ward, 2025; Liang et al., 2024), so human phrasings closer to that dialect would be matched more often and accumulate more labels, favoring LLM-preferred wording in the validity comparison. However, our null finding, here, runs against this: the formal, LLM-preferred phrasings show no validity advantage (44 wins against 44), which is the opposite of what a dialect-matching bias would produce.

5 Linguistic Associations

Finding #4: Register and metaphoricity differ by source. How a criterion is worded is associated with the class of source that states it. Institutional guidelines are the most formal, lay critics

in the middle, and the rules communities write for themselves are the plainest class in nearly every class (Appendix H.1). Metaphoricity also differs by class: community rules are the least figurative (12% of names against 21% for lay critics; Appendix H.1) (Fig. 7). We read this as two independent axes of speaker position: register correlates with institutional voice, while metaphoricity correlates with function. Language that shares adjudication guidelines, rather than descriptive suggestions, is less figurative at any register: when matched on judged formality, community rules are less figurative than lay speech at four of the five formality levels (for analysis, see Appendix H.1).

Finding #5: Norm specificity is associated with enforcement. Criteria differ in how many scenarios they are applied to: *general* norms apply across contexts, *specific* norms tend to be document-particular, and *hyper-specific* norms apply to localized patterns in documents (Fig. 2; for analysis, see Appendix H.2). We have two findings: first, that again, the more specific a norm is, the more it is tied to standards that are adjudicated by human judges. A community's written rules are 3.5× more likely to be hyper-specific than individual speech[8] (17.4% vs. 5.0%; Appendix H.2)[9]

Finding #6: Choices of lexicon reflect community identity. The reuse of an established name, we find, is rare in every domain: most naming events introduce a phrasing not already attested (Appendix H.3). The same construct, we find, is lexicalized differently by different communities: e.g. the recognition moment in humor (Figure 1) is expressed as a *get it moment* in a popular article (NPR, 2014) and the Greek *anagnorisi* in academic literary theory (Aristotle, c. 335 BCE/1902). Communities that tend to vary the most are creative writing, journalism, and humor (renaming rates .50, .48, .46; Table 20). Domains whose criteria come from more distinct websites rename more: with the same number of documents drawn from each domain, the number of distinct source sites correlates with the renaming rate at ρ=+.57 (p=.07). Appendix H.5 reports the earliest dated source of each theme.

[8]This can be seen as a version of the rules-vs-standards trade-off (Kaplow, 1992; Ehrlich and Posner, 1974)

[9]This is consistent under multiple relabeling runs: 3.6× under GPT-5.6 and 3.7× under Claude, relabeling 1,000 names per class.

6 Discussion

Our results indicate that when we design rubrics for an LLM, our choice words has an impact; the judge is sensitive to that choice and less sensitive to the wording around it. A construct's names effect half the total agreement, mostly because they differ in the requirement they state; when two names also state the same checkable requirement, the effect is small. Since LLMs tend speak and prefer the institutional register, while the communities whose work they judge write their own standards in plainer, more enforceable language, the names LLMs choose differ in register from the names communities use, although we find no evidence that this difference by itself lowers agreement with community judgments. While there does not seem to be a clear pattern a-priori that can help us predict which prompts work well, our results suggest that this is a potent level to check. Indeed, rubrics designers and model training developers should test the stability and community agreement of each criterion name before relying on it. Additionally, an LLM's preference to speak in the formal register likely presents unexplored challenges to prompt-optimization frameworks; in most frameworks (Spangher, 2026; Yang et al., 2024; Opsahl-Ong et al., 2024; Agrawal et al., 2025), a proposer LLM suggests and refines prompts for a generator LLM. As LLMs prefer institutional registers, broader prompt spaces should be explored.

7 Conclusion

We presented the first cross-domain census of quality-focused evaluative criteria, mined to a measured notion of saturation, organized into a validated four-level hierarchy, and analyzed as a linguistic system. The lexicon is socially stratified: the class of source that states a norm is associated with its register, metaphoricity, and specificity. LLM judges live inside this system — they speak its most institutional stratum, and the single lexical choice of a criterion's name changes what they measure: a construct's names move a judge's rankings by about half the agreement scale, mostly because they differ in the requirement they state. We release the census, hierarchy, corpora, and silver labels as a benchmark for studying how written evaluative criteria are worded and applied.

Limitations. All measurement here is LLM-judged; we mitigate with external gold validation per instrument, camouflaged anchors in every batch, and cross-family agreement checks, but construct identity at the construct grain remains instrument-relative — a stricter or looser merge threshold shifts inventory sizes, though not the closed-head/open-tail shape. Extraction recall is .69 against enumerable gold on structured pages but .28 against a multi-family panel union on prose, so absolute coverage is well below the census's observed size even where sampling has saturated; the domain comparisons we draw are unaffected, since one extractor is applied identically to all domains (Appendix A.3). Silver labels are community outcomes, not ground truth for quality. And our census covers the *written, public* lexicon: criteria communities never write down are invisible to it by construction.

Four further limits deserve their own statement. (1) The cross-domain correlations (codification with naming coincidence, and their relatives) are descriptive: eleven domains are not independent draws, and no cross-domain directional test was preregistered; the preregistered tests in this paper are within-domain. (2) The behavioral experiments are narrower than the census. The name-swapping and silver-label experiments concentrate on humor, the domain with the densest community commentary; whether their effect sizes transfer to the other ten domains is untested, except that the name swap replicates in creative writing (Appendix G). (3) The register instruments score criterion *names*, not full rubric texts. A rubric's surrounding prose has a register of its own, which we manipulate only in the paraphrase control; the causal claim is about the name. (4) The census is conducted in English, with older sources read through their standard English translations. Register strata, and the Latinate/Germanic contrast in particular, are language-specific, so the social stratification we describe is a fact about the English evaluative lexicon rather than about evaluation as such.

Reproducibility and data. Appendix I lists the released artifacts, model identifiers, prompts, and raw judge outputs; Appendix J lists every preregistration with its freeze date and status. The raw runs of the first name-swap and form-effect experiments were lost before release, and both results are re-established on persisted data. We release source URLs, extracted criteria, and evidence quotes rather than full page text, with commenter usernames removed.

References

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2025. Gepa: Reflective prompt evolution can outperform reinforcement learning. *Preprint*, arXiv:2507.19457.

Adam Alami and Oliver Krancher. 2022. How scrum adds value to achieving software quality? *Empirical Software Engineering*, 27(165).

Aristotle. c. 335 BCE/1902. *Poetics*. Macmillan, London. Original work composed c. 335 BCE. Translated by S. H. Butcher as *The Poetics of Aristotle*, Macmillan, 1902; Project Gutenberg eBook #1974. Recognition (*anagnorisis*) is treated in ch. 11, 1452a.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL*.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of COLING*.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW).

Juliet M. Corbin and Anselm Strauss. 1990. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Sage.

David Corson. 1985. *The Lexical Bar*. Pergamon.

Lee J. Cronbach and Paul E. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302.

Renata Gonçalves Curty, Rebecca Greer, and Torin White. 2021. Teaching undergraduates with quantitative data in the social sciences at university of california santa barbara.

Gerard de Melo. 2014. Etymological Wordnet: Tracing the history of words. *Proceedings of LREC*.

Bradley Efron and Ronald Thisted. 1976. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447.

Isaac Ehrlich and Richard A. Posner. 1974. An economic analysis of legal rulemaking. *The Journal of Legal Studies*, 3(1).

Casey Fiesler, Joshua Jiang, Rebecca McCann, Kyle Frye, and Jed R. Brubaker. 2018. Reddit rules! Characterizing an ecosystem of governance. In *ICWSM*.

Sigmund Freud. 1905/1916. *Wit and Its Relation to the Unconscious*. Moffat, Yard and Company, New York. Original work published as *Der Witz und seine Beziehung zum Unbewussten*, Deuticke, 1905. Translated by A. A. Brill, 1916.

Christina L. Gagné, Thomas L. Spalding, and Daniel Schmidtke. 2019. LADEC: The large database of English compounds. *Behavior Research Methods*, 51.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of EMNLP*.

Andrew Gelman and Donald B. Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Irving J. Good. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of LREC*.

Tom S. Juzek and Zina B. Ward. 2025. Why does ChatGPT "delve" so much? Exploring the sources of lexical overrepresentation in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*.

Louis Kaplow. 1992. Rules versus standards: An economic analysis. *Duke Law Journal*, 42.

Neha Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *NAACL*.

Scott Kirkpatrick, C. Daniel Gelatt, and Mario P. Vecchi. 1983. Optimization by simulated annealing. *Science*, 220(4598):671–680.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Domain agnostic real-valued specificity prediction. In *Proceedings of AAAI*.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Proceedings of AAAI*.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *International Conference on Machine Learning (ICML)*.

Travis Lloyd, Joseph Reagle, and Mor Naaman. 2025. "there has to be a lot that we're missing": Moderating AI-generated content on Reddit. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)*.

Annie Louis and Ani Nenkova. 2011. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of IJCNLP*.

Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barrett, and Ruth Reeves. 1998. NOMLEX: A lexicon of nominalizations. In *Proceedings of EURALEX*.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL Workshop on Multiword Expressions*.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of *SEM*.

National Center for O*NET Development. 2025. O*net 30.0 database. Technical report, U.S. Department of Labor, Employment and Training Administration. Work Activities content model: Task Statements, Detailed, Intermediate and Generalized Work Activities.

NPR. 2014. 'new yorker' cartoon editor explores what makes us get it. NPR, All Things Considered, 24 March 2014.

Krista Opsahl-Ong, Michael J. Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of NAACL-HLT*.

Jim Pitman and Marc Yor. 1997. The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2).

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1).

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of IJCNLP*.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*.

Eleanor Rosch, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology*, 8.

Abhilasha Sancheti, Aparna Garimella, Balaji Vasan Srinivasan, and Rachel Rudinger. 2022. Agent-specific deontic modality detection in legal language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2024. Fora: A corpus of multi-party discussions with annotation of discussion facilitation strategies. In *ACL*.

Alexander Spangher. 2026. Articulation upper bounds: Scaling behaviors for prompts. In preparation.

Gerard Steen, Aletta Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476).

Asif Kamal Turzo, Fahim Faysal, Ovi Poddar, Jaydeb Sarker, Anindya Iqbal, and Amiangshu Bosu. 2023. Towards automated classification of code review feedback to support analytics. In *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*.

Sriram Venkatapathy and Aravind Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proceedings of HLT-EMNLP*.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. Technical report, University of Pennsylvania. LDC2019T05.

Galen Weld, Amy X. Zhang, and Tim Althoff. 2022. What makes online communities 'better'? Measuring values, consensus, and conflict across thousands of subreddits. *ICWSM*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. In *International Conference on Learning Representations (ICLR)*.

Appendix contents

A Corpus construction **12**
A.1 Collection pipeline 12
A.2 Example page parses and units of analysis . 12
A.3 Extraction validity 12

B Saturation: model, measurement, and certification **17**
B.1 The franchise model 17
B.2 Saturation across strata 17
B.3 Ensuring saturation: mixing, not depth . . . 18

C Hierarchy construction **20**
C.1 The algorithm 21

D External validation against human-coded hierarchies **21**
D.1 What is available, and what is not 22
D.2 Design 22
D.3 Results 23
D.4 Baselines 24
D.5 Partition-level agreement 25
D.6 Limits 25

E Behavioral dataset construction **25**

F Instruments **26**
F.1 Instrument reliability 26

G Experiments on LLM judges: details **27**
G.1 Rearticulation 27
G.2 Name swap: design and per-family results . 27
G.3 Matched-meaning pairs and human annotation 29
G.4 Replication across domains 30
G.5 Form effects and register matching 32

H Linguistic associations: details **33**
H.1 Register by speaker class and domain 33
H.2 Specificity by speaker class 34
H.3 Codability of the naming process 35
H.4 Codification level across domains 35
H.5 Norm age per domain 37

I Released artifacts **38**

J Preregistration records **39**

K Hierarchy subtrees by domain **39**

L Annotated name pairs **43**

A Corpus construction

A.1 Collection pipeline

This section documents the web search of §2. Figure 8 gives the pipeline end to end, from seeded search through to the canonicalized census banks.

Figure 9 breaks the 59,028 successfully typed source pages down by page type and domain — the provenance mix is a register signature in its own right, with adjudicated domains drawing on formal guidelines and standards while creative domains draw on wikis, how-tos, and blogs.

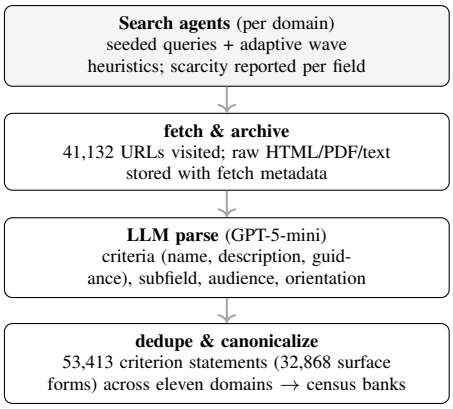


Figure 8: The web-retrieval pipeline behind the census. Later lay-critic waves reuse the same pipeline with Good–Turing missing mass as the adaptive stopping rule.

A.2 Example page parses and units of analysis

This section shows what a single parsed record looks like at both ends of the pipeline, so the reader can see exactly what the extractor produces and what provenance rides on every criterion.

Two complete (abridged only in long text fields) parse records. Every census criterion carries the first kind of provenance; every lay-corpus criterion carries the second, including the extracted author and the coded affiliation type.

Units of analysis: one criterion end to end. The paper uses four units. A *record* is one extracted statement with its parsed name, description, and guidance; its *head term* is the token in the source text that the lexical analyses count; a *surface form* is the normalized head term; a *construct* groups surface forms that state one requirement; and the behavioral experiments score a coarser construct partition built so that each construct has enough labeled documents. Table 2 follows one humor criterion through all of them.

A.3 Extraction validity

We measured the census extractor's precision and recall against external gold rather than only self-referential fidelity checks.

Item-level P/R on a known-denominator gold. Sixty subreddit pages (404 enumerable gold rules, stratified by community size) were reconstructed in native layout and run through the production extractor blind; extracted items were matched to gold rules by an anchor-gated judge. Precision against gold is 1.0 (136 of 136 production extractions

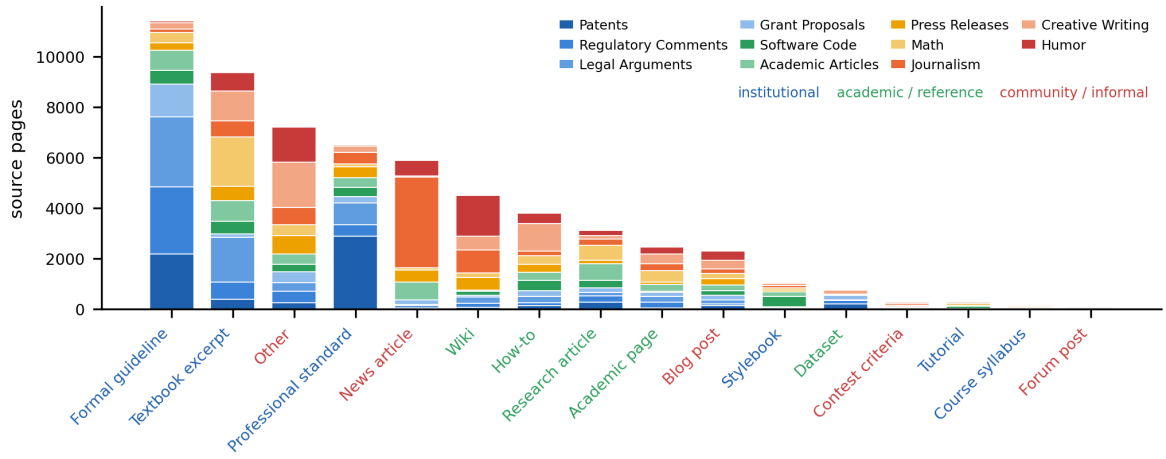


Figure 9: Where the criteria come from: source pages by page type, stacked by domain (59,028 successfully typed pages). The provenance mix is itself a register signature: adjudicated domains (patents, regulatory comments, legal argument, grants) draw 40–50% of their criteria from formal guidelines and professional standards, while humor and creative writing draw theirs from wikis, how-tos, and blogs — the same institutional-to-informal gradient that the register instruments recover from the criterion names themselves (§H.1 — convergent evidence from an independent signal).

match a real rule, one-sided 95% upper bound on the error rate 2.2%; 398 of 398 for the all-norms pass, bound 0.75%) and an all-norms extraction pass recovers essentially every rule on the page (.995, 181 of 182, on quality-rule gold, .98 (217 of 222) on governance gold). The production pass, which targets quality criteria only, recovers .69 [.62, .75] of quality-rule gold: since the all-norms pass shows the extractor *sees* these rules, the loss is localized entirely to the quality-vs-governance boundary decision, not to detection. Governance leakage into the quality pass is .05 of governance rules (11 of 222), which is .08 of the production pass's own items. Matching a real rule is not the same as reproducing it: two judge families rate .74 and .79 of the production extractions as faithful to the gold wording and .98–.99 as faithful or minor drift, with one or two of 136 unfaithful. Every batch contributing to these figures passed its own anchor gate; one batch that initially failed was rejudged under a fresh gate rather than dropped.

The matching judgment replicates exactly across judge families. Re-running the anchor-gated match on one batch (86 gold rules) with a second model family reproduced all 86 match votes, in both the quality and all-norms passes, and reproduced the frozen votes of the original run identically — so the recall figures above are not a judge-family artifact. What does *not* replicate as cleanly is the quality-vs-governance label itself: the two families each dissent from the frozen labels on 8 of 86 rules, on *disjoint* sets (agreement .91 with the frozen labels each, .81 with each other), and their dissents are systematic rather than noisy — one family reads duplicate-posting rules as governance, the other reads link-hosting and topic-scope rules that way. The boundary, not the detector, is where the residual disagreement in this appendix lives.

Page-level detection. All 60 rule-bearing pages are detected. Against 30 near-miss negatives (real community description pages screened norm-free), the false-positive rate is $3/30$, all low-confidence borderline scope statements; far negatives (unrelated prose) produced $4/30$ flags. Blind adjudication of all 30 far pages — each judged from its text alone, with verdicts frozen before the detector flags were opened — found that 16 carried genuine norm statements and had been mislabeled norm-free, including all four flagged pages. On the corrected labels the far-negative false-positive rate is $0/14 = 0$: the flags were detections on mislabeled positives, not extractor errors. Near negatives therefore remain the harder stratum (10% against 0%), the direction we preregistered. The size of the label correction is itself the caution here: a negative stratum assembled by topical distance from rule pages is not reliably norm-free, and we report the far rate only on the adjudicated denominator.

(a) A professional-standard page (census pipeline)


```
{
  "url": "https://www.kirkusreviews.com/indie/",
  "source_file": "kirkus_indie_reviews.html",
  "fetched_at": "2026-05-10",
  "title": "Kirkus Reviews - Fiction Reviewing Criteria",
  "orientation": "professional_standard",
  "community": "Kirkus Indie editorial program",
  "intended_audience": "self-published, indie, and unpublished authors seeking a
     paid Kirkus Indie book review",
  "subfield": "kirkus_indie_review_submission",
  "subtask_keywords": ["kirkus", "indie", "book_review", "self_published",
     "submission_guidelines"],
  "source_text": "# Kirkus Reviews - Fiction Reviewing Criteria ## Review Format
     - 250-300 words per review - Plot summary + concise unbiased opinion
     (positive, negative, or mixed) - Identifies both strengths and weaknesses
     ## Elements Reviewers Assess 1. Plot development ...",
  "rubrics_metrics": [
    {"name": "Summary content is fundamental",
     "description": "Summary content is a fundamental element of Kirkus' reviews,
        because Kirkus aims to help consumers, booksellers, librarians,
        publishers, agents, movie producers, and other influencers discover new
        books and authors.",
     "guidance": "Reviews include enough summary content to give context to
        criticism and help readers decide if the book will interest them."},
    {"name": "How Kirkus chooses a reviewer",
     "description": "A Kirkus editor analyzes your book and considers elements
        such as content, style, genre, and sub-genre. We then match your book
        with a reviewer who is a content and genre expert...",
     "guidance": "Editors assign titles based on a reviewer's professional
        experience, expertise, and reading interests."},
    "... 18 more criteria"
  ]
}
```


(b) A lay-corpus criterion with author attribution (W13 pipeline)


```
{
  "url": "https://www.reddit.com/r/nonprofit/comments/[redacted]/",
  "task": "grant-funding",
  "subfield": "nonprofit grant applications",
  "head_term": "powerful data visuals",
  "key_terms": ["tell your story", "stronger application"],
  "description": "Powerful data visualizations can tell the story and strengthen
     an application.",
  "quote": "if you can develop some powerful data visuals to tell your story, it
     can make for a much stronger application",
  "author_name": "[redacted]",
  "author_affiliation": "[redacted]",
  "author_type": "credentialed_individual"
}
```


Figure 10: Example provenance records. (a) The parser's output for a professional page. Of the 74 Kirkus records it produced, two entered the census (Dialogue, Originality); the two shown here did not, and two judge families label them as review procedure and as a requirement on the review rather than on the book; (b) a lay-corpus criterion with the extracted author, free-text affiliation, and the anchor-gated affiliation-type code (`institution` / `credentialed_individual` / `lay_individual`). The commenter's username, inferred affiliation, and thread identifier are redacted; the release removes usernames throughout.

Unit	Value
Source page	`artsundivided.com/post/2022-02-16-an-uphill-battle-steve-martins-book-born-standing-up` (blog post; provenance rung 4, practitioner; raw HTML)
Source passage	"To implement the new concept called originality that I had been first introduced to in Showmanship for Magicians, and was now presenting itself again in my classes in literature, poetry, and philosophy, I would have to write everything in the act myself. Any line or idea with even a vague feeling of familiarity or provenance had to be expunged. There could be nothing that made the audience feel they weren't seeing something utterly new."
Record key	`humor::raw::67f58d957c86.html::1`
Parsed name	"Originality of material" (a paraphrase of the source, both judges)
Parsed description	"There could be nothing that made the audience feel they weren't seeing something utterly new; he had to write everything in the act himself and expunge any line…"
Head term	"originality" (verbatim in the source, both judges; the token the lexical counts use)
Surface form	`originality`, form node 2636 ("Basic Originality")
Construct (Table 1)	`humor_R1_g389` "Basic Originality"
Construct (scoring partition)	`a54` "Originality and genuine craft over formula or clout-chasing" (10 census leaves)
Seated comment	"this is very original" (r/standupcomedy; production match, high confidence)
Silver label	that document is positive for `a54` (746 positive documents)

Table 2: One criterion traced through every unit of analysis. The relation between the two partitions is many-to-many (the ten census leaves of `a54` sit in several construct nodes), so the scoring construct is reached through the record, not through the Table 1 construct.

Prose-page recall. Structured rule lists are the easy case. On real prose pages from the census, we benchmark the production extraction against the union of exhaustive multi-family enumeration passes. Raw union recall is grain-mismatched — the reference passes enumerate micro-criteria, and production extraction emits criterion-grain units — so we collapsed the panel's adjudicated micro-criteria into criterion-grain units under an anchor-gated same-requirement relation. Two of the fifteen adjudicated pages were found on audit to have been served to the panel and the extractor as unparsed PDF byte streams rather than as text; both scored 100% recall on two- and three-unit denominators, the signature of a spurious ceiling, and both are excluded here. On the remaining thirteen pages, 534 micro-criteria collapse to 203 criterion-grain units (2.63 micro-criteria per unit) and recall is $56/203 = .28$ at matched grain against $.15$ raw. This is a *ceiling-relative* number and not absolute recall: the panel union is what several model families could find, not everything norm-bearing on the page, so $.28$ bounds recovery against panel-findable content. It is also the loosest number in this appendix, and the gap between it and the $.69$ measured against enumerable gold is informative rather than contradictory — gold rule pages state their norms explicitly and in a list, while prose pages diffuse them, so extraction recall is itself a function of how explicitly a source codifies. The items the panel found but the production pass missed are *not* lower-register than the ones it recovered: judged formality of missed items 5.27 (n=149) against 5.36 for found items (n=85), one-sided Mann–Whitney p=.44, A=.494 (anchor gate passed). Our preregistered prediction was that misses would skew low-register; it does not hold, which together with the item-level result above places the loss at the quality-versus-governance boundary rather than at register or detection.

We twice attempted an experimental version of that register test and report both failures. A synthetic hard-page test — gold rules woven into distractor prose — was *voided by its own adversarial realism gate*: every generated page was identifiable as synthetic. A splice-based redesign, holding prose and rule sentences real so that only their combination is synthetic, and running the realism gate *first* under a registered stopping rule, was voided the same way: $0/20$ spliced pages passed as real while $20/20$ untouched controls did, so no downstream extraction was run and no recall number exists to report. The register prediction is therefore *untested*, not falsified; the observational null above is the only evidence we have on it. We note the recurring obstacle for anyone attempting this design: current judges detect spliced provenance essentially perfectly, even when every sentence in the page is real.

What the saturation certificate does and does not cover. These figures compose with, rather than duplicate, the stopping apparatus of Appendix B.3. The re-heating probe certifies that the *sampler* is not narrow: it draws a wave from a previously unsampled source stratum and asks whether novelty jumps. It cannot detect a narrow *extractor*, because a blind spot in what we pull off a page is invariant to which pages we fetch. This section is the missing half of that certificate, and the two together scope the census claim: the corpus is mined to saturation over the criteria that are reachable online *and* visible to the extractor. Read that way, the recall figures set the census's absolute coverage below its observed size; cross-domain comparisons rest on one instrument ap-

plied identically to every domain, with one residual caveat, that recall itself covaries with codification, taken up at the end of this appendix. The results in this paper are correspondingly comparative: shares, ranks, and contrasts across domains, not absolute inventory sizes.

Concept coverage. Against the published macro-norms of Chandrasekharan et al. (2018), our community-adjacent inventories cover .83 (humor) and .92 (journalism) of macro-norm concepts; quality-focused domains further from community governance cover .67 — expected, since macro conduct norms are governance, not quality.

What these measurements establish. Taken together, these measurements give the extractor a simple profile: precision-safe and recall-limited, with recall a function of how explicitly a source codifies (1.0 precision and a 0% adjudicated far-negative rate; .69 recall where norms are stated as enumerable lists, .28 where they are diffused through prose). That coupling means extraction recall is not independent of this paper's subject matter: the census recovers a larger share of a domain's criteria the more codified the domain is, so even one instrument applied identically to every domain does not behave identically on every domain, and extraction error is correlated with the codification gradient of Appendix H.4. The direction of the resulting bias, however, is conservative for what we report. Under-recovering the diffuse tail of a weakly codified domain retains that domain's most explicitly stated criteria, the subset most likely to have conventionalized names, which can only push its observed naming coincidence *up*, toward the well-codified domains; the reported association (low institutional share with low coincidence) is attenuated by this bias, not produced by it. The one direction in which extraction error inflates rather than deflates (spurious singletons, which drive the Good–Turing and Chao1 estimates) is governed by precision, not recall, and precision is the quantity these legs pin at 1.0. What the legs do not settle is whether visibility is register-biased, i.e. whether casually phrased norms are systematically the ones in the missed mass; the observational comparison above is null, and the experimental version of that question remains open.

What the extractor extracts. On a stable-hash sample of 50 census criteria per domain (n=550), two judge families labeled what each criterion governs: the submitted work, commentary on it, a procedure for evaluating it, the conduct of people, or something else. Criteria not about the work are .10 [.08, .13] (Claude) and .08 [.06, .11] (GPT-5.6), κ=.50. By domain they are concentrated in regulatory comments (.29 and .20) and patents (.20 and .12), where formal requirements (fee rules, statutory listing criteria, compliance with a funding announcement) can be read either as properties of the submitted document or as process, and they are near zero in software code and creative writing. The census carries no evaluand field; these shares bound how much of each domain's inventory concerns the work itself.

Where the names come from. The same sample was judged for name provenance against each record's evidence quote. The parsed name is the parser's wording more often than the source's (source wording .25, paraphrase about half, generated .08, by both judges), but the parsed name is not what the lexical analyses count. They count the head term, the token that the verbatim pass attributes to the source, and head terms are verbatim in the source in .96–.97 of records (n=185); lay-critic and community-rule terms, judged against their quotes, are verbatim in .98–1.00 (n=100 each). Table 2 follows one record from passage to head term.

Documents written by the collecting LLM. During collection the search agent sometimes wrote a markdown document in place of a fetched page: a summary of a long discussion thread, an abridged copy of a structured guideline, or a synthesis of several pages (one such file is titled as a synthesis of the most-upvoted advice in a joke-writing thread, and its recorded source URL is not a permalink). These documents went through the same extractor as fetched pages. They account for 9,666 of the 53,413 census criteria (18%) and for 5,118 of the 17,453 named rows that enter the lexical analyses (29%); Table 3 gives the share per domain, from .07 of named rows in software code to .60 in legal arguments. Judged on 66 such documents, Claude reads .67 and GPT-5.6 1.00 of them as syntheses in the LLM's own words rather than transcriptions (the disagreement is over abridged copies of structured guidelines), so between .20 and .29 of the counted names are verba-

Domain	share (named)	n named	share (all)
Software Code	.07	5,497	.05
Academic Articles	.14	1,116	.10
Patents	.24	1,054	.11
Grant Proposals	.35	701	.19
Regulatory Comments	.38	573	.27
Creative Writing	.39	1,748	.20
Math	.39	1,023	.14
Humor	.40	1,491	.21
Journalism	.44	1,002	.20
Press Releases	.52	2,161	.39
Legal Arguments	.60	1,087	.37
All	.29	17,453	.18

Table 3: Share of each domain's census that comes from documents the collecting LLM wrote rather than fetched. *share (named)*: share of named rows, the rows that enter the lexical analyses, with their number n; *share (all)*: share of all census forms.

tim from text that an LLM wrote. The exposure is uneven by source class: 14–16% of named rows from institutional sources against 37–49% from academic, practitioner, and community sources.

Three analyses are affected differently. The register contrast of §5 compares the institutional class with the lay-critic and community-rule corpora; the latter two are separate collections whose terms are verbatim from human text, so the contrast is exposed only through its institutional class, where 12 of the 113 names occur only in LLM-written documents and the class mean is 4.24 with or without them (3.86 [3.44, 4.28] on the 36 names that never occur in such documents, a subset that also drops the most widely used names). The coda-bility table (Appendix H.3) is exposed at the full share; recomputed without these documents, the domain ordering of reuse given a known construct is unchanged (Spearman .96 across domains): humor .17→.12, creative writing .18→.16, math .19→.17, software code .64→.65, with patents (.42→.53), press releases (.33→.40), and regulatory comments (.34→.64) rising on 40–60% fewer records. The hierarchy inherits these records as surface forms, so constructs whose only support is an LLM-written document are the ones at risk; the release flags every record by source layer so that any analysis can be rerun on fetched pages alone.

B Saturation: model, measurement, and certification

This section collects the sampling machinery in one place: the hierarchical model that defines novelty at each grain (§B.1), the measured per-domain rates and missing mass (§B.2), and the devices that certify a plateau as population coverage rather than sampler narrowness (§B.3).

B.1 The franchise model

Model the census as a Chinese Restaurant Franchise (Teh et al., 2006): each criterion *draw* enters the surface-form restaurant; each *new* surface form is in turn a customer of the construct restaurant; each new construct a customer of the theme restaurant; each new theme of the category restaurant. At level ℓ with n_ℓ customers and k_ℓ occupied clusters, the Pitman–Yor predictive (Pitman and Yor, 1997) is

$$\Pr(\text{new}) = \frac{\theta_\ell + d_\ell k_\ell}{\theta_\ell + n_\ell}, \quad \Pr(j) = \frac{n_{\ell j} - d_\ell}{\theta_\ell + n_\ell},$$

with the CRP as the $d_\ell{=}0$ case. Because level ℓ's customers are exactly level $(\ell{-}1)$'s new clusters, per-draw novelty chains multiplicatively: $m_\ell^{\text{draw}} = \prod_{j \le \ell} m_{j|j-1}$. Each conditional is estimated nonparametrically by the Good–Turing plug-in (Good, 1953) $\hat{m}_{\ell|\ell-1} = f_1^{(\ell)}/k_{\ell-1}$ (singleton clusters over customers); at fitted parameters the two estimators agree (peer review, construct level: GT .134 vs. PY .130). The fit itself: per level, the discount d_ℓ and concentration θ_ℓ are chosen to best reproduce the observed growth of cluster counts over draws, and the fitted predictive is read out at the census's terminal (n_ℓ, k_ℓ). The discount governs growth: $\mathbb{E}[k(n)] \sim n^d$ for $d > 0$ (power law — no finite saturation point) vs. $\theta \log n$ for $d{=}0$. Empirically the conditionals behave very differently by level as sampling deepens: $\hat{m}_{\text{form}|\text{draw}}$ falls .81 → .45 over the census; $\hat{m}_{\text{construct}|\text{form}}$ is nearly *flat* (.18 → .14 — the power-law signature, fitted $d{=}.76$); $\hat{m}_{\text{theme}|\text{construct}}$ is ≈0 throughout. The accumulation curves of Figure 11 are the cumulative view of the same data; the franchise conditionals are their derivatives.

B.2 Saturation across strata

This section shows the full per-domain saturation picture summarized in §2. Figure 11 plots the conditional novelty rates $p(\text{unseen} \mid \cdot)$ at each level of the hierarchy over the course of collection, and Table 5 gives the terminal Good–Turing missing mass at every grain: the surface level stays open, constructs settle to a low steady rate, and themes and categories exhaust.

Figure 12 stress-tests the closed head from below: 56% of newly collected lay criteria are constructs absent from the professional inventory, yet

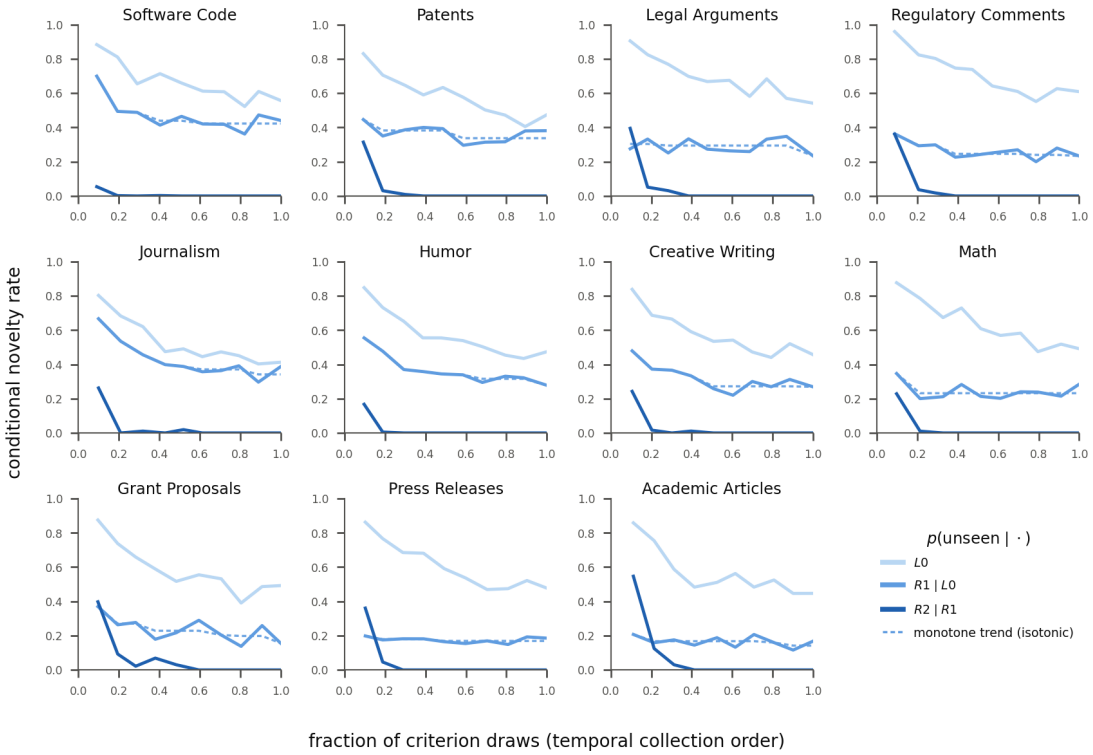


Figure 11: CRF conditional novelty rates up the ladder, per domain (windowed derivative of the accumulation data; panels ordered by construct openness). In every domain the pattern is identical: the surface rate (new form per draw) declines slowly and stays high; the construct rate (new construct per new form) settles onto a low *plateau* — the dashed isotonic guide shows the underlying slow decline the jagged empirical rate scatters around — and the theme rate (new theme per new construct) collapses to zero within the first fifth of sampling. Two cautions on reading the construct plateau as "saturation": (i) it is a nonzero *steady-state* discovery rate, not exhaustion — under the fitted Pitman–Yor law (d=.76, §B.1) construct novelty decays as a power law and never reaches zero, so new constructs keep arriving, just ever more slowly; and (ii) the plateau height is a property of the *search process*, not only the domain — it measures how fast our seeded collector re-encounters known constructs, which is why ensuring the plateau reflects the population and not the sampler needs the frontier controls of §B.3. Novelty propagates exactly one level up the franchise and then dies at the theme grain.

only 4.7% of those escape the frozen theme taxonomy. Lay discourse renews the surface of the lexicon without opening its conceptual territory.

B.3 Ensuring saturation: mixing, not depth

The novelty rates above are only trustworthy if our collector actually *reaches* the population it is trying to enumerate. A Good–Turing missing mass estimated from one narrow seam of sources will look reassuringly small long before the domain is covered: coverage statistics bound the probability of an unseen type *within the support of the current sampler*, and say nothing about mass the sampler cannot propose (Good, 1953; Efron and Thisted, 1976). Saturation is therefore a *mixing* problem, not a *depth* problem — deeper sampling from the same seeds cannot certify coverage, only broader seeding can. We use three devices, borrowed from species-sampling, MCMC convergence diagnostics, and stochastic search.

(1) Multiple search seeds as parallel chains. Each collection wave is launched from an independently seeded batch of search queries (distinct entry venues, source types, and query phrasings), the discovery analogue of running multiple MCMC chains from dispersed starts. We compare *between-seed* and *within-seed* construct discovery: if fresh seeds keep surfacing constructs that deep sampling of old seeds does not, the between-chain variance dominates and the process has not mixed, by analogy to the Gelman–Rubin diagnostic (Gelman and Rubin, 1992). The analogy is heuristic: we computed no $\hat{R}$ statistic, and the comparison of between-seed and within-seed novelty was read from the wave-level rates of Table 6 rather than from a stored diagnostic. Reported

Domain	$\hat{m}_{\mathrm{f}}$	$\hat{m}_{\mathrm{c}\mid\mathrm{f}}$	$\hat{m}_{\mathrm{t}\mid\mathrm{c}}$	new constr.
Software Code	.52	.36	.00	18.6%
Creative Writing	.42	.23	.00	9.5%
Grant Proposals	.46	.20	.00	9.0%
Humor	.43	.26	.00	11.1%
Legal Arguments	.56	.25	.00	14.0%
Math	.49	.20	.00	9.5%
Journalism	.40	.29	.00	11.5%
Regulatory Comm.	.59	.23	.00	13.8%
Patents	.44	.31	.00	13.7%
Academic Articles	.43	.13	.01	5.8%
Press Releases	.47	.16	.00	7.2%

Table 4: The hierarchy as a Chinese Restaurant Franchise: each level's clusters are customers of the level above, so novelty at level ℓ is a *conditional* missing mass and per-draw novelty chains multiplicatively. Columns: probability a draw is a new surface form; that a new form is a new construct; that a new construct is a new theme; and the chained per-draw new-construct rate. Novelty propagates exactly one level up the franchise and then dies ($\hat{m}_{\mathrm{theme}\mid\mathrm{constr}}{\approx}0$ everywhere) — the closed-head/open-tail law as franchise conditionals. A Pitman–Yor fit at the construct level (peer review: $d{=}.76$) implies power-law construct growth — a slow tail rather than a hard knee, consistent with the saturation campaign's measured 2.5 new constructs per 100 draws after adversarially hardened seating (vs. 5.8% naive plug-in: raw partitions overstate construct novelty $\sim$2.3$\times$). Subscripts: f form, c construct, t theme.

Domain	Form	Construct	Theme	Category
Software Code	.52	.21	.00	.00
Creative Writing	.42	.11	.00	.00
Grant Proposals	.46	.11	.00	.00
Humor	.43	.12	.00	.00
Legal Arguments	.56	.16	.00	.00
Math	.49	.11	.00	.00
Journalism	.39	.13	.00	.00
Regulatory Comm.	.59	.16	.00	.00
Patents	.44	.16	.00	.00
Academic Articles	.43	.07	.00	.00
Press Releases	.46	.09	.00	.00

Table 5: Good–Turing missing mass (f_1/N) by stratum and domain — the estimated probability that the next criterion drawn is a never-seen type at that grain. The surface grain stays open (.39–.59); constructs are closing (.07–.21); themes and categories are exhausted (.00 everywhere): the closed-head/open-tail law.

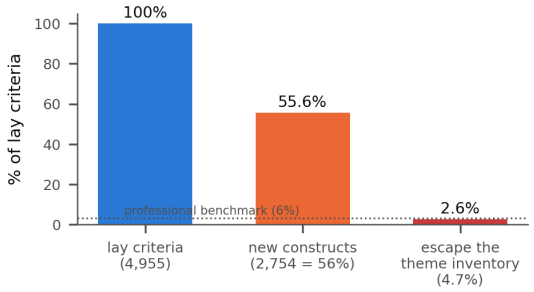


Figure 12: The closure law under lay stress-test: 56% of newly collected lay criteria are constructs absent from the professional inventory, yet only 4.7% of those escape the frozen theme taxonomy — effectively at the professional benchmark. Lay discourse renews the lexicon's surface, not its conceptual territory.

plateaus are only taken as population-level once between-seed novelty falls to within-seed noise.

(2) An annealing schedule over the frontier. The seed schedule narrows over waves, by analogy to annealing (Kirkpatrick et al., 1983): early waves run "hot" — deliberately diverse strata (many venue/funder classes) to spread support across the population — and later waves "cool" toward exploitation, deepening within the discovered clusters. High temperature guards against the local minimum of a single source community that mints its own dialect; cooling makes the missing-mass estimate converge. The temperature is the *diversity* of the current frontier, and our central empirical finding is that construct novelty tracks that temperature, not the number of draws. The schedule was set by hand per wave (the stratum column of Table 6); there is no temperature parameter or acceptance rule, so the term describes the order in which strata were sampled, not a specified algorithm.

(3) A re-heating probe as the saturation certificate. Because a plateau under a cooling schedule can be either true coverage or a frozen frontier, we certify it by *re-heating*: injecting one fresh, previously unsampled stratum and testing whether the construct rate jumps back above the stop threshold (Fig. 14). This turns saturation into a falsifiable prediction. Grant proposals fail the probe — adding a new funder class (defense, equity, climate) drives the rate from 1.4 to 2.8 new constructs per 100 draws, while a matched control that resamples *old* classes stays at 1.4 — so grants are *not* saturated: mission-specific criteria re-open the frontier. The rise is measured on one wave of 566 draws (16 new constructs; 95% interval 1.8–4.5 per 100), and the control is the following wave rather than a concurrent one, so the difference (1.4 points, interval -0.3 to 3.2) is suggestive rather than decisive. Rates in this appendix are new constructs per 100 draws after adversarial re-seating; the frozen campaign rule was a rate below 2.0 for two consecutive waves, and the body's $\lambda{=}.3$ is a

per-new-form conditional, so the two thresholds are not directly comparable. Table 6 records every wave of the three campaigns.

Legal argument passes — switching from appellate to trial, contract, regulatory, and international venues leaves the rate flat (0.9), evidence of a shared argument-quality core that is venue-general. The known-frame simulation of this appendix shows how a biased sampler can overstate coverage and validates the estimators on a frame whose truth is known; it does not certify coverage of the open web. Academic review plateaus in place across five waves without a probe-sized reservoir of untried strata. Whether a domain's construct frontier closes is thus a substantive fact about how divergent its sub-communities' criteria are, and the re-heating probe is what distinguishes it from an artifact of where we happened to look.

Validation on a known-denominator frame. Because these devices are only as good as their theory, we validated all of them on a frame where truth is exactly computable: the AIRules corpus of 99,967 subreddits (467,554 rule tokens; 243,450 surface types), whose subscriber counts define exact sampler propensities (Lloyd et al., 2025). Simulating a search-like (popularity-biased) collector against a uniform one (200 replicates/cell, preregistered) shows: (i) *the deception gap* — the biased sampler's sampler-relative missing mass understates the population unseen mass by $+.14$ at n=1k, growing to $+.45$ at n=20k, so biased search *overstates* saturation and deeper sampling makes the overstatement worse (Fig. 13a); (ii) Good–Turing is accurate *at its own estimand* (within its concentration band in 100% of replicates) — the estimator is fine, the estimand is the sampler's; (iii) capture–recapture lower bounds are always valid but nearly uninformative when all lists share one bias (3–4% of true richness), while *bias-diverse* lists at identical budget triple the bound (Fig. 13b) — list diversity, not depth, is what buys population coverage; (iv) propensity reweighting (Horvitz–Thompson with propensities estimated from observables) recovers 85% of true richness where support is full but stalls when a stratum is unreachable — no weighting scheme corrects π=0; and (v) the reachability-floor bound fan contains truth exactly up to the true minimum propensity and breaks immediately above it (Fig. 13c). The deception gap is not a surface-string artifact: repeating the simulation at the *construct* grain (a Codex-seated partition of the frame's humor sub-frame, 4,082 forms into 2,439 constructs, anchor-gated) makes it *worse* — at n=2k the popularity sampler reports .04 missing mass, appearing fully saturated, while the true construct-level unseen mass is .60 (a $15\times$ overstatement vs. $2.8\times$ at the surface) — because constructs concentrate in large communities even more than surface forms do, so a search-like sampler is most flattered at exactly the grain we care about. Together these calibrate how to read our field numbers: sampler-relative rates are trustworthy as measured; population statements require the diversity devices above.

C Hierarchy construction

This section walks through every step of the clustering process sketched in §2, with the flow chart in Fig. 15. The guiding principle throughout: construct identity is a judgment about *normative content*, so every merge decision is an LLM judgment run under anchor gates, and every convenience heuristic (lexical dedup, retrieval shortlists) is used only to *propose* candidates, never to decide them.

Step 1: surface normalization. Extracted criterion names are normalized (case, punctuation, whitespace) and exact/near duplicates collapsed. Everything surviving is a surface form. No semantic judgment is made at this step.

Step 2: candidate shortlisting. For each surface form we retrieve candidate constructs by lexical overlap and token containment against existing construct head forms. This bounds the judgment budget; a form with no plausible candidates goes directly to Step 4.

Step 3: anchored same-construct judgment. An LLM judge decides, for each (form, candidate construct) pair, whether the form states the same normative requirement — the same thing a rater would check — with definitions and member examples shown. Batches carry camouflaged known-member positives and known-distinct negatives; a batch failing its anchor gate is discarded and re-run, never patched.

Step 4: adversarial minting. A form provisionally judged *new* does not mint a construct directly. A second, adversarial pass is shown the full inventory and instructed to find *any* existing construct that covers the form, defaulting to seating. Only

Campaign	Wave	Draws	New constructs	Rate / 100	95% CI	Stratum
Academic review	1	890	22	2.47	1.64–3.71	seeded angles
	2	999	9	0.90	0.47–1.70	narrowed angles
	3	849	6	0.71	0.32–1.53	knee (second wave below 2)
	4	607	4	0.66	0.26–1.68	fresh angles
	5	766	9	1.17	0.62–2.22	recent, AI, and method angles
Grants	1	524	4	0.76	0.30–1.95	federal and foundation
	2	507	5	0.99	0.42–2.29	knee called, later retracted
	3	437	6	1.37	0.63–2.96	same funder classes
	4	566	16	2.83	1.75–4.54	new funder classes (re-heating probe)
	5	427	6	1.41	0.65–3.03	old classes resampled (control)
Legal	1	375	4	1.07	0.42–2.71	appellate
	2	426	4	0.94	0.37–2.39	appellate
	3	477	9	1.89	1.00–3.55	appellate
	4	660	6	0.91	0.42–1.97	switch to trial, contract, regulatory, and international venues

Table 6: The three construct-level collection campaigns, wave by wave: draws, new constructs after adversarial re-seating, the rate per 100 draws, a Wilson 95% interval computed from the counts (draws treated as independent, so the intervals are lower bounds on width), and the source stratum sampled. The frozen rule stopped a campaign when the rate stayed below 2.0 per 100 for two consecutive waves, or after eight waves. The other eight domains have the census-time ladder statistic only and no campaign.

forms surviving this pass mint new constructs. On our field campaigns this pass re-seats the majority of provisional "new" verdicts — naive one-pass seating overstates construct novelty by roughly $2.3\times$.

Step 5: audited repair rounds. The partition is then stress-tested in both directions: random within-construct pairs are re-judged for over-merge, and nearest cross-construct pairs for under-merge; confirmed errors trigger merge/split repairs, and repair candidates are screened and confirmed by separate judges. Theme and category levels are built on top of the frozen construct level by the same judged process at coarser grain.

Grounding the levels. The levels are not ad hoc: they mirror the strata of established qualitative methodology — surface forms are verbatim data extracts; constructs correspond to *codes* (Braun and Clarke, 2006) and to psychometric constructs (Cronbach and Meehl, 1955); themes to Braun–Clarke *themes*; categories to the axial/selective categories of grounded theory (Corbin and Strauss, 1990). This correspondence is what makes external validation possible: independent literatures have produced human item-to-code and code-to-theme assignments (O*NET work activities, PDTB discourse relations, code-review feedback, the Scrum and UCSB codebooks, and DISAPERE review aspects; Table 7), and Appendix D measures whether our pipeline recovers them.

C.1 The algorithm

The hierarchy is not produced by clustering an embedding space. Construct identity is a judgment about normative content: two criteria can share almost every content word and impose different requirements (*cite sources accurately* vs. *cite sources completely*), and two criteria can share no vocabulary and impose the same one (*anagnorisi* vs. *get it moment*). Embeddings order the first pair above the second. The pipeline therefore uses lexical retrieval only to *propose* candidates and an anchored LLM judgment to *decide*, with an explicit repair stage because a single pass both over-merges and over-mints.

Algorithm 1 runs once per level with $\rho_{\text{construct}}$ (SAME CONSTRUCT), ρ_{theme} (SAME THEME), ρ_{category} (SAME CATEGORY); the form level is built the same way under SAME CRITERION. Each level is measured against *its own* relation, never the form level's, because precision at the form level and precision at the theme level are different questions.

D External validation against human-coded hierarchies

The hierarchy in §2 is built by LLM judgment, so it could encode one pipeline's priors rather than a structure a human coder would recognize. We test it against item-to-category assignments made by people, in literatures with no connection to ours.

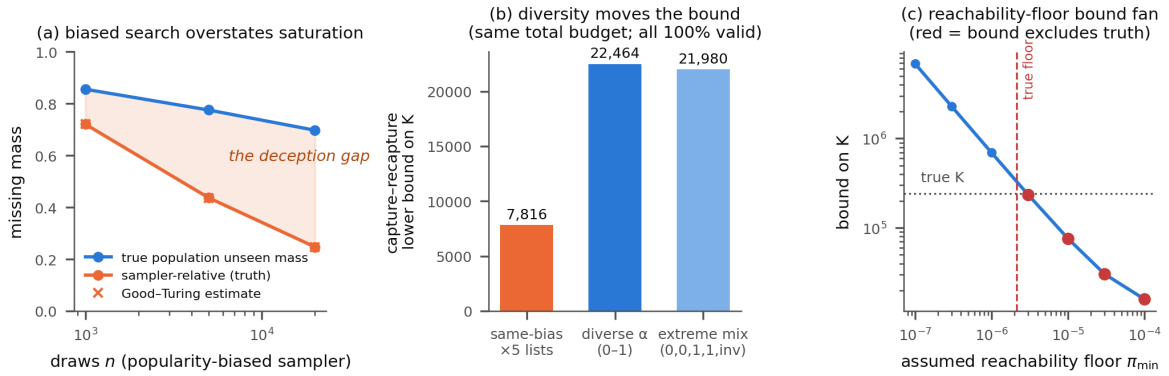


Figure 13: Sampling-theory validation on a known-denominator frame (AIRules; 99,967 subreddits, 243,450 rule types, subscriber-weighted propensities; 200 replicates/cell). (a) A popularity-biased (search-like) sampler's own missing mass (orange; Good–Turing tracks it exactly) falls far below the true population unseen mass (blue): the deception gap grows with depth — biased search overstates saturation. (b) Capture–recapture lower bounds at identical total budget: five same-bias lists are valid but weak; five bias-diverse lists nearly triple the bound while remaining 100% valid — diversity, not depth, moves the population bound. (c) The reachability-floor bound fan: assuming every type has propensity $\geq \pi_{\min}$ yields an upper bound on richness that provably contains truth for all $\pi_{\min}$ at or below the true floor (dashed red) and breaks immediately above it.

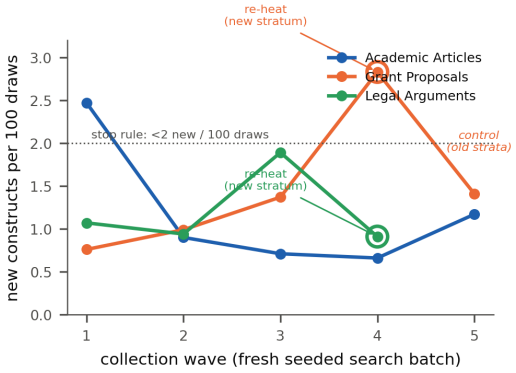


Figure 14: The re-heating probe (recent W15 campaigns). Each wave is a freshly seeded search batch; y is new constructs per 100 draws after adversarially hardened form-to-construct seating. Open circles mark *re-heating* waves that inject a new sub-population stratum. Grants jump on re-heat (new funder class) and fall back under the old-strata control — not saturated; legal stays flat across a venue-class switch — saturated and venue-general; academic review plateaus in place. A flat response to re-heating is the saturation certificate; a jump falsifies it.

D.1 What is available, and what is not

We searched for corpora carrying human-assigned hierarchical codes and found a consistent pattern: published qualitative work releases *either* the codebook (how codes nest) *or* the coded excerpts (which item went where), rarely both at depth. Table 7 lists the six corpora we obtained. Only two — O*NET and PDTB — supply three or more levels *and* item-level links. The hand-coded Reddit rule corpus of Fiesler et al. (2018) releases only its 24-category codebook, the coding behind Weld et al. (2022) is unreleased, and the Fora corpus (Schroeder et al., 2024) is access-gated under a no-redistribution license; we record these as unavailable rather than substituting weaker proxies.

D.2 Design

Each corpus supplies gold links at one or more *rungs*. A rung pairs one of our frozen level relations with one gold link: the form rung asks our SAME CRITERION prompt about two surface items whose gold link is co-membership in one code; the construct rung asks our SAME CONSTRUCT prompt about two code names whose gold link is co-membership in one theme; the theme rung does the same one level up. Prompt text is imported programmatically from the modules that built the corpus rather than retyped, so the arm measures the deployed instrument.

Negatives determine what the measurement means. We stratify them in equal parts: a **hard** negative draws its two items from different parents under the *same* grandparent (two O*NET tasks that both inspect logs for defects, under one generalized activity but different detailed activities); an **easy** negative draws from different grandparents. Random negatives alone push every cell near ceiling and separate nothing, so hard-negative AUC is the reported quantity; pooled AUC moves with

Corpus	Domain	Levels	Items	Links
O*NET (National Center for O*NET Development, 2025)	work activities	4	18,797	yes
PDTB 3.0 (Webber et al., 2019)	discourse relations	3	∼53k	yes
Code review (Turzo et al., 2023)	code review feedback	2	1,829	yes
Scrum (Alami and Krancher, 2022)	software process	2	131	yes
DISAPERE (Kennard et al., 2022)	peer review	1	734	yes
UCSB (Curty et al., 2021)	qualitative codebook	4	93 codes	no

Table 7: Human-coded hierarchies used for external validation. *Links* means the corpus records which surface item was assigned to which code, not merely how the codes nest.

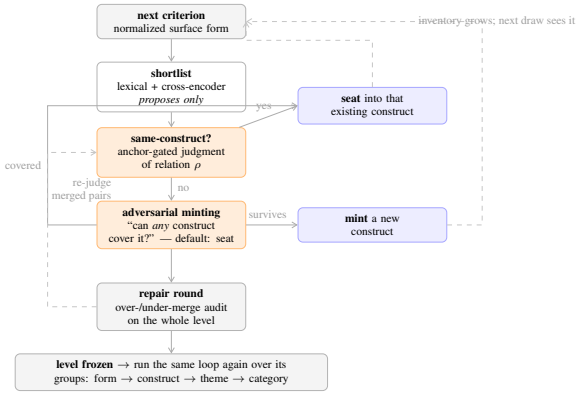


Figure 15: Hierarchy construction is an iterative seat-or-mint loop, not a one-pass clustering. Each criterion is shortlisted by heuristics that only *propose* (gray), then decided by anchor-gated LLM judgments (orange): if it matches an existing construct it is seated, otherwise it must survive an adversarial pass that defaults to seating before a new construct is minted (blue). Every seat or mint changes the inventory the next draw is judged against — the property that makes the process a Chinese Restaurant Franchise rather than a partition of a fixed set. Dashed arrows are feedback: periodic repair rounds re-judge the whole level, and once a level is frozen the same loop runs over its groups at the next grain up.

Algorithm 1 Building one level of the hierarchy from the level below. Lexical retrieval only *proposes* candidate groupings; every merge is decided by an anchor-gated LLM judgment of the level's own relation, and a group must survive an adversarial re-seat and a repair audit before it is named and passed up as input to the next level. Run three times, with ρ set to SAME CONSTRUCT, SAME THEME, SAME CATEGORY.

Require: nodes V_L with names+glosses; level relation ρ_{L+1}

1: **Retrieve.** For each v, take a shortlist $N(v)$ by TF–IDF and cross-encoder rank. ▷ proposal only; never decides
2: **Propose.** If $|V_L| \leq 220$, one prompt sees all nodes and returns a global grouping. Else bucket by TF–IDF neighbourhood; each node is assigned in its own self-centred bucket, so no group chains across buckets.
3: **Judge.** For each proposed pair, an LLM applies ρ_{L+1} on $\{0, 1, 2\}$ with six camouflaged known-answer anchors per batch. **Discard** the batch if < 5 anchors are correct.
4: **Seat.** Accept a merge only at score 2; ties resolve downward (*uncertain* ⇒ *not same*).
5: **Re-seat.** Re-present each accepted group with its strongest non-member as an adversarial distractor; drop members the judge no longer places inside.
6: **Repair.** Screen candidate under-merges among sibling groups, confirm each independently under the strict protocol, and apply only confirmed merges.
7: **Name.** Name each multi-member group from its members; singletons inherit their member's name. Names and reps become V_{L+1}, the input to the next level.
8: **Measure.** Against a frozen node-pair eval judged at ρ_{L+1}, report recall $= P(\text{co-clustered} \mid \text{same})$ and precision $= P(\text{same} \mid \text{co-clustered})$.

the hard/easy mix and is not quoted alone. A cell where easy far exceeds hard indicates a prompt reading topical distance rather than level membership. Each batch carries six camouflaged known-answer anchors and is discarded if fewer than five are correct. Readout is rank AUC of the $0/1/2$ prompt score against the gold binary, with bootstrap intervals.

One design fault is worth recording because it is easy to reproduce. Our first PDTB run returned AUC 1.000: we had rendered sense labels as their full dotted paths (*Comparison / Contrast*), while the construct gold link *is* a shared path prefix, so prefix comparison scores perfectly without reading the labels. Labels are now rendered leaf-only (*Result + Belief* vs. *Reason + Belief* is a positive pair; vs. *Result* is a hard negative) and the cell was re-judged. Any label-grain gold whose node names encode their own ancestry carries this hazard.

D.3 Results

Two results follow from Table 8.

Construct identity transfers across domains. Hard-negative AUC is .892 on discourse relations and .887 on U.S. labor-market work activities — two schemes built by different communities, for different purposes, with no shared vocabulary — and the on-domain code-review cell sits at .868

Corpus	Rung	n	AUC	95% CI	AUC_{hd}
PDTB	construct	122	.956	[.91,.99]	**.892**
O*NET	construct	140	.915	[.87,.96]	**.887**
Code rev.	construct	82	.868	[.80,.94]	—
UCSB	construct	140	.849	[.78,.91]	**.772**
Scrum	construct	103	.726	[.63,.82]	—
O*NET	form	140	.765	[.69,.83]	**.692**
Scrum	form	140	.723	[.65,.80]	**.644**
PDTB	form	140	.706	[.62,.78]	**.577**
DISAPERE	form	140	.552	[.50,.60]	—
Code rev.	form	135	.549	[.50,.60]	**.563**
PDTB	theme	126	1.000†	—	—
UCSB	theme	140	.736	[.62,.83]	—
O*NET	theme	133	.694	[.63,.76]	—

Table 8: Frozen level prompts against human gold links. Dashes mark cells whose gold parent is the top of its tree, so no grandparent exists to draw hard negatives from; those AUCs rest on easy negatives and are not comparable to the boldface column. †PDTB's theme rung discriminates 14 relation types across 4 semantically disjoint classes with no hard negatives available; the ceiling reflects an easy discrimination over a small label inventory and carries no information about the instrument.

on easy negatives. We had registered the opposite prediction, that agreement would be highest on the on-domain gold. It is not: at the construct level the on-domain and off-domain cells fall inside each other's intervals. The SAME CONSTRUCT relation our pipeline applies is not specific to evaluative criteria.

The claim is bounded by its fourth cell. Scrum's construct rung reaches only .726, against easy negatives, where the comparable code-review figure is .868. Scrum is the smallest corpus here (35 codes over 131 segments) and states its codes at a high level of abstraction (*Conformity to business needs*, *Code quality*), which leaves the code-to-theme link thinly determined by the code name alone. The agreement at .89 rests on the two corpora deep enough to admit hard negatives; extending it to every human coding scheme is not warranted.

Our surface level is finer than most published coding schemes cut. The form rung scores .56–.69 against hard negatives across four corpora where the two hard-negative construct cells score .89. The contrast holds *within* three of the four corpora that supply both rungs (code review .563 → .868; O*NET .692 → .887; PDTB .577 → .892), so for those it cannot be attributed to domain difficulty or text genre. The reason is

Corpus	Rung	prompt	tfidf	mpnet	bge
O*NET	construct	.887	.794	.911	.888
PDTB	construct	.892	.920	.918	**.944**
O*NET	form	.692	.604	**.717**	.693
Scrum	form	**.644**	.642	.612	.636
PDTB	form	.577	**.642**	.505	.506
Code review	form	.563	.518	.553	**.711**

Table 9: Hard-negative AUC, prompt vs. similarity baselines, same pairs.

visible in the data: two review comments both coded *Naming Convention* concern different identifiers in different files, and two O*NET tasks under one detailed work activity come from different occupations. Our form relation — would applying either item make the same pass/fail distinction? — declines those merges, and so scores near chance against a gold that treats them as one code. For these three, the finest rung cut falls between our form and construct levels.

Scrum is the exception, and it clarifies the mechanism. There the two rungs are level (.723 vs. .726 pooled; on matched easy negatives the form level is slightly *higher*, .779 vs. .726). Scrum's units are single-point interview quotes and its codes name one requirement apiece, so its bottom rung already sits near the grain our form relation targets and the rung effect disappears. The size of the gap elsewhere tracks how coarsely a corpus coded its bottom level, not a property of the judge.

D.4 Baselines

The arm-A numbers are only interpretable next to what a cheaper method scores on the same pairs. We ran TF–IDF cosine (the lexical control, since our protocols instruct the judge not to infer sameness from shared vocabulary) and three sentence encoders, on the identical judged pairs with the identical strata.

Table 9 does not show the prompts dominating. Sentence embeddings match or exceed them on four of six cells; BGE reaches .888 on O*NET constructs against the prompt's .887. Two asymmetries are worth stating. The prompt emits three ordered levels while cosine is continuous, so the prompt loses rank resolution to ties and is handicapped in AUC. And the ordering inverts on PDTB forms, where embeddings collapse to chance (.51, .51) while the prompt reaches .577 and TF–IDF .642: discourse-relation type is not similarity between the two spans, so an encoder measuring top-

ical closeness has nothing to read. Where the gold link is semantic, encoders are strong; where it is relational, they fail and the prompt does not.

The comparison that matches our actual claim is at the partition level, because the pipeline must produce clusters, not score pairs. Average-linkage agglomerative clustering on BGE embeddings, *given the gold number of clusters* — a parameter our pipeline is never told — scores adjusted Rand .005 on code-review codes, −.000 on DISAPERE aspects, .049 on Scrum codes, .203 on Scrum themes, and .118/.267 on O*NET tasks/activities. Our pipeline reaches .097 on DISAPERE and .08–.22 across the Scrum cells while inferring its own granularity. The two tasks therefore separate: at scoring one pair an off-the-shelf encoder is competitive with our judge, and at building a partition without being told how many clusters exist it is not.

D.5 Partition-level agreement

Where a corpus supplies two coded levels we also seat its items blind and compare partitions directly. On Scrum, an unprompted fine pass produced 39 clusters against 35 human codes (ratio 1.11) and a coarse pass 11 clusters against 14 human themes (ratio 0.79): the granularity our prompts settle on, with no target count supplied, matches the human one within roughly 20% at both levels. Partition agreement itself is weaker — adjusted Rand .097 on DISAPERE aspects and .08–.22 across the Scrum cells, all above 1,000-permutation nulls at $p \leq .002$, with homogeneity exceeding completeness throughout, the signature of a finer partition nested inside a coarser gold.

D.6 Limits

None of these corpora is double-coded, so the agreement two human teams would reach on the same pairs is unknown and every number here is a contrast — rung against rung, corpus against corpus — rather than an accuracy. Cells whose gold parent tops its tree admit no hard negatives. Precision and recall at the prompt's own decision point are uninformative in most cells (recall .02–.05), because the protocols resolve ambiguity downward by construction; the rank statistic carries the measurement.

E Behavioral dataset construction

This section details the text corpora and silver labels of §2. For each of six domains we collect the community's commentary on individual works (Table 10). From each comment we parse stated norm applications — a criterion invoked about a specific piece — with the same verbatim-evidence discipline as the census extractor, then seat each application onto its construct with the anchored seating protocol of Appendix C. Aggregation per item yields the silver labels used in §4: 1.01M parsed norm applications, of which 64,820 seat onto a construct in a batch whose anchors pass, touching 665 distinct constructs (Table 10).

Checks on the mapping. Four checks discipline the mapping. (1) Every matching batch includes camouflaged known-answer anchors (application→construct pairs with hand-verified or consensus-adjudicated labels, plus known-wrong pairs that the judge must abstain on); batches failing them are discarded and re-run, and passing batches score ≥.96 on both anchor types. (2) Because the same judgment machinery builds the hierarchy and matches comments into it, any grouping-level claim risks circularity from clustering freedom. A random-grouping null, reshuffling scored constructs into 500 random families matched on the observed size distribution, showed family-level aggregation effects indistinguishable from the null in four of five audited domains; those claims were retracted, and every number in this paper is computed at the leaf level, where the matcher has no grouping freedom. (3) A judge from a different family than the matcher (GPT-5.6) read 50 seated matches per domain (300): the comment invokes the assigned construct in .77 [.72, .82] and at least partly in .92 [.88, .95]; by domain, yes / yes-or-partly: software code .78 / .86, creative writing .76 / .98, humor .86 / .90, math .80 / .94, academic articles .70 / .92, press releases .74 / .92. The humor matcher was trained partly on GPT-5.6 labels, so that domain's check is not fully independent. (4) Because the labels are weakly supervised, the linking can be biased toward the LLM's own dialect (Juzek and Ward, 2025; Liang et al., 2024): human phrasings closer to it would be matched more often, accumulate more labels, and favor LLM-preferred wording in the validity comparison. The register null of §4 runs against this: the formal, LLM-preferred phrasings show no validity advantage (44 wins against 44, and 45 against 39 in PREREG-15-v2), the opposite of what a dialect-matching bias would produce. The residual con-

Domain	Corpus	Kind	Parsed	Seated	Constr.
Humor	Reddit joke & standup	critique threads	77,378	8,010	117
Creative Writing	r/WritingPrompts	feedback comments	4,929	2,877	106
Software Code	GitHub PR + CodeReview.SE	review comments	611,785	19,997	132
Math	Math StackExchange	answer comments	8,238	3,133	88
Academic Art.	ICLR/NeurIPS OpenReview	peer reviews	277,420	19,049	86
Press Releases	release→news pairs	news coverage	26,996	11,754	136
Total			1,006,746	64,820	665

Table 10: Commentary corpora behind the silver labels. *Parsed* = norm applications extracted from human commentary; *Seated* = applications matched to a construct in anchor-passing batches; *Constr.* = distinct constructs receiving at least one seated application, at the grain of the scoring partition (coarser than the construct partition of Table 1; humor has 285 against 1,282). Two exclusions keep the commentary human-written: creative writing omits 329K LLM-written craft rationales that were extracted alongside the Reddit feedback, and math omits contest editorials, which are author self-narration rather than third-party feedback. The software-code seated count is from the frozen correlation table; a later rerun gives 18,044 and the two were never reconciled.

cern that the shared judgment distorts hierarchy and labels in the same direction at the leaf level is bounded by the external validation of the construct relation against six human-coded corpora (Appendix D) and by this bias direction. Because the labels were drawn from comments, positive statements may also be overrepresented; we read them as an indicator of norm transmission in the community rather than as ground truth.

F Instruments

All measurements in this paper are made by LLM judges under a fixed discipline: every instrument has (i) a frozen definition and prompt, (ii) deterministic known-truth *anchors* camouflaged in every judging chunk with a hard acceptance gate (failed chunks are discarded, never patched), (iii) validation against external labeled gold where such gold exists, and (iv) a cross-family agreement check on our own items. Preregistered contrasts are frozen — direction, single test, and null — before any confirmatory run.

Register instruments. *Formality* (1–7 register elevation) and *etymological stratum* (Germanic/Latinate/Greek/mixed) follow the lexical-bar and register traditions (Corson, 1985; Biber, 1988); the deployed judge is validated against human lexical-formality ratings (Pavlick and Nenkova, 2015; Brooke et al., 2010) (ρ=.85) and the Etymological Wordnet (de Melo, 2014) (86–90%). *Nominalization* (derived abstract nominals, the signature of institutional prose) validates against NOMLEX (Macleod et al., 1998) (97.9% under our suffix-restricted definition).

Metaphoricity instruments. *Metaphoricity* (figurative borrowing vs. literal naming) is a term-level analog of MIP/MIPVU (Pragglejaz Group, 2007; Steen et al., 2010; Lakoff and Johnson, 1980); the deployed judge scores 88.7% on type-level idiom gold drawn from MAGPIE (Haagsma et al., 2020), with additional metaphor resources (Mohammad et al., 2016; Birke and Sarkar, 2006; Gao et al., 2018) reserved for context-level extensions. *Semantic transparency* (compositional vs. opaque) validates against human compositionality ratings (Gagné et al., 2019; Reddy et al., 2011; Venkatapathy and Joshi, 2005; McCarthy et al., 2003).

Specificity instrument. Criterion *specificity* (general / specific / hyper-specific) operationalizes the breadth of the evaluand class over which a criterion's verdict can vary; its functional reading is the rules-vs-standards trade-off (Ehrlich and Posner, 1974; Kaplow, 1992), its cognitive reading basic-level category structure (Rosch et al., 1976), and its linguistic reading sentence specificity (Louis and Nenkova, 2011; Li and Nenkova, 2015; Ko et al., 2019), against whose labeled data the instrument is benchmarked, with WordNet depth and information content (Resnik, 1995) as mechanical corroborators.

F.1 Instrument reliability

This section reports, per instrument, accuracy of the deployed judge against external labeled gold and chance-corrected cross-family agreement on our own items — the two numbers that together say whether an instrument measures its construct and whether that measurement is model-idiosyncratic. Table 11 is the sum-

mary grid (machine-readable: `instrument_calibration_20260723.json`).

Metaphoricity on corpus names. The idiom gold validates type-level detection, not the judgment of a criterion name in context. Human annotators labeled a stable-hash sample of corpus names (60 lay-critic and 40 community-rule names drawn so that 35 are figurative by the judge), each shown with its source quote, as figurative, literal, or undecidable, with the judge's label hidden until the name was answered. On the 37 annotated names, agreement with the judge is .76 and κ=.49; the annotators call 16 figurative against the judge's 11. Agreement is .90 on lay-critic names and .56 on community-rule names, where the disagreements are terse rule phrasings the annotators read as figurative and the judge as literal (*very low effort posts*, *key information*, *barely edited*).

G Experiments on LLM judges: details

This section gives the design details and full per-family numbers behind the rearticulation, name-swap, and form-effect experiments of §4.

Three properties are kept apart throughout: *consistency* (whether a judge's item rankings agree across runs or across names), *agreement* with the community's own judgments (AUC against the silver labels), and *validity* as a measure of quality, which neither establishes, since a judge can reproduce biased community judgments consistently.

G.1 Rearticulation

Design. Every construct with at least five attested human namings is included; each family is sampled as many times per construct as there are human namings, so the emitted-name distribution is directly comparable to the human one at matched N. Emitted names are scored by the same register instruments as the corpus. Human name probabilities are occurrence counts of extracted head terms (near-duplicate quotes collapsed), not counts of distinct names; in the original run each sample saw a different source record's definition of the construct.

Rearticulation robustness. Forty constructs (ten each from humor, creative writing, math, and journalism; 37 after three rewrites failed an equivalence check) were renamed by GPT-5.6 and Claude Sonnet 5, 325 samples per variant per family with every raw sample kept, under four definition variants: the original per-record definitions (ORIGP), one fixed original definition per construct (ORIGC), a plain-language rewrite (PLAIN), and a rewrite that keeps about 9% of the original's content words (NEUTRAL); rewrites were written by GPT-5.6 and checked for meaning by Claude. Table 12 reports the raw modal-name share, per-construct entropy, and judged formality (anchor gates passed in 18 of 18 chunks). Under the original per-record design the contrasts with human names are directional but not significant at this size (p=.10–.21); under any fixed definition the collapse is large ($p<2\times10^{-6}$), and PLAIN and NEUTRAL match ORIGC, so the definition's wording does not drive it. A fixed definition measures sampling variability on one input while a human inventory spans many documents, so the two designs are not like for like. Compared within the four domains that have institutional names (institutional mean 4.54), the families' emitted names are level with or slightly below that mean (GPT-5.6 $-.06$ to $+.02$; Claude $-.19$ to $-.13$; GLM-4.7 in the original run $-.14$) and human names for the same constructs are .43 below it; every family is above the community mean in each domain.

Entropy comparison. Raw emitted-name samples were retained for one family (GLM-4.7); the other four retained only aggregate head-share statistics, which is why the body reports the head-share contrast for all five families (.11 human vs. .19 LLM) and the entropy contrast for one. Over the 149 constructs with at least five human namings in both sources, mean per-construct Shannon entropy is 2.66 bits for the human attested distribution against 2.51 for the model (Wilcoxon p=.0008; the model is less diverse on 71 constructs, more diverse on 44, tied on 34). The difference concentrates in creative writing ($+.29$ bits); humor, math, and journalism are near zero ($+.05$, $+.09$, $+.05$). Descriptive, not preregistered.

G.2 Name swap: design and per-family results

Name swapping. The judge scores a fixed item set under a scoring prompt; across arms only the criterion's name changes, drawn from that construct's attested register variants: pairs come from constructs with both a Germanic and a Latinate judged variant in the register bank and from lay-

Instrument	Description	Ext. gold	Judge	Cross-family
formality	elevation of the name as a register choice, 1 (casual everyday speech) – 4 (neutral professional) – 7 (technical-institutional)	ρ=.85 (human)	Sonnet	ρ .84–.87
etymology	dominant root stock of the content words: Germanic (*help, craft*) / Latinate (*clarity, rigor*) / Greek (*metaphor*) / mixed	86–90% (EtymWN)	Sonnet	κ=.71
nominalization	whether the name contains a derived abstract nominal (*-tion, -ity, -ness, -ment, -ance*) — the signature of institutional prose	97.9% (NOMLEX)	Sonnet	κ=.77
metaphoricity	whether the name is a figurative borrowing from another domain (*punch, flow, hook*) vs. literal (*citation accuracy*)	88.7% (MAGPIE)	GLM	κ .58–.70
transparency	whether the meaning is composable from the parts (*excessive setup*) vs. opaque/idiomatic (*kill your darlings*)	AUC .83 (LADEC)	GLM	κ .62–.71
polarity	whether the phrasing names a fault or prohibition (*no toilet humor*) vs. a positive quality (*clarity*) vs. a neutral topic	F1 .82 (LexDeMod)	Codex	—
specificity	breadth of the evaluand class: general / document-particular / surface-bound (rules-vs-standards scope)	ρ=.30 (adoption breadth)†	Codex	κ=.24 (weak)

Table 11: Reliability of the register instruments: accuracy of the deployed judge against external labeled gold, and chance-corrected cross-family agreement (Sonnet/GPT-5.6/GLM) on our own items (n=300/instrument). †Specificity validates against *empirical scope* — a rule's adoption breadth across 99,969 subreddits (ρ=.30, p=7×10^{-11}; rules in $\geq$100 communities: 55% general / 7% hyper-specific; singletons: 23% / 22%); the gradient survives length control essentially unchanged when computed separately among short ($\leq$4-word) and among long (5–12-word) names (ρ=+.21/+.20, n=1,071, both $p<10^{-5}$) — while its nulls against sentence-specificity and privacy-vagueness gold are *discriminant* validity: normative scope is not descriptive detail. Its remaining weakness is cross-family boundary placement (κ=.24), so per-item bucket calls carry a caveat; distributional contrasts are supported. Polarity's external gold is prohibition detection on lease-clause deontic labels.

Family	Definition	Modal share	Entropy	Formality
Human	attested names	.264	2.59	4.11
GPT-5.6	ORIGP	.295	2.47	4.51
	ORIGC	.711	0.93	4.47
	PLAIN	.689	1.02	4.50
	NEUTRAL	.700	0.98	4.55
Claude	ORIGP	.312	2.43	4.34
	ORIGC	.687	1.07	4.41
	PLAIN	.667	1.19	4.42
	NEUTRAL	.667	1.21	4.37

Table 12: Rearticulation under four definition variants (37 constructs, two families): raw modal-name share, mean per-construct entropy in bits, and mean judged formality of emitted names, against the human attested names for the same constructs.

versus-professional matched heads, with a judged formality difference of at least one SD. About 20 pairs were planned and 14 were run (seven humor, seven creative writing); the rule that selected these 14 was not recorded, and meaning equivalence and score direction were not checked for them (PREREG-25 below). Within-name reliability is a split-half statistic: the Spearman correlation between the mean item ranking of two runs and that of the other three, averaged over the ten splits of the five runs; between-name agreement is the Spearman correlation between the five-run mean rankings under the two names; Δ is their difference, the disruption attributable to the name alone. The control arm paraphrases the scoring instructions sentence by sentence (name, scale, and output format unchanged; one paraphrase; GPT-5.6 only) while keeping the name fixed. Table 13 gives the per-family values and Table 14 the per-pair values. The Δ contrasts were preregistered (direction and test frozen before the confirmatory run on GPT-5.6; the other four families were added afterwards under the same frozen design and are reported separately); the paraphrase control disrupts nothing (Δ=$-.004$ against $+.505$ for the name swap, p=1.2×10^{-4}).

Specification. Items: 30 per domain, identical across both names and all seven pairs of the domain; humor items are one-line jokes, and creative-writing items come from a corpus disjoint from the silver-label corpus. Each run scores every item 1–7; items with other values are dropped. Judges: GPT-5.6 (`gpt-5.6-sol` through the

Family	Within	Between	Δ	p
Gemma-4	.938	.410	$+.525$	2.4×10^{-4}
GPT-5.6	.913	.408	$+.505$	6.1×10^{-5}
Claude	.758	.261	$+.497$	4.3×10^{-4}
GLM-4.7	.669	.451	$+.218$	8.5×10^{-4}
Llama-8B	.500	.510	$-.010$	.64

Table 13: Name-swap effects per judge family: within-name run-to-run reliability, between-name agreement, and their difference Δ. Under swapped names every family lands in a similar range (.26–.51); what varies is how much within-name reliability there was to lose. Llama-8B has the lowest within-name agreement (.50; chance is 0), so a swap has less agreement to remove. Gemma-4's between-name value is a mean over the 12 of 14 constructs where it is defined.

Codex CLI, default sampling, five fresh sessions per name), Claude (Sonnet; the model identifier was not recorded; five fresh contexts), GLM-4.7 (Z.ai API, temperature 1.0, one call per item, about 75% of items parsed), and Gemma-4-31B and Llama-3.1-8B (vLLM, seed 1000 plus the run index). Ties take average ranks. A correlation is undefined when fewer than 10 items overlap or a ranking is constant; this removes Gemma-4's pairs 1 and 13, where one arm collapsed to a near-constant score, from its Wilcoxon test (n=12) but not from its permutation count or level shift. The register-asymmetry range p=.22–.90 covers four families, since Gemma-4's is undefined, and 6.1×10^{-5} and 2.4×10^{-4} are the smallest one-sided values attainable at n=14 and 12. The raw runs, item files, and prompt texts of PREREG-11 and PREREG-12 were lost when a working directory was wiped; the per-construct results survive in the stored result files, and PREREG-25 tests the effect on persisted data.

Length and register distance. Across the 85 register arms of PREREG-15-v2, the name-swap disruption is unrelated to the difference in name length (Spearman ρ=.09, p=.43), to total name length ($\rho=-.01$), and to the formality difference between the two names (ρ=.05, p=.66), so neither length nor register distance accounts for it. Familiarity and specificity were not measured for these forms.

G.3 Matched-meaning pairs and human annotation

PREREG-25: matched-meaning name swap. Frozen 2026-10-12 before any adjudication or scoring. Materials: the 86 humor constructs of PREREG-15-v2 with their 40-item sets and all 858 phrasings. Adjudication: GPT-5.6 and Claude Sonnet 5, blind to every score and AUC and shown names only, judged every within-construct pair with a formality difference of at least one SD (1,567 pairs) for equivalence (same requirement / overlapping / different) and every form (799) for score direction (higher is better / worse / unclear); each batch carried camouflaged anchors (paraphrase pairs, cross-metric pairs, explicit fault names) and passed at 10 of 10 in all 54 batches. Agreement: κ=.62 on three-way equivalence and .67 on same-versus-not; .59 on three-way direction and .73 on worse-versus-not. Both families call 113 of 1,567 pairs the same requirement and 17 of 799 forms fault-oriented. Selection: per construct, the eligible pair (same requirement, both forms higher is better) with the largest formality difference; 35 constructs qualify. Scoring: each name under the frozen scaffold twice and the less formal name under two paraphrased scaffolds once each, in GPT-5.6 and Claude Sonnet 5; the statistics and the primary test are those of the name swap above, with a TOST equivalence margin of .10 for the paraphrase control. Results (Table 15): the primary test is supported in both families, the paraphrase control is equivalent to zero within .10 for GPT-5.6 only, and the AUC differences between the two names are small (median .03 and .07; 3 and 7 of 35 constructs differ after Benjamini–Hochberg correction) with no register direction (formal wins 14 and 19 of 35). Excluding the two constructs scored first under a superseded pair selection leaves the primary at p=.0009 and .004. All 420 runs are persisted with their prompts.

Human annotation. Two human annotators jointly annotated 97 pairs blindly, on a page that shows the two names, the judges' question, and the three answers, with the judge verdicts hidden until a pair is answered: the 10 published pairs (6 same, 3 overlapping, 1 different), the 14 name-swap pairs (10 same, 4 overlapping), 38 of the 86 register arms (23, 13, 2), and all 35 PREREG-25 pairs (35 same). Against each judge family on the 83 pairs both saw, three-way κ=.22 (GPT-5.6) and .21 (Claude), same-versus-not κ=.39 and .39; against the two families jointly, same-versus-not κ=.37, agreement .66. Nearly every disagreement is a pair the annotators call the same and the judges call overlapping (e.g. *Punch up, not down*

Pair	Domain	Low name / high name	W	B_n	B_s	Δ_n	Δ_s	shift$_n$	shift$_s$
0	humor	playing to the room / aptum	.873	.685	.740	.188	.133	1.03	0.49
1	humor	trim the fat / pleonasm	.825	.216	.872	.609	−.047	3.92	0.27
2	humor	wordplay / paronomastic	.964	.934	.979	.029	−.016	0.35	0.12
3	humor	long walk / cumulative humor	.817	−.321	.698	1.139	.119	2.15	0.20
4	humor	button / logical mechanism	.935	.915	.946	.020	−.011	1.37	0.71
5	humor	callback / reincorporation	.867	.759	.938	.108	−.071	2.64	0.15
6	humor	blow / presentation	.931	−.320	.965	1.251	−.034	2.41	0.02
7	creative writing	start close to the end / in media res	.898	.447	.949	.451	−.051	1.10	0.10
8	creative writing	show don't tell / specificity at moment of action	.958	.680	.963	.279	−.005	0.37	0.08
9	creative writing	keep your cast small / dramatis personae	.954	−.077	.983	1.031	−.028	1.31	0.28
10	creative writing	reaching the reader / profluence	.938	.913	.896	.025	.041	0.41	0.03
11	creative writing	struggle / progressive complication	.962	.302	.969	.660	−.007	1.25	0.23
12	creative writing	bad guy close in / escalation	.943	.534	.982	.409	−.039	1.45	0.03
13	creative writing	world of hats / originality	.921	.046	.964	.875	−.043	2.46	0.17
Mean			.913	.408		.505	−.004	1.59	0.21

Table 14: The 14 name-swap pairs (PREREG-11/12, GPT-5.6). W: within-name split-half reliability; B_n: between-name agreement; B_s: agreement between the original and paraphrased scoring instructions under the same name; Δ_n=$W-B_n$, Δ_s=$W-B_s$; shift: mean absolute score change (points) under the name swap and under the paraphrase. The median Δ_n is .43; six pairs exceed .5 and five are below .2, so the mean of .505 mixes near-synonymous and non-synonymous pairs. Three negative B_n values (pairs 3, 6, 9) are what a pair naming different requirements, or one with reversed score direction, would produce.

	GPT-5.6	Claude
within-name agreement	.889	.763
between-name agreement	.769	.660
between-scaffold agreement	.851	.683
Δ_{name} / Δ_{scaffold}	.120 / .043	.104 / .045
constructs with $\Delta_{\text{name}}>\Delta_{\text{scaffold}}$	26 / 35	25 / 35
primary Wilcoxon p	.0004	.0015
TOST, Δ_{scaffold} within $\pm.10$	p=.002	p=.13
$\lvert\Delta\text{AUC}\rvert$ median (IQR)	.03 (.01–.10)	.07 (.03–.12)
$\lvert\Delta\text{AUC}\rvert$ max	.20	.44
CIs excluding 0, raw / BH	9 / 3	10 / 7
formal / informal wins	14 / 21	19 / 16

Table 15: PREREG-25 results on 35 verified-equivalent name pairs (humor), per judge family. Agreement values are Spearman correlations of item rankings; the paraphrase control uses two independently verified rewordings of the scoring instructions.

and *Tonal discipline: target systems, not individuals*; *No toilet humor* and *Sophisticated humour and absence of offensive material*). Name-swap disruption on the register arms by the annotators' label: same .60 (n=23), overlapping .71 (n=13), different .66 (n=2); the 14 name-swap pairs by the annotators' label: same .51 (n=10), overlapping .49 (n=4). The paper adopts the construct-level reading as its definition of the same criterion, and reports the judges' requirement-level reading alongside it as the stricter standard. Under the construct-level reading the disruption does not depend on the label, which is what that reading predicts if a construct's names differ in requirement below the construct grain. Table 23 lists every annotated pair.

Judges calibrated to the construct-level reading. We re-ran both judge families over all 1,567 pairs and the 14 name-swap pairs with the same question, a calibration paragraph stating the construct-level reading, and 32 of the annotators' labeled pairs as examples (3 different, 7 overlapping, 22 same); the remaining 65 labeled pairs were held out. All batches passed their anchors. On the held-out pairs the calibrated judges agree with the annotators at κ=.49 (GPT-5.6) and .41 (Claude) three-way and .66 and .64 on same-versus-not; on the 21 held-out pairs outside PREREG-25 and the name-swap set, where the two readings differ most, κ=.38 and .36 on same-versus-not against .07 and .13 for the strict judges. Under the calibrated standard both families call 380 of 1,567 pairs the same construct (113 under the strict standard, all retained): 24 of the 86 register arms (10 outside the examples), 5 of the 10 published pairs, 8 of the 14 name-swap pairs, and all 35 PREREG-25 pairs, so the PREREG-25 result is unchanged. Among the register arms the disruption is Δ=.54 for calibrated-same pairs (n=24; .44 on the 10 outside the examples) against .58 for overlapping and .76 for different, and the same-versus-not difference is not significant under any standard (Mann–Whitney p=.15–.54). Exploratory; the examples' labels are the annotators' by construction.

G.4 Replication across domains

The name swap across domains. Table 16 collects the name-swap replications: for each do-

main, the number of constructs with a verified pair under the construct-level standard, the disruption from the name swap and from the paraphrase control in each judge family, and the one-sided test. The design is that of PREREG-26 throughout (up to 20 or 25 constructs per domain by stable hash, six runs per construct per family); humor's construct-level run (PREREG-30) uses the PREREG-15-v2 item sets, and its requirement-level run is PREREG-25.

PREREG-30: humor under the construct-level standard. Frozen before selection and scoring. From the 86 humor constructs with item sets, those with a pair both families call the same construct under the calibrated prompt, with both names higher-is-better, excluding pairs used as examples in that prompt (68 qualify), capped at 20 by stable hash; formality differences run 1.5 to 3.5. Three selected pairs coincide with PREREG-25 pairs and reuse those runs. Claude Sonnet 5 (n=19; one construct dropped after two failed runs): Δ_{name}=.13 against Δ_{scaffold}=.01, p=.002, 14 of 19, paraphrase equivalent to zero within .10. GPT-5.6 (n=19 of 20): .30 against .06, p=1.9×10^{-6}, 19 of 19, equivalent. The formality difference does not predict the disruption (ρ= − .30 and −.20, both p>.2). The GPT-5.6 runs used the laptop companion, in which the session reads the item file itself; the difference from the PREREG-25 harness is recorded with the runs.

PREREG-26: the name swap in creative writing. Frozen 2026-10-12 before any judging, with one logged amendment (the construct-level standard as primary) recorded before any adjudication reply was read. Materials: the 284 phrasings of 26 creative-writing constructs and a pool of 251 WritingPrompts stories. Formality: every phrasing judged twice under the register protocol in 14 anchor-gated chunks (all passed; the two passes agree at ρ=.84). Adjudication: all 1,466 within-construct pairs and 284 forms, under both the construct-level and the requirement-level prompt, by GPT-5.6 and Claude Sonnet 5 (100 batches, all passed; κ=.62 and .61 on equivalence, .66 and .71 on direction); both families call 322 pairs the same construct and 119 the same requirement. Selection: per construct, the pair both families call the same with both names higher-is-better and the largest formality difference; 26 constructs under the construct-level standard, 21 under the requirement-level one, with 5 in common. Scoring: 40 stories per construct by stable hash, the frozen scaffold and two paraphrases checked for meaning by GPT-5.6, six runs per construct per family (264 runs per family, all valid on the first attempt). One construct (a resolved ending) is dropped because the stored stories are 400-character openings and every run scored it constant. Results: construct-level selection, GPT-5.6 Δ_{name}=.29 (median .14) against Δ_{scaffold}= − .01, p=2.6×10^{-6}, 22 of 25; Claude .33 (.22) against .01, p=3.2×10^{-5}, 21 of 25; the paraphrase effect is equivalent to zero within .10 in both (TOST p<10^{-3}). Requirement-level selection, GPT-5.6 .12 against .00 (p=.001, n=20); Claude .19 against .03 (p=.005, n=19). The formality difference between the two names does not predict the disruption (ρ= − .04 and −.20). All runs, prompts, and adjudication replies are persisted.

PREREG-27: math and journalism. The same design, frozen per domain before judging, with up to 20 constructs by stable hash and the construct-level selection scored. Math (Math StackExchange answers; 24 constructs, 803 pairs; 216 pairs the same construct by both families, κ=.55; 120 runs per family): GPT-5.6 Δ_{name}=.36 against Δ_{scaffold}=.01, p=1.9×10^{-5}, n=18, paraphrase equivalent to zero within .10; Claude .25 against .06, p=.003, n=16, equivalence not shown (p=.19). Journalism (homepage headlines with context; 24 constructs, 775 pairs; 133 same by both, κ=.59; 96 runs per family): GPT-5.6 .16 against .10, p=.48, n=14; Claude .15 against −.01, p=.005, n=13. Items are headlines only, so criteria about article bodies are hard to apply, and criteria that cannot apply to a text (programming examples or graphics in math) scored constant and were dropped, which is why n is below 20. The formality difference between the two names does not predict the disruption in either domain.

PREREG-28: software code, press releases, academic articles. One note frozen for the three domains before any judging; 20 constructs per domain by stable hash with up to eight names each; items are the judged objects (the first 1,000 characters of GitHub pull-request diffs, press-release texts, ICLR 2024 titles with abstracts), 200 per domain by stable hash. All adjudication batches and formality chunks passed their gates; both paraphrases were judged equivalent to the scaffold in every domain. Pairs the same construct by both

Domain	Standard	n	GPT-5.6 Δ_{name}	GPT-5.6 Δ_{scaf}	GPT-5.6 p	Claude Sonnet 5 Δ_{name}	Claude Sonnet 5 Δ_{scaf}	Claude Sonnet 5 p
Humor	requirement (PREREG-25)	35	.12	.04	.0004	.10	.05	.0015
Humor	construct (PREREG-30)	19 / 19	.30	.06	1.9×10^{-6}	.13	.01	.002
Creative Writing	construct (PREREG-26)	25	.29	$-.01$	2.6×10^{-6}	.33	.01	3.2×10^{-5}
Creative Writing	requirement	20 / 19	.12	.00	.001	.19	.03	.005
Math	construct (PREREG-27)	18 / 16	.36	.01	1.9×10^{-5}	.25	.06	.003
Journalism	construct (PREREG-27)	14 / 13	.16	.10	.48	.15	$-.01$	.005
Software Code	construct (PREREG-28)	16 / 15	.25	.01	.005	.17	.06	.042
Press Releases	construct (PREREG-28)	17 / 16	.28	.01	7.6×10^{-6}	.23	.04	.003
Academic Articles	construct (PREREG-28)	15 / 13	.06	$-.05$	.015	.17	.05	.029
Legal Arguments	construct (PREREG-29)	8 / 6	.34	$-.01$	.008	.26	$-.03$	.078
Patents	construct (PREREG-29)	5 / 4	.68	.00	.094	.62	.06	–
Grant Proposals	construct (PREREG-29)	14 / 11	.26	.08	.045	.17	.03	.062
Regulatory Comments	construct (PREREG-29)	6 / 5	.22	.03	.016	.23	$-.04$	.031

Table 16: The name swap by domain. n: constructs with a verified pair (two values where the families differ); Δ_{name}: within-name minus between-name Spearman agreement; Δ_{scaf}: within-name minus between-scaffold agreement under the paraphrase control; p: one-sided Wilcoxon for $\Delta_{\text{name}}>\Delta_{\text{scaf}}$. In the last four domains many criteria fit none of the items and score constant, so n is small; – marks a test not run ($n<5$).

families: code 95 of 470 (κ=.60), press 113 of 467 (.68), academic 129 of 414 (.68); constructs with an eligible pair: 17, 17, and 15. Results (GPT-5.6; Claude): code Δ_{name}=.25 against Δ_{scaffold}=.01, p=.005, n=16; .17 against .06, p=.042, n=15. Press .28 against .01, $p=7.6\times10^{-6}$, n=17; .23 against .04, p=.003, n=16. Academic .06 against $-.05$, p=.015, n=15; .17 against .05, p=.029, n=13. The paraphrase effect is equivalent to zero within .10 for GPT-5.6 in all three domains and for Claude in press releases only. No formality-gap correlation is significant. Caveats: constructs whose criteria cannot apply to a diff opening or an abstract scored constant and were dropped, and the academic bank mixes criteria for papers with criteria for review writing.

PREREG-29: legal arguments, patents, grant proposals, regulatory comments. One note frozen for the four domains before any judging; up to 20 constructs per domain by stable hash with up to six names each (names cut off at the 90-character storage limit, 30 to 40% of names, are excluded); items are 200 texts per domain by stable hash: statements of the facts in Title VII cases (first 1,000 characters), patent abstracts, NIH FY2024 project abstracts, and public comments on proposed rules (400 to 1,000 characters). All formality chunks and adjudication batches passed their gates, and both paraphrases were judged equivalent to the scaffold in every domain. Pairs the same construct by both families: legal 27 of 141 (κ=.64), patents 20 of 132 (.42), grants 27 of 171 (.35), regulatory comments 27 of 182 (.57); the same requirement: 4, 7, 11, and 5. Constructs with an eligible pair: 11, 13, 16, and 8. Criteria that fit none of the items (an indictment is not evidence, applied to statements of fact) score constant in every run and are dropped, so the test keeps 4 to 14 constructs per family. Results (GPT-5.6; Claude): legal Δ_{name}=.34 against Δ_{scaffold}= $-$ .01, p=.008, n=8; .26 against $-.03$, p=.078, n=6. Patents .68 against .00, p=.094, n=5; .62 against .06, n=4, no test. Grants .26 against .08, p=.045, n=14; .17 against .03, p=.062, n=11. Regulatory comments .22 against .03, p=.016, n=6; .23 against $-.04$, p=.031, n=5. The paraphrase effect is equivalent to zero within .10 for GPT-5.6 in legal and patents only. The formality difference between the two names does not predict the disruption in any domain. Most selected pairs differ by at most one point of judged formality, and the names are extracted rule statements rather than named criteria, so these four domains test the construct-level claim with the least power.

G.5 Form effects and register matching

This section expands the form-effect result of §4. Table 17 lists the largest confidence-interval-clean effects in each direction when the same construct is scored under its most informal versus most formal phrasing: single phrasing choices move agreement with community silver labels by up to .77 AUC (.73 once the three fault-oriented forms are reversed), in both directions; neither register nor, once score direction is corrected, prohibition fram-

ing predicts the winning direction.

PREREG-15-v2: register matching with fresh arms. Because the item-level data of PREREG-15 were lost, we froze a fresh design (2026-08-28) on the same construct set: 858 phrasings judged for formality twice under the register protocol (38 of 38 anchor batches passed), each construct's least and most formal phrasing as arms (85 constructs after the 1-SD difference rule; one construct had no item set), 20 positive and 20 negative community-labeled documents per construct drawn by stable hash, and GPT-5.6 scoring every item twice per arm (340 runs). The frozen primary test asks whether the arm closer to the humor community's own register (A=2.75) is the more valid one: Spearman $\rho(\Delta d, \Delta\mathrm{AUC})$=.01, one-sided p=.55; with the lay-critic (3.08) or pooled community (3.20) anchor, ρ= − .02, p=.43. Secondaries: informal wins 45, formal 39, one tie (p=.59); $|\Delta\mathrm{AUC}|$ median .11 (IQR .03–.21, max .78); 31 of 85 per-construct bootstrap intervals exclude zero, 23 after Benjamini–Hochberg; mean AUC .71 (informal) and .69 (formal), so the design is not at floor. Exploratory: reversing the three fault-oriented arm forms leaves ρ= − .02 (p=.42) and the win split at 47 to 37; the polarity contrast on the three constructs with one fault-oriented arm moves from $+.51$ to $-.24$. Only 4 of the 85 arm pairs are verified equivalent, so this experiment, like PREREG-15, measures the choice among a construct's phrasings. An even win split rules out a strong directional tendency but is not evidence that the two arms are equivalent.

Register and judge consistency. Within-name reliability is higher under the institutional-register arm than the vernacular arm in all five families (GPT-5.6 .935 vs. .892; Gemma-4 .957 vs. .919; Claude .767 vs. .750; GLM-4.7 .671 vs. .666; Llama-8B .501 vs. .499), but no family's asymmetry is significant on its own (p=.22–.90) and the cross-family sign pattern is only suggestive (5/5 families, exact p=.0625). Descriptive.

Register distance and effect size. Across the 88 pairs, the register difference between a construct's two phrasings does not predict which arm wins (signed ΔAUC: ρ=.08, p=.45) but weakly predicts how much is at stake ($|\Delta\mathrm{AUC}|$: ρ=.22, p=.042). Mean AUC is essentially identical across arms (.675 informal, .676 formal). Descriptive.

H Linguistic associations: details

H.1 Register by speaker class and domain

This section gives the per-domain register measurements behind the pooled contrast in §5. Table 18 reports mean judged formality of criterion names by rule-writer class in each domain, including the two reversal domains, which we show rather than pool away.

Metaphoricity by function at matched register. The dissociation claim of §5 is that adjudicated speech sheds metaphor at any register, i.e. that community rules' low metaphoricity is not a byproduct of their low register. Each name carries a single judged formality score on the integer 1–7 scale, so the matching is exact: we compare community-rule and lay-critic names at the same score, pooling only the sparse tails (1 with 2, and 6 with 7) for sample size, which yields five strata (n=998 community-rule and 1,000 lay-critic names with both judgments). Community rules are less figurative at four of the five formality levels (each Fisher p<.02; the exception is the sparse top level, n=75), with Mantel–Haenszel OR=.41 across strata (χ^2=49.3, p=2.2×10^{-12}) against a crude unstratified OR=.54. Controlling register strengthens the contrast rather than explaining it away. Official-class terms are excluded here: they have term-level formality judgments but no term-level metaphoricity judgments. Descriptive, not preregistered.

Domain composition, matched constructs, and authorship. Community rules are plainer than lay critics in 7 of 10 domains, and the 95% cluster-bootstrap interval (resampling communities and source pages) excludes zero in software code (-0.36), humor (-0.36), and math (-0.46); press releases reverses ($+0.37$ [0.06, 0.67]). Averaging within-domain differences gives -0.14 [-0.23, -0.05], against -0.39 [-0.46, -0.32] pooled, because humor alone contributes 1,202 of the 3,332 community rules. Institutional names are more formal than both other classes in every domain where they occur (1.32 [1.04, 1.58] points above community rules and 1.06 [0.80, 1.31] above lay critics, averaged over humor, math, and journalism), and 0.86 [0.17, 1.50] points above lay names for the same construct (20 constructs named by both; community rules are not seated on constructs, so they cannot be matched). Authorship differences are small: organization-authored in-

	Construct (informal form ↔ formal form)	AUC_{inf}	AUC_{for}	Δ
informal wins	does the joke land at the target audience ↔ use of crude/sexual/scatological material	.84	.34	$+.50$
	punch up, not down ↔ presence of the target (working rubric)	.92	.42	$+.50$
	keep it simple; don't overthink ↔ specificity and lyricism in writing	.94	.45	$+.49$
	storytelling with long setups and twist endings ↔ exclusion — pure congruity	.57	.10	$+.47$
	strong button (sketch) ↔ credits interrupted by stingers	.95	.50	$+.45$
formal wins	hack / hacky (anti-pattern) ↔ originality (surprise effect)	.10	.87	$-.77$
	Rule 12: be brutal ↔ economy of language	.24	.96	$-.73$
	use of humor / comedic tone ↔ humor / comedic effectiveness	.51	1.00	$-.49$
	you don't have to be funny all the time ↔ humor quotient	.36	.76	$-.40$
	failure mode: R2 too hard to find ↔ object is a perceived incongruity	.27	.61	$-.35$

Table 17: Largest CI-clean form effects against community silver labels (humor; $\Delta = AUC_{inf} - AUC_{for}$). CI-clean: the construct's 95% bootstrap interval for Δ (resampling items) excludes zero, without multiplicity correction. Both directions occur (18 vs 20 of 88); register does not predict the winner. Equivalence verdicts (GPT-5.6 / Claude) for these ten pairs: no pair is judged the same checkable requirement by both families, while two human annotators judge 6 of the 10 to name the same construct (a2, a32, a50, a64, a118, a142) (a2 same/overlap; a64 overlap/same; five pairs overlap/overlap; three pairs different by at least one family). Forms both families judge fault-oriented: *hack / hacky* (a64), *pure congruity* (a142), *R2 too hard to find* (a86); reversing their AUCs gives $\Delta = +.03$, $-.33$, and $+.12$ and the non-prohibition advantage falls from $+.18$ ($p = .0065$) to $+.05$ ($p = .19$), so the prohibition contrast is confounded with score direction. AUC cells are exact values from `paper1_figdata.json`.

Domain	Comm. rules	Lay critics	Instit. guides
Software Code	3.36	3.68	–
Creative Writing	3.14	3.30	4.73
Grant Proposals	3.36	3.70	–
Humor	2.75	3.08	3.83
Legal Arguments	3.60	3.55	–
Math	3.23	3.70	5.42
Journalism	3.62	3.57	4.17
Regulatory Comments	3.64	3.79	–
Patents	–	3.71	–
Academic Articles	3.84	3.91	–
Press Releases	3.42	3.06	–

Table 18: Mean judged formality of criterion names by rule-writer class per domain (9,292 anchor-gated judgments). Community rules are plainer than lay criticism in 7 of the 10 domains with both classes; legal arguments, journalism, and press releases reverse (heterogeneity is real; the pooled contrast is LOO-robust, $p = .0038$). Cells are exact values from `paper1_figdata.json`; '–' = $n < 10$ judged official-class terms in that domain (official names concentrate in a few domains; the pooled official mean is 4.24). Pooled contrast: community < lay, Fisher $p = .0038$.

stitutional names average 4.29 ($n = 62$) and those by credentialed individual experts 4.57 ($n = 21$); credentialed lay critics average 3.81 ($n = 224$) and other lay critics 3.57 ($n = 3{,}982$). The per-domain institutional counts behind Table 18 (12, 35, 24, and 11 names for math, journalism, humor, and creative writing in the figure data) could not be reproduced exactly from the files on disk; a modal-domain assignment gives 10, 35, 28, and 9, with means within 0.15 of the table.

Metaphoricity by class. Every community-rule and lay-critic name carries a name-level metaphoricity judgment: 12% of community-rule names and 21% of lay-critic names are figurative. Institutional names have bank-level judgments only: .14 for the 29 strongly institutional variants plotted in Fig. 7 and .35 over all 109 scored institutional names, so the institutional class is not placed on the metaphoricity axis on equal terms.

H.2 Specificity by speaker class

This section breaks the specificity distribution of Fig. 2 down by speaker class, showing the enforcement gradient discussed in §5: community rules, which must be adjudicated, are several times more hyper-specific than individual lay speech.

Diffuseness and generality. Each parsed source document carries a micro-topic tag (e.g. `kirkus_indie_review_submission`), which we treat as the subfield of the criteria on that page. Distinct subfields per 1,000 criteria (from 35 in code review to 64 in humor) correlate with the share of general norms at $\rho = +.82$ ($p = .002$, $n = 11$; Fig. 16), but this normalization does not make domains comparable: nearly every source page has its own tag, so the measure is the inverse of criteria per page ($\rho = -.82$ with that quantity). With equal budgets the correlation is $-.13$ (rarefaction to 1,000 criteria per domain, 1,000 draws) and $-.04$ to $-.12$ under coverage standardization (common coverage .5 to .9); the

Speaker class	n	general	specific	hyper-sp.
Institutional/professional (census)	53,413	48.9%	44.9%	6.2%
Lay critics	4,787	33.0%	62.0%	5.0%
Community rules	3,174	30.0%	52.6%	**17.4%**

Table 19: Norm specificity by speaker class (census buckets; lay-critic and community-rule buckets anchor-gated, 50/52 chunk gates). Community rules — which must be *enforced* — are 3.5× more hyper-specific than individual lay speech, the rules-vs-standards signature (Ehrlich and Posner, 1974; Kaplow, 1992) and the same adjudication-function axis that makes community rules literal at lay register. The census row includes 4,950 creative-writing forms that were never labeled and default to general (without them: 43.7% / 49.5% / 6.8%, n=48,463), and census labels do not replicate when a 300-item sample is relabeled by GPT-5.6 and Claude (both call about 4% general; κ=.03–.08 with the deployed labels, .70 with each other). The community and lay rows do replicate: on 1,000 names per class the community-to-lay hyper-specific ratio is 3.0× on the deployed labels, 3.6× [2.7, 5.0] under GPT-5.6, and 3.7× [2.8, 5.2] under Claude; on those names the two families agree with the deployed labels at κ=.57 (GPT-5.6) and .49 (Claude).

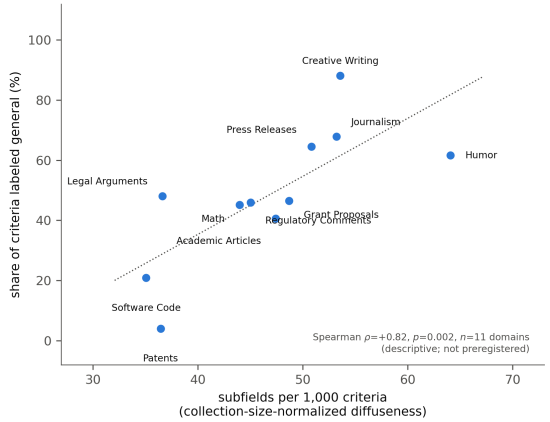


Figure 16: Sub-community diffuseness against norm generality. x: distinct subfields per 1,000 criteria; y: share of the domain's criteria judged *general*. Spearman ρ=+.82, p=.002, n=11 domains. Descriptive. The association does not survive equal-budget rarefaction (ρ=−.13); see text.

raw count gives −.18, and excluding creative writing changes none of these signs (+.87 normalized, −.16 rarefied). Rarefied richness at 1,000 criteria spans only 708 (code review) to 835 (notice and comment). We therefore report no association between diffuseness and generality. Descriptive, not preregistered.

H.3 Codability of the naming process

This section quantifies how conventionalized each domain's naming process is: when a fresh named criterion is drawn, how often it introduces a new concept, renames a known concept, or reuses a known concept's established name. Reuse is rare everywhere — most naming events introduce a phrasing not already attested — which is the mechanism behind the open surface tail of Table 1.

Does variation track community diversity? The domains that rename most are creative writing (.50), journalism (.48), and humor (.46). Against per-domain diversity measures built from source metadata (distinct community subreddits, distinct source sites, and the Shannon entropy of the source-site distribution), the renaming rate correlates positively with all three (ρ=+.36 to +.59; strongest against distinct source sites, p=.055): directionally, more diverse speaker pools relexicalize more, though nothing reaches significance at n=11. A second variation measure, the forms-per-construct ratio, ranks academic articles, press releases, and math highest and correlates negatively with the same diversity measures (ρ=−.41 to −.50); the two measures answer different questions, since the renaming rate isolates the naming process while the ratio also reflects sampling depth. Descriptive, not preregistered. Because distinct-site counts grow with collection size, we also compare at equal budgets: with 961 documents drawn from each domain (the smallest domain's count; 500 draws), the mean number of distinct source sites correlates with the renaming rate at ρ=+.57 (p=.069), against +.61 (p=.048) on the raw counts; the entropy of the site distribution gives +.28 (p=.41). The association does not depend on collection size, but eleven domains leave it suggestive.

H.4 Codification level across domains

Domains differ in *who states* their criteria, not only in how those criteria are worded. We code every source document in the census onto a five-rung ladder of institutional codification — official guideline, professional standard, academic, practitioner, community folk — and report the mix per

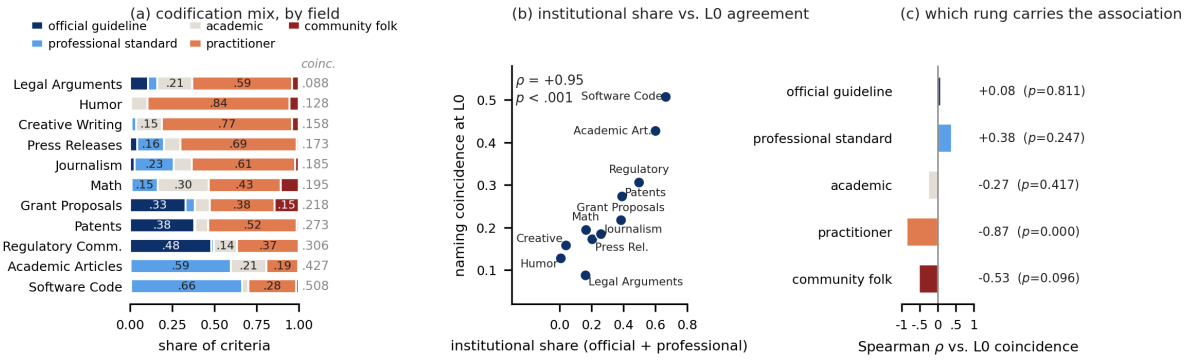


Figure 17: Codification level across the eleven domains, and its relation to naming agreement. **(a)** Share of each domain's criteria stated by a source document at each rung of the codification ladder, criterion-weighted (a criterion carries the rung of the document it was extracted from), with domains ordered by naming coincidence (right-hand column). Blue rungs are the institutional side of the validated binary split. **(b)** Institutional share against coincidence: the two move together across domains. **(c)** Each rung's share correlated separately with coincidence. No institutional rung carries the association on its own; *practitioner* share is what runs against it, so the composite in (b) tracks coincidence largely by being its complement. Sensitivity: weighting by document rather than by criterion gives ρ=+.72, p=.013. These correlations are descriptive and were not preregistered; see the caveats below.

domain in Fig. 17a. The unit is the criterion rather than the page, matching Fig. 2: each extracted criterion carries the rung of the document it came from, so a domain whose criteria concentrate in a few long checklists is scored by where its criteria actually come from.

How the ladder was coded. The unit is the source document, and every rung is judged from the document's own text. We first tested the shortcut of mapping our scrape-time page-type tags onto the ladder and rejected it: on a validation set of 311 documents stratified over all eleven domains and page types, blind-coded by two Sonnet judges with camouflaged anchors (6/6), tag-vs-judge agreement was only .60 exact (.69 at three classes), because page-type tags encode *genre* rather than *authority* (a journal's reviewer guidelines and a federal regulation both look like "formal guideline"). The deployed coder (GLM-4.7, prompt identical to the validation judges) reads a 1,600-character excerpt and assigns one rung under frozen definitions: (1) *codified/official*, issued by a governing authority and binding (statutes, agency guidelines, patent-office manuals, formal contest rules); (2) *professional standard*, authoritative within a profession but not binding (stylebooks, professional-body standards, journal reviewer guidelines); (3) *academic* (research articles, textbooks, course pages); (4) *practitioner/secondary* (how-to guides, tutorials, trade press, wikis); (5) *folk/informal* (personal blogs, forum threads, informal community advice). Judges are instructed to code from textual content and voice, not URL fragments. Calibrated against the 311-document reference, the coder's three-class agreement is .777, below our .85 instrument bar, while the binary institutional split ($\{1,2\}$ against $\{3,4,5\}$) agrees at .869; this is why panel (a) marks that boundary, every confirmatory use collapses to the binary split, and the five-rung mix is shown descriptively. Criteria inherit their document's rung through an exact key join (100% joined in all eleven domains).

The spread is wide. Humor draws essentially nothing from official guidelines (.00) and .84 of its criteria from individual practitioners; regulatory comments draw .48 from official guidelines and academic articles .59 from professional standards. The top of the ordering also splits into two institutional forms rather than one. Grant proposals, patents, and regulatory comments are *statutory*: their criteria come from official guidelines — funder program rules, the patent examination manual, rulemaking guidance. Academic articles and software code are *professional-standard* domains: reviewer forms, style guides, and linter rule documentation. Both are institutional, through different organs.

Codification level tracks naming agreement across domains (Fig. 17b, ρ=+.95). But the per-rung breakdown in Fig. 17c shows the associa-

Domain	new concept	renames concept	reuses name	reuse |known
Academic Articles	.49	.34	.17	.33
Creative Writing	.39	.50	.11	.18
Grant Proposals	.48	.41	.11	.21
Humor	.45	.46	.09	.17
Journalism	.38	.48	.14	.22
Legal Arguments	.48	.41	.11	.21
Math	.54	.37	.09	.19
Patents	.41	.34	.25	.42
Press Releases	.43	.38	.19	.33
Regulatory Comm.	.51	.32	.17	.34
Software Code	.51	.18	.32	.64

Table 20: Codability of the naming process: when a fresh named criterion is drawn, the probability it introduces a new concept, renames a known concept, or reuses a known concept's known name. Reuse of established names is rare everywhere (.09–.32) — most naming events introduce a phrasing not already attested — with Software Code the most conventionalized lexicon and Humor/Math the least. The first three columns are unconditional event probabilities and sum to one; the last is the probability that a restatement of an already-known construct reuses its established name, reuses/(renames + reuses) (exact binomial 95% intervals within $\pm.06$ for every domain). Recomputed without the LLM-written source documents of Appendix A.3, the domain ordering of the last column is unchanged (Spearman .96): humor .17→.12, creative writing .18→.16, math .19→.17, software code .64→.65, while patents (.42→.53), press releases (.33→.40), and regulatory comments (.34→.64) rise on 40–60% fewer records.

tion does not come from institutional authority: official-guideline share is uncorrelated with coincidence (ρ=+.08), professional-standard share only weakly so (ρ=+.38), and it is practitioner share that runs against it (ρ=−.87). What suppresses shared naming is a domain whose criteria are written by individual practitioners working independently. Legal arguments and mathematics make the point directly: near-identical institutional shares (.16 both) but coincidence of .088 and .195, because mathematical writing has shared proof-writing conventions and legal argument quality has no comparable common instrument. Software code is the converse, with almost no official-guideline content and the highest coincidence in the census.

Caveats. Four limits bound what Fig. 17 supports. (i) The correlations are *descriptive*: they were computed after the domain ordering was known, on n=11 domains that are not independent draws, and no directional test was preregistered. The preregistered tests in this line are *within*-domain — whether author pairs sharing a codification class name the same thing alike more than cross-class pairs — and they separate by grain: null at the surface-form grain, supported at the construct grain. (ii) The five-rung code is descriptive only; every confirmatory use collapses to the binary split, as above. (iii) Criterion weighting favours dense documents, and professional-standard sources are the densest (7.2 criteria per document against 2.9–4.2 elsewhere) — though official-guideline sources are the *sparsest*, so the weighting does not uniformly favour institutional provenance. The document-weighted correlation is weaker but same-signed (ρ=+.72). (iv) Coincidence and institutional share can both rise from boilerplate repeated across documents; our near-duplicate guard operates on quoted evidence spans and would not catch paraphrased reuse across sources in the same class.

H.5 Norm age per domain

Figure 18 shows three domains and Figure 19 the remaining eight. For each domain we show the themes with the oldest attested sources alongside themes appearing only in recent sources; ancient anchors are genuine, and recent-only rows carry a small-n caveat discussed in the caption.

Table 21 gives the per-domain attestation profile. The ordering is systematic: practices that needed no institution are attested from antiquity, practices born of early-modern institutions start in the 17th–18th centuries, and practices younger than 1900 have no earlier themes to show.

Construct age within domains. Domain age (above) and construct age are different questions, so we also compare themes *within* each domain: each theme's earliest attested source year against the properties of its names. Coverage: 3,308 of 3,816 themes have an attestation year; 2,163 of those have at least five specificity-labeled leaves. Specificity is the one property that moves robustly with age: older themes are more general (pooled within-domain rank correlation ρ=+.22, $p<10^{-21}$; the direction holds in nine of ten domains and is individually significant in six). Part of this comes from source counts: a theme with more dated sources tends to have an earlier earliest date (ρ=−.55) and more general names (ρ=+.20). Controlling for the number of dated sources and the theme's breadth (its number of

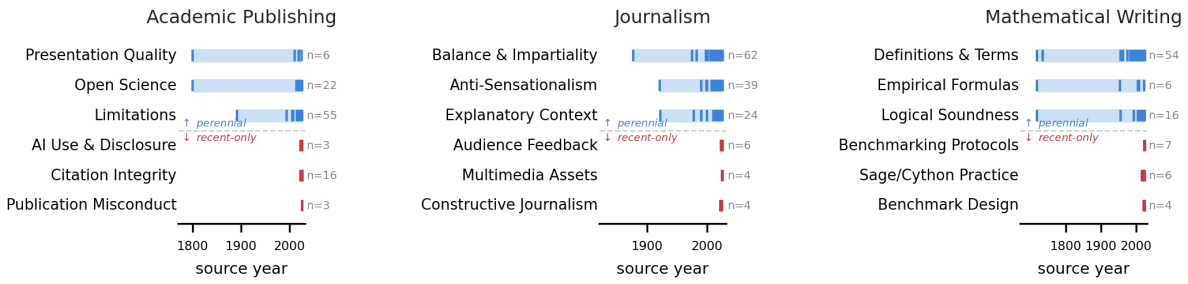


Figure 18: Across eleven areas of normatively evaluated creative output we collect explicit evaluative criteria published online, and date each theme by its earliest dated source found in our search. Three domains are shown: academic publishing, journalism, and mathematical writing. In each, some themes are anchored in sources centuries old while others appear only in recently dated ones — the emergence and lexicalization of norms this paper studies. Bars run from a theme's earliest to its latest attested source; n is the number of dated sources supporting it.

Domain	Earliest	Median	Pre-1900
Legal Arguments	∼350 BCE	2018	.19
Creative Writing	∼335 BCE	2019	.22
Humor	95 CE	2019	.05
Regulatory Comm.	1628	2018	.05
Math	1718	2015	.08
Press Releases	1790	2021	.01
Patents	1793	2020	.05
Academic Articles	1800	2021	.04
Journalism	1877	2018	.02
Software Code	1948	2019	.00
Grant Proposals	1962	2020	.00

Table 21: Attestation profile per domain: earliest dated source year, median source year, and share of themes with at least one pre-1900 attestation (theme = aspect stem; misdated Gutenberg pages filtered). Every median is contemporary; what varies is how deep the oldest themes reach, and that depth tracks the age of the practice itself.

constructs), the correlation is $+.11$ [.06, .16], and it is positive in every tertile of source count ($+.09$, $+.16$, $+.26$). Metaphoricity shows a weak, directionally consistent tilt toward older themes being more figurative (ρ=.09, p=.005, n=919; no single domain significant alone; with the same controls $+.05$ [$-.02$, .12], indistinguishable from zero), and formality and hyper-specificity show no age relationship. One method caveat: the formality and metaphoricity joins are approximate (scored terms matched to theme leaves by lexical overlap, since the two corpora share no key), so those two rows are weaker evidence than the specificity row, which joins exactly. Descriptive, not preregistered.

What a date means. A theme's date is the composition date of the earliest source our search found that states it. This is neither the earliest attestation of the norm in history nor the date of its English wording: the oldest source in the census, Aristotle's *Rhetoric* (c. 350 BCE, about 2,375 years ago), is read in a modern English translation, so it dates the requirement, not the phrasing. A few parser-assigned dates are also wrong (e.g. a Project Gutenberg biography dated 1877 among the journalism sources).

I Released artifacts

The release (code and experiment materials: `https://github.com/alex2awesome/one-construct-many-names`; dataset: `https://huggingface.co/datasets/alex2awesome/CreativePreferences`) contains: (1) the 53,413 census criteria, each with the criterion's name, description, application guidance, verbatim evidence quote, source URL, page type, community, audience, and date; (2) the hierarchy, as form→construct→theme→category membership files per domain; (3) the register, metaphoricity, specificity, and polarity judgments for every judged name; (4) the commentary corpora as norm applications (comment span, matched construct, document id) with usernames removed, and the silver labels per construct; (5) the materials of every behavioral experiment whose runs were persisted (prompts, item sets, raw judge outputs; PREREG-15-v2 and PREREG-25 in full); and (6) the collection, clustering, and analysis code. Judges are GPT-5.6 (`gpt-5.6-sol` through the Codex CLI), Claude (Claude Sonnet 5 through Claude Code), Gemma-4-31B and Llama-3.1-8B (served locally with vLLM), and GLM-4.7 (through the Z.ai API).



Figure 19: Perennial vs. recent-only norms for the remaining eight domains (same construction as Figure 18). Ancient anchors are real: rhetorical strategy in legal argument traces to ∼350 BCE; humor's audience-targeting to 95 CE. Recent-only rows should be read with the small-n caveat above: some are credible recent formalizations (SEO in press releases, crowd-work reporting in comedy), while others reflect corpus recency rather than lexicon novelty.

J Preregistration records

Each preregistration was written to the project registry, which is version-controlled, before any data for its test existed; Table 22 lists those behind results in this paper. An analysis is *confirmatory* when its hypothesis, statistic, and test were frozen before the run; everything else in the paper is labeled descriptive or exploratory where it appears.

K Hierarchy subtrees by domain

For each of the 11 domains, Fig. 20–30 below show a subtree of the certified hierarchy (Appendix C), in a wide layout of the kind used for the humor example in Fig. 1. For each domain we pick the evaluative category with the most member themes and show its two richest themes (by number of member constructs); a dashed edge to an unboxed "· · ·" marks that the category has further member themes not drawn here. Each shown theme in turn displays up to two of its richest constructs (ranked by number of named surface forms) whenever the theme has at least two named constructs that each carry at least two named forms; where that support does not exist, the theme instead shows only its single richest construct, with as many named forms as the data provides (2–3 where available). Nine of the eleven domains support this two-construct branching in both shown themes (code-review, creative-writing, grant-funding, humor, legal-outcome-prediction, math-stackexchange, news-homepages, notice-and-comment, patents); peer-review and press-releases branch in only

ID	Question	Frozen	Status
3	LLM names collapse onto a modal term	2026-07-22	confirmatory
10 / 10R	register by source class	2026-07-23	confirmatory; 10R a disclosed rerun
11	name swap lowers rank agreement	2026-07-23	confirmatory
12	prompt paraphrase control (GPT-5.6)	2026-07-23	confirmatory
13	metaphoricity by source class	2026-07-23	confirmatory
15	form effects on silver labels	2026-07-23	confirmatory (two-sided)
17	polarity explains form effects	2026-07-23	exploratory (hypothesis formed after inspecting PREREG-15)
20	stopping rule on a known frame	2026-07-24	confirmatory
21	extraction validity	2026-07-24	confirmatory
22	hierarchy vs. human-coded corpora	2026-07-24	confirmatory
23	frozen level prompts vs. gold	2026-07-25	confirmatory
15-v2	register matching (fresh arms)	2026-08-28	confirmatory
25	matched-meaning name swap	2026-10-12	confirmatory
26	name swap in creative writing	2026-10-12	confirmatory (one logged amendment before judging)
27	name swap in math and journalism	2026-10-12	confirmatory
28	name swap in software code, press releases, academic articles	2026-10-12	confirmatory
29	name swap in legal arguments, patents, grants, regulatory comments	2026-10-12	confirmatory
30	name swap in humor under the construct-level standard	2026-10-12	confirmatory

Table 22: Preregistrations behind results in this paper: registry ID, the question tested, freeze date, and status.

one of their two shown themes, because the other theme's construct partition is here fully degenerate (every construct a singleton surface form) — a genuine property of those cells, not a display choice, and visible directly in those two panels as a theme with only one child construct. All category, theme, and construct names are taken verbatim from the frozen taxonomies produced by the classification pipeline of Appendix C (`outputs/lexicon/derive_then_classify_v1/<task>/{R2,R3}/taxonomy_*.json`); none are generated for this figure. Leaf shading gives judged formality, darker = more formal, freshly scored per exact surface-form text (not joined from the register bank, which measures only at the coarser construct/head-variant level) by the same 1–7 instrument as the register bank (§F; full prompt and per-form scores released at `outputs/lexicon/prereg23_l0_formality.json`), so every leaf in every panel is shaded. The short italic tag under a form, where present, is its dominant etymological stratum or metaphoricity, joined from the register bank (§F) where it covers a given surface form; elsewhere it is omitted rather than inferred. The machine-readable data behind every panel, including the full (untruncated) strings, the per-theme branch-support audit, and the member-theme audit trail for each chosen category, is released at `outputs/lexicon/prereg23_task_subtrees_v3.json` and `outputs/lexicon/prereg23_r3_names_v3.json`.

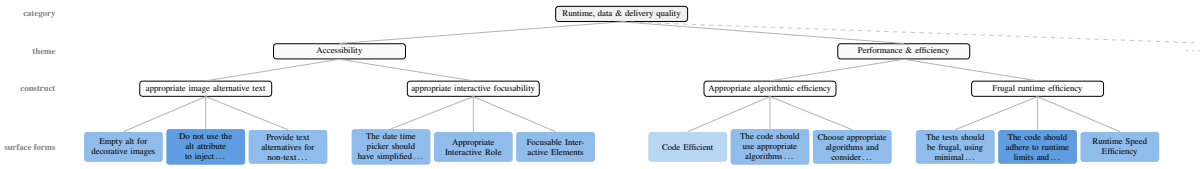


Figure 20: Software Code, rooted at the category *Runtime, data & delivery quality* (2 of its 8 themes shown).

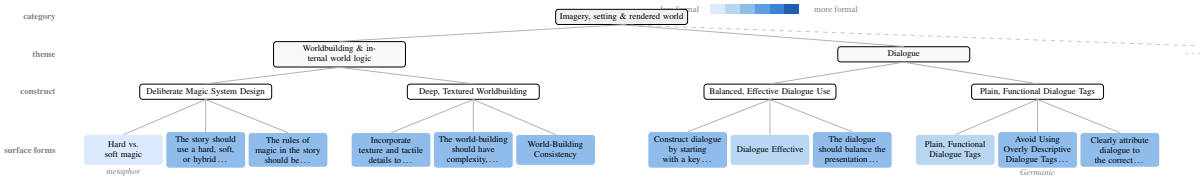


Figure 21: Creative Writing, rooted at the category *Imagery, setting & rendered world* (2 of its 8 themes shown).

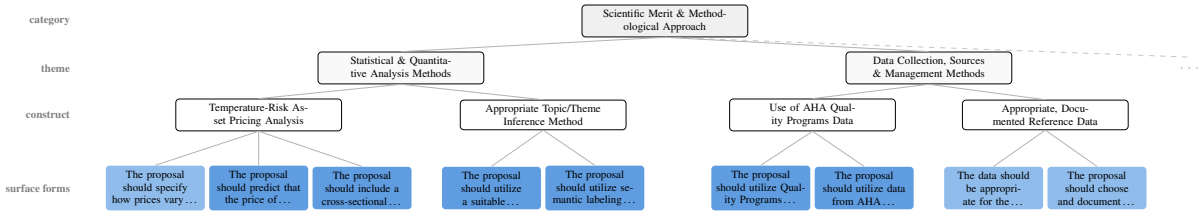


Figure 22: Grant Proposals, rooted at the category *Scientific Merit & Methodological Approach* (2 of its 11 themes shown).

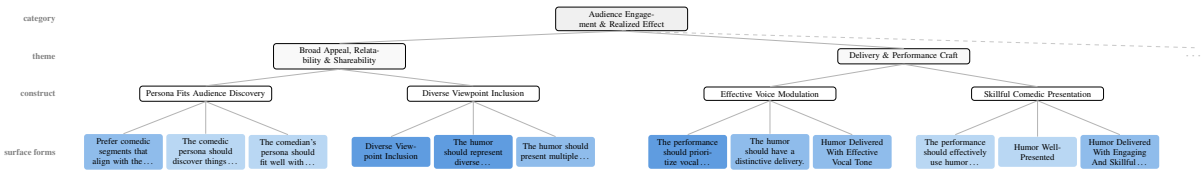


Figure 23: Humor, rooted at the category *Audience Engagement & Realized Effect* (2 of its 8 themes shown).

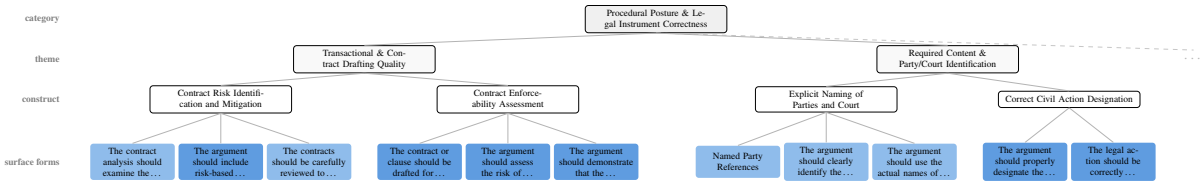


Figure 24: Legal Arguments, rooted at the category *Procedural Posture & Legal Instrument Correctness* (2 of its 13 themes shown).

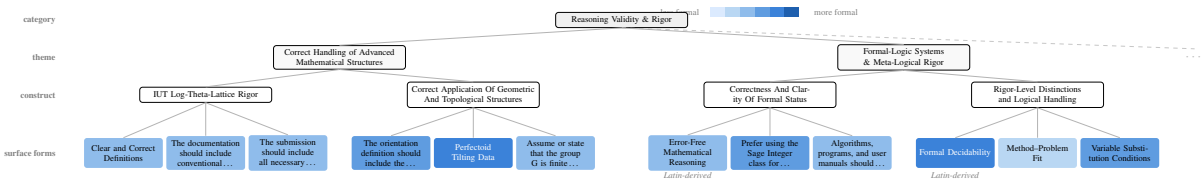


Figure 25: Mathematical Writing, rooted at the category *Reasoning Validity & Rigor* (2 of its 11 themes shown).

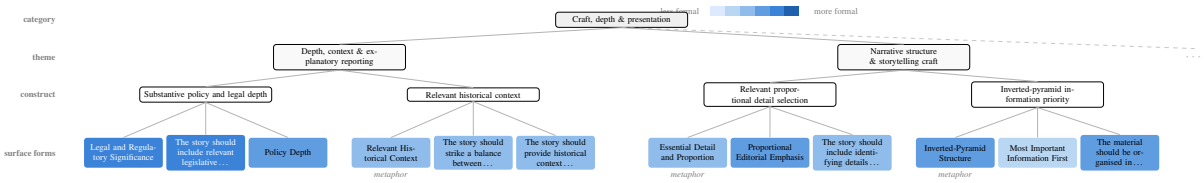


Figure 26: Journalism, rooted at the category *Craft, depth & presentation* (2 of its 8 themes shown).

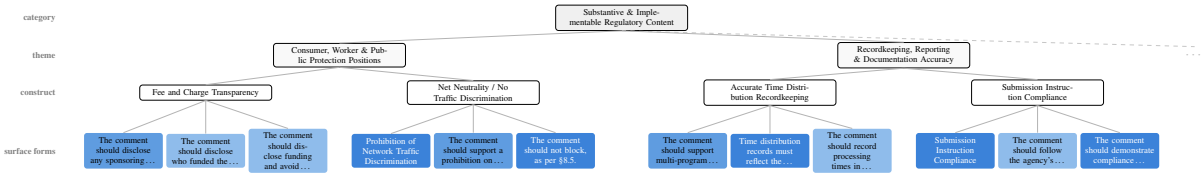


Figure 27: Regulatory Comments, rooted at the category *Substantive & Implementable Regulatory Content* (2 of its 10 themes shown).

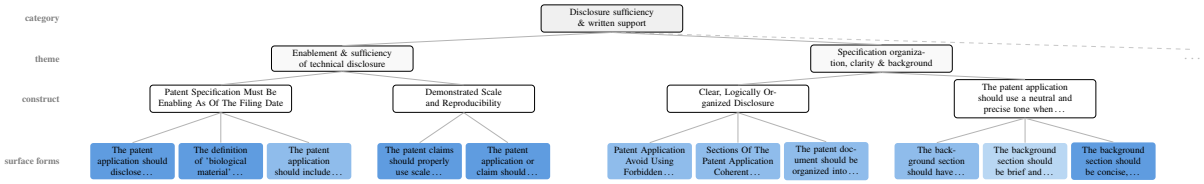


Figure 28: Patents, rooted at the category *Disclosure sufficiency & written support* (2 of its 8 themes shown).

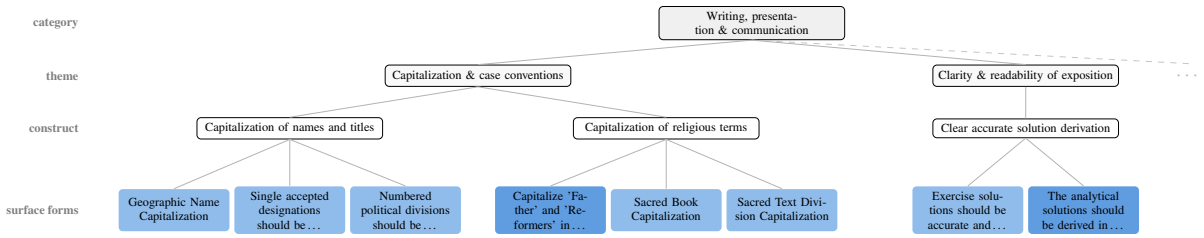


Figure 29: Academic Publishing, rooted at the category *Writing, presentation & communication* (2 of its 10 themes shown).

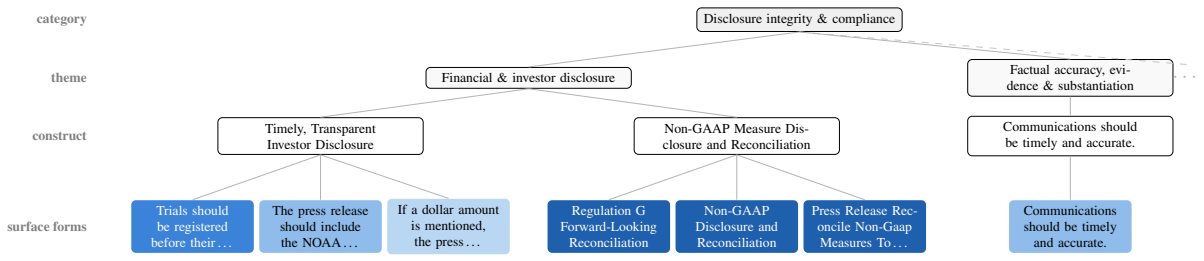


Figure 30: Press Releases, rooted at the category *Disclosure integrity & compliance* (2 of its 9 themes shown).

L Annotated name pairs

Table 23: Every name pair annotated by human annotators (97), sorted by the formality difference between the two names (Δf, on the 1–7 scale; blank where a name was not in the judged form pool). Set: T3 = the published Table 17 pairs; P11 = the name-swap pairs of Table 14; arms = the least/most formal arms of PREREG-15-v2; P25 = the PREREG-25 pairs. Annotators = construct-level label; Judges = GPT-5.6 / Claude under the requirement-level question (same, over[lap], diff[erent]). Pairs with $\Delta f \geq 3$ are the vastly different register cases.

Set	Id	Name A (less formal)	f_A	Name B (more formal)	f_B	Annotators	Judges
arms	a38	Inside humor	2.5	Affirmation: reliance on a shared sensus communis	7.0	same	over/same
arms	a40	Don't cater to the room	2.0	Audience interaction and segmentation as performance material	6.5	overlap	diff/diff
arms	a54	Make your jokes stand out	1.5	Personalization and presence of lived experience	6.0	overlap	over/diff
T3	a50	You don't have to be funny all the time	1.5	Humor quotient / entertainment vs. analysis balance	5.5	same	over/over
arms	a0	Punch up, not down	2.0	Tonal Discipline / Target Systems, Not Individuals	6.0	same	over/over
arms	a1	Make the fake news actually funny	2.0	Enjoyability/entertainment value as an evaluative criterion	6.0	same	over/over
arms	a10	Know your joke and make sure the reader knows your joke	2.0	Accuracy and nuance of political/media criticism	6.0	overlap	diff/diff
arms	a41	Stay in the bit / Don't step on laughs	2.0	Necessary condition: appropriate pacing	6.0	same	over/over
arms	a50	You don't have to be funny all the time	1.5	Farcical scale and authenticity of laughter	5.5	same	diff/diff
arms	a71	Trim the fat	2.0	Precision, professionalism, and hilarity	6.0	same	over/diff
P25	a140	Eliminate lengthy build-ups; go straight to the punch line	2.5	Brevity (essential quality)	6.0	same	same/same
T3	a36	Keep it simple; don't overthink	2.5	Specificity and lyricism in writing	6.0	overlap	diff/diff
arms	a16	Six Rules of Humor – Make it relevant	3.5	Necessary condition – shared sensus communis (explicit)	7.0	same	over/over
arms	a29	Gross-out / boundary-pushing humor	3.0	Subject suitability / moral boundary consideration	6.5	same	over/over
arms	a36	Keep it simple; don't overthink	2.5	Specificity and vivid imagery	6.0	overlap	diff/diff
arms	a46	Pop-culture references and playful wordplay as comedic fuel	3.5	Intertextuality and cultural literacy	7.0	same	over/over
arms	a49	Edginess / lowbrow tone	3.0	Balancing atrocity and festivity; racial context sensitivity	6.5	same	over/over
arms	a64	hack / hacky / comic's clichés (anti-pattern)	2.5	Reject client/performer insertions of clichés	6.0	same	over/same
arms	a66	Beginner's Rubric: Does this joke fit your POV (would only YOU write it)?	2.5	Persona as Performance Foundation	6.0	same	over/over
P25	a202	Omit needless words	4.0	Brevitas (brevity)	7.0	same	same/same
P25	a206	Does the switch deliver something unexpected?	3.5	Incongruity conceptualized as Surprise	6.5	same	same/same
P25	a222	Rule 6: Don't get caught by the logic police	2.5	Premise test – internal consistency	5.5	same	same/same
P25	a281	Is there a clean turn (twist)?	2.5	Punchline is incongruous with the prediction	5.5	same	same/same
arms	a6	No toilet humor	2.5	Sophisticated humour and absence of offensive material	5.5	same	over/over
arms	a8	Taste and crudeness	3.5	Catharsis and social release	6.5	overlap	diff/diff
arms	a17	Audience fit	3.5	Dependence on context and contradiction for humour (audience inference requirement)	6.5	same	over/over
arms	a19	Audience cringes with recognition ('we've all been THAT person')	3.5	Relatability / observation-based comedy as benign violation	6.5	same	over/over
arms	a20	Trying to be funny instead of being true (anti-pattern)	2.5	Comedy demands acute observation and sensitivity	5.5	overlap	over/over
arms	a52	Create content people actually want to watch	3.0	Autonomy enables risk-taking and distinctive voices	6.0	overlap	diff/diff
arms	a58	Stand-up comedy mode	3.0	Cross-cultural breadth and comparative perspective	6.0	different	diff/diff
arms	a79	Kibblesmith Rubric – will it still be funny tomorrow morning?	3.0	Topicality / attention to current events	6.0	same	over/over
P25	a1	Make the fake news actually funny	2.0	Optimize for funny (working rubric phrasing)	4.5	same	same/same
P25	a16	Sufficient shared background	4.5	Necessary condition – shared sensus communis (explicit)	7.0	same	same/same
P25	a17	Audience fit	3.5	Respect audience cultural sensibilities and align tone to readership	6.0	same	same/same
P25	a71	Trim the fat	2.0	Does the joke compress to its minimum?	4.5	same	same/same
P25	a100	Give the heckler a chance; fire a warning shot before gunning him down.	2.5	Prefer a moderate, non-abusive opening response	5.0	same	same/same
P25	a112	Six Rules of Humor – Be original	3.5	Originality / innovation / subversion of expectations	6.0	same	same/same
P25	a118	Word economy: every word earns its place (Is This Funny? rubric)	3.5	Brevity / Economy	6.0	same	same/same
P25	a155	Sufficient shared background	4.5	Necessary condition – shared sensus communis (explicit)	7.0	same	same/same
T3	a0	Punch up, not down	2.0	Working Rubric – Presence of the target	4.5	overlap	over/over
T3	a64	hack / hacky / comic's clichés (anti-pattern)	2.5	Originality (surprise effect)	5.0	same	over/same
arms	a7	Blue comedy	3.0	Typical audience and developmental appeal	5.5	same	diff/diff
arms	a12	Use – Humor	3.5	Political / Social Commentary	6.0	overlap	over/diff
arms	a18	Avoiding 'low-hanging fruit' topical jokes	3.5	Topical/contextual resonance and target alignment	6.0	same	over/over
arms	a31	Timing / 'Too soon!' sensitivity	3.5	Ethical and legal considerations when dealing with real historical atrocities	6.0	same	over/over
arms	a32	Blue comedy	3.0	Use of crude/sexual/scatological material	5.5	same	same/same
arms	a43	Blend of writing and drawing ('novel-in-pictures')	3.5	Cover art: unified, sophisticated, provocative design	6.0	different	over/diff
arms	a65	Tell stories where you're the idiot – but trying	3.0	Self-deprecation and personal material	5.5	same	over/over
P25	a0	Don't exploit or trivialize marginalized people	4.0	Artist responsibilities in depicting marginalized communities	6.0	same	same/same
P25	a20	Laugh should come from recognition ('I thought I was alone')	3.5	Recognizability of human truth (Mankoff selection)	5.5	same	same/same
P25	a31	Timing / 'Too soon!' sensitivity	3.5	Sensitivity and timing of material (controversy management)	5.5	same	same/same
P25	a40	Crowd work (audience interaction)	4.0	Audience interaction and participation	6.0	same	same/same

Set	Id	Name A (less formal)	f_A	Name B (more formal)	f_B	Annotators	Judges
P25	a66	Is there a clear comic persona?	4.0	Persona as Performance Foundation	6.0	same	same/same
P25	a91	Mine your own life for comedy ideas	2.5	Use of personal experience as material	4.5	same	same/same
P25	a113	Surprise	4.0	Surprise (violation of expectation)	6.0	same	same/same
P25	a142	Mimic normal tone/wording for context, then break it with unexpected twist	4.0	Use reversal (subvert expectations) as a comedic mechanism	6.0	same	same/same
P25	a144	Puns and word play	2.5	Pun - community-approved engine	4.5	same	same/same
P25	a162	Find a unique angle – avoid the obvious read	3.5	Originality of conception – avoid repetition of stock situations	5.5	same	same/same
P25	a177	Word play and puns	2.5	Wordplay and double-meaning exploitation	4.5	same	same/same
T3	a32	Does the joke land at the target audience's 'clean' level?	3.5	Use of crude/sexual/scatological material	5.5	same	over/over
T3	a142	Storytelling with long setups and twist endings	3.5	Exclusion – Pure congruity (predictable punchline)	5.5	same	diff/diff
arms	a33	Balance pain and humor – timing for turning trauma into jokes	4.5	Sensitivity of comedic depiction of suffering (ethical considerations)	6.5	overlap	over/same
arms	a34	Anchor nonsense with truthful elements	4.0	Authenticity/realism in depicting depression	6.0	overlap	over/over
arms	a45	Core principle: surprise (set up implied expectation then shatter it)	5.0	Cognitive elaboration and comprehension difficulty as contributors to humor	7.0	overlap	diff/diff
arms	a55	Politeness work	4.5	Audience complicity and false moral superiority	6.5	overlap	diff/diff
arms	a62	Host role / performance style	4.0	Permissible performance modes: manic, smutty, abusive, improvisatory	6.0	overlap	over/over
P25	a19	Audience cringes with recognition ('we've all been THAT person')	3.5	Recognition-based landing (not provocation)	5.0	same	same/same
P25	a38	Necessary condition – return to common sense	5.5	Affirmation: reliance on a shared sensus communis	7.0	same	same/same
P25	a41	Comedic timing and delivery via subtle cues	4.5	Delivery: intonation, inflection, attitude, timing (beat)	6.0	same	same/same
P25	a80	Puns and word play	2.5	Puns as comedy device (punchlines and parodies)	4.0	same	same/same
P25	a86	N – Normality (situation is actually okay)	4.0	No serious practical concern engaged	5.5	same	same/same
P25	a120	Recurring motifs, settings and running gags	3.5	Narrative devices and recurring gags	5.0	same	same/same
P25	a122	Use historical accuracy and avoid spreading false claims	4.0	Factually defensible / research-vetted	5.5	same	same/same
P25	a178	Use specific, recognizable details	4.5	Specificity and vivid imagery	6.0	same	same/same
P25	a196	Strong button (sketch)	3.0	Rubric question: Is the button satisfying?	4.5	same	same/same
P25	a198	Surreal humor and absurd elements	4.5	Surreal humour and the absurd when mixing serious with trivial intentionally	6.0	same	same/same
P25	a205	Set expectations then subvert (pull the rug out)	3.0	Punchline-driven structure: setup, expectation, twist	4.5	same	same/same
arms	a2	Use of humor / comedic tone	4.0	Humor effectiveness (laugh frequency/impact)	5.5	same	over/over
arms	a24	Avoid cynicism; preserve warmth and wit	4.5	Seriousness and straight-faced storytelling restore humanity	6.0	same	over/over
T3	a196	Strong button (sketch)	3.0	Credits interrupted by stingers or extra scenes (stinger after or during credits)	4.0	different	over/diff
T3	a118	Rule 12: Be brutal	3.5	Economy of language – eliminate unnecessary words	4.5	same	over/over
T3	a2	Use of humor / comedic tone	4.0	Humor / comedic effectiveness	5.0	same	same/over
T3	a86	Failure mode: R2 too hard to find	4.5	Object is a perceived incongruity	5.5	overlap	over/diff
P11	0	playing to the room		aptum		same	–
P11	1	trim the fat	2.0	pleonasm		same	–
P11	2	wordplay		paronomastic		same	–
P11	3	long walk		cumulative humor		same	–
P11	4	button		logical mechanism		overlap	–
P11	5	callback		reincorporation		same	–
P11	6	blow		presentation		same	–
P11	7	start close to the end		in media res		same	–
P11	8	show don't tell		specificity at moment of action		same	–
P11	9	keep your cast small		dramatis personae		overlap	–
P11	10	reaching the reader		profluence		overlap	–
P11	11	struggle		progressive complication		same	–
P11	12	bad guy close in		escalation		same	–
P11	13	world of hats		originality		overlap	–