TACTICS: Taxonomy-Aware Test Suites for Fine-Grained Machine Translation Evaluation

Prasanth Bathala, Anubhav Shrimal, Sukhdeep Singh Kharbanda, Pradyumna Lanka, Rohit Dhaipule
Translation Services, Amazon
{pbathala, shrimaa, kharsukh, pradyl, rdhaipu}@amazon.com

Abstract

Large-scale machine-translation (MT) systems are typically evaluated on random samples from a corpus whose distributional composition is an artifact of how it was assembled. Such a sample inherits the phenomena the collection happens to contain rather than the full space a system must handle, spanning rule-governed conventions (terminology, punctuation, currency formatting) and context-dependent phenomena (tone, honorifics, document-level coherence), and thus provides no coverage guarantee for assessing robustness. We propose **TACTICS** (**T**axonomy-**A**ware **C**overage-op**T**imized **I**ntelligent **C**orpus **S**ampling), which recasts coverage as an explicit objective. TACTICS induces a hierarchical taxonomy from a locale style guide, classifies segments against it, and selects a fixed-budget subset jointly optimizing coverage of rare categories, document-level coherence, and distributional fidelity to the full corpus. Applied to MT evaluation across four translation directions, TACTICS improves coverage of rare categories over lexical and embedding-based selection. By targeting the phenomena that separate systems, TACTICS makes a fixed evaluation budget go further, recovering the true system ranking from far fewer segments than random sampling wherever a real quality gap exists and never signaling a difference where none exists.

1 Introduction

Modern machine translation (MT) systems, including both dedicated neural MT engines and LLM-based translation pipelines, are increasingly deployed at scale across many language pairs and locales (Bahdanau et al., 2015; Vaswani et al., 2017; Brown et al., 2020; Touvron et al., 2023; OpenAI, 2023). At this scale, evaluation is a bottleneck. Human assessment of every translated segment is prohibitively expensive, so systems are typically evaluated on a random sample drawn from a large corpus. The implicit assumption is that such a sample provides a representative picture of system quality. But a representative sample is not the same as a diagnostic one. A random sample reproduces the corpus faithfully, yet provides no guarantee of adequate coverage of the phenomena that discriminate between systems.

The difficulty is that these discriminating phenomena are also the ones a corpus represents least evenly. Most of a translation corpus consists of short, common segments that nearly any system translates correctly, whereas the cases that genuinely stress a system, such as terminology constraints, locale-specific formatting, and date or currency conventions, are comparatively rare. This matters because an evaluation, whether by automatic metric (Sai et al., 2022) or human protocol such as MQM (Lommel et al., 2014), is only as diagnostic as the set on which it is computed. When easy segments dominate that set, the differences concentrated in the harder ones are averaged away, and two systems of unequal quality can appear all but indistinguishable.

Existing selection methods do not close this coverage gap. Most target training-data curation rather than evaluation, whether through quality filtering (Gao et al., 2021; Marion et al., 2023; Tirumala et al., 2023), importance resampling on lexical or n-gram statistics such as DSIR (Xie et al., 2023), or embedding-space diversity such as ClusterCLIP (Shao et al., 2024). These signals capture frequent lexical patterns or broad semantic variation, but none of them explicitly models the prescriptive rules that govern localization, such as whether a brand term is preserved, punctuation follows locale conventions, or currency and dates are formatted correctly. A sample can therefore be lexically or semantically diverse and still fail to cover the phenomena on which systems actually differ.

We propose **TACTICS** (**T**axonomy-**A**ware

Coverage-op**T**imized **I**ntelligent **C**orpus **S**ampling), a data-selection method that constructs compact, stratified evaluation sets by treating coverage as an explicit objective rather than a byproduct of how the corpus was sampled. It formulates evaluation-set construction as a constrained optimization problem over a hierarchical taxonomy induced from prescriptive style guides. We instantiate this for MT, though the approach applies to any evaluation that must cover a structured set of phenomena under a fixed budget. TACTICS balances three properties (Figure 1).

(i) **Coverage.** Every taxonomy category, including rare ones, is adequately represented under a fixed budget, unlike random sampling.

(ii) **Coherence.** Segments are selected with their document neighbors, so each is evaluated in context rather than in isolation.

(iii) **Fidelity.** The sample matches the full corpus in category composition and difficulty.

We evaluate TACTICS on a large-scale localization corpus across four translation directions. At a fixed budget of $B = 5{,}000$, TACTICS raises the average number of taxonomy categories with at least 20 sampled segments from 56.3 to 73.3 relative to random sampling, while preserving high correlation with the full-corpus category distribution (Pearson $r = 0.949$ to 0.982) and close alignment on edit rate. It also increases the average segments per document from 1.9 to 3.2, yielding less fragmented samples. Most importantly, these subsets are more diagnostic. Where a real quality gap exists, TACTICS recovers the true system ranking from fewer segments than random, with a higher paired t-statistic at every budget on en→hi, and raises no false signal on directions where the systems are tied.

Our contributions are:

- We introduce a **style-guide-grounded taxonomy induction pipeline** for representing localization phenomena relevant to MT evaluation.
- We formulate fixed-budget evaluation set construction as a **taxonomy-aware constrained sampling problem** over coverage, coherence, and fidelity.
- We show that TACTICS **covers rare phenomena** and **recovers true system quality differences** more reliably than random, lexical, and embedding-based selection.

2 Related Work

2.1 MT Evaluation and Test Set Design

MT evaluation relies on automatic metrics such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006), chrF (Popović, 2015), COMET (Rei et al., 2020), and BLEURT (Sellam et al., 2020), whose reliability for ranking systems has been studied at scale (Kocmi et al., 2021), alongside human frameworks such as MQM (Lommel et al., 2014; Freitag et al., 2022). These estimate quality on a test set but not whether the set is diagnostic, that is, whether it contains the phenomena on which systems differ. This depends on how the set is built. The WMT shared task builds its test sets by drawing source texts from a chosen domain, usually online news, and commissioning reference translations for the task, leaving targeted phenomenon evaluation to separately contributed test suites (Barrault et al., 2020). Its composition is therefore fixed by the choice of domain and source rather than by explicit coverage of these phenomena, and reproducing it for a new language pair or domain requires substantial manual effort. Benchmarks like FLORES-101 and FLORES-200 scale instead by sampling their sentences from a corpus such as Wikipedia (Goyal et al., 2022; NLLB Team et al., 2022), but inherit whatever the corpus contains, with no guarantee that the discriminating phenomena are covered. TACTICS addresses both limitations by making coverage of a style-guide-derived taxonomy an explicit objective when sampling a corpus, so those phenomena that discriminate systems are represented by construction rather than by chance or manual curation.

2.2 Data Selection and Sampling

Prior data-selection work has largely focused on training data, including quality filtering (Gao et al., 2021; Marion et al., 2023), pruning (Tirumala et al., 2023), importance resampling such as DSIR (Xie et al., 2023), and embedding-based diversity such as ClusterCLIP (Shao et al., 2024). These methods select examples using lexical statistics, embeddings, or training utility.

A separate line of work selects data for *evaluation* rather than training, so that systems can be tested on fewer examples. One group reduces a benchmark to a small subset that still reproduces its system ranking, keeping the examples that mod-

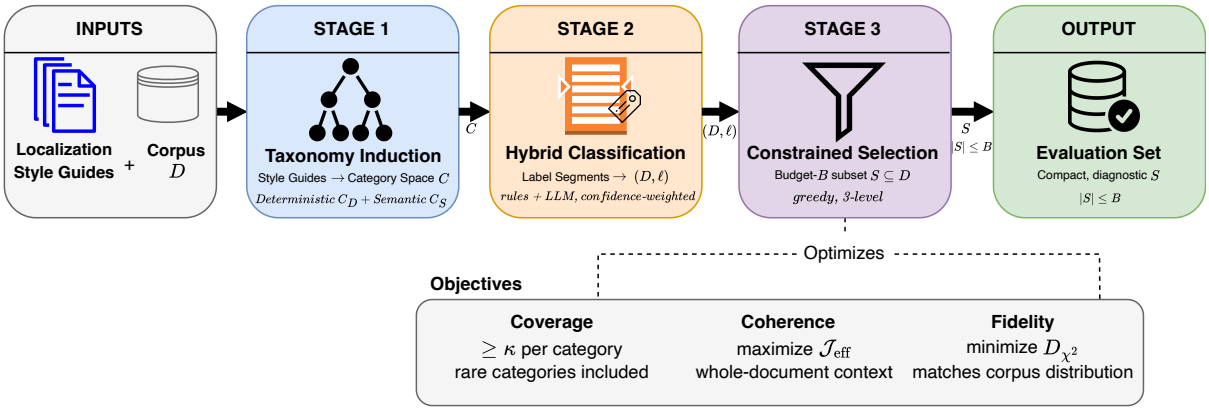


Figure 1: Overview of TACTICS. Stage 1 induces a taxonomy (category space C) from style guides; Stage 2 labels segments into C, yielding (D, ℓ); Stage 3 selects a budget-B subset S jointly optimizing coverage ($\geq \kappa$), coherence ($\mathcal{J}_{\text{eff}}$), and fidelity (D_{χ^2}).

els find hardest or most discriminating (Vivek et al., 2024; Maia Polo et al., 2024; Rodriguez et al., 2021; Vania et al., 2021; Perlitz et al., 2024). Another selects examples adaptively, spending human judgments on the comparisons that best separate systems (Chaganty et al., 2018; Mohankumar and Khapra, 2022; Kossen et al., 2021). More recently, Zouhar et al. (2025) pick segments for human evaluation of NLG systems from metric-score variance, output diversity, and item difficulty. All of these rely on signals read from model outputs. The selection is therefore tied to the systems being evaluated, does not transfer easily to new ones, and cannot ensure that the phenomena a system should handle are covered.

TACTICS inverts this dependence. We cast evaluation-set construction as taxonomy-aware stratified sampling under a fixed budget, with strata drawn from a prescriptive style-guide taxonomy rather than surface metadata, domains, or embedding clusters. The taxonomy is fixed independently of any model, so the resulting evaluation set is reusable across systems, including ones built later. This approach suits MT, where much of localization quality depends not on meaning but on explicit rules, such as preserving brand terms, formatting dates and currencies for the locale, or choosing the correct honorific. These rules are set by style guides, yet selection by lexical or embedding similarity fails to capture them (Section 4.3).

2.3 Automated Taxonomy Construction

Recent LLM-based taxonomy induction methods build taxonomies bottom-up from observed data, e.g., TnT-LLM (Wan et al., 2024) and Chain-of-Layer (Zeng et al., 2024). Such methods discover corpus-salient patterns, but localization evaluation also depends on prescriptive rules that may be scarce in a given corpus, such as currency formatting, brand preservation, quotation marks, or date conventions. TACTICS takes a complementary top-down approach. It induces taxonomy structure from localization style guides and grounds it in the corpus through constrained classification and sampling, so evaluation sets cover both common patterns and rare style-guide-defined phenomena that matter for system comparison.

3 Methodology

3.1 Problem Formulation

Let $D = \{s_i\}_{i=1}^{n}$ denote translation segments, with each segment belonging to a document $J \in \mathcal{J}$. Let T be a hierarchical taxonomy with category set $C = C_D \cup C_S$, where deterministic categories C_D are rule-matchable and semantic categories C_S are LLM-inferred. Each segment has taxonomy labels $\ell(s) \subseteq C$ and a post-edit score $e(s) \in [0, 1]$, the fraction of a machine translation that a human corrected (approximated by TER). The score can come from any MT system and serves only to approximate the segment's difficulty.

Given a budget $B \ll |D|$, we select a stratified evaluation subset $S \subset D$, $|S| \leq B$, by enforcing coverage constraints and optimizing the coherence and fidelity objectives below.

Coverage (constraint). For a minimum threshold $\kappa \geq 1$, each category with at least κ segments

in the corpus must be represented in the sample:

$$\forall c \in C: \quad |\{s \in S : c \in \ell(s)\}| \geq \kappa. \quad (1)$$

Coherence (maximize). Sampling entire documents preserves the context surrounding each segment. Beyond this, we favor samples spread across many documents rather than a few, measured by the effective number of documents:

$$\mathcal{J}_{\text{eff}}(S) = \exp\Big(-\sum_{j=1}^{|\mathcal{J}|} p_j \log p_j\Big), \quad (2)$$

where $p_j = |S \cap J_j|/|S|$ is document J_j's share of the sample. Exponentiating rescales the entropy into an interpretable document count without changing the optimum.

Fidelity (minimize). We align the sample S with the corpus D in category composition and difficulty. The category distribution P_C gives the relative frequency of each taxonomy category in C, and P_e is a histogram of the post-edit score $e \in [0, 1]$ over fixed bins, both empirical distributions computed on the corpus (P) and the sample ($\hat{P}$). We minimize their chi-squared divergence,

$$D_{\chi^2}(\hat{P}_C \,\|\, P_C) + \lambda\, D_{\chi^2}(\hat{P}_e \,\|\, P_e), \quad (3)$$

with $\lambda > 0$ trades off category against difficulty alignment.

3.2 Method Overview

TACTICS formulates stratified evaluation-set construction as constrained selection over a taxonomy-induced representation of the full corpus. The framework has three stages (Figure 1): taxonomy induction defines the category space, segment annotation maps the corpus into this space, and constrained sampling selects a fixed-budget subset that balances coverage of rare phenomena, document-level coherence, and fidelity. This yields compact evaluation sets that preserve both the linguistic structure and the key statistical properties of the full corpus.

3.2.1 Taxonomy Extraction (Stage 1)

We induce a hierarchical taxonomy T from locale-specific style guides to define the category space C used for sampling. This is necessary because the corpus does not explicitly annotate linguistic phenomena, and embeddings or metadata alone do not capture prescriptive constraints such as terminology, branding, formatting, or locale-specific

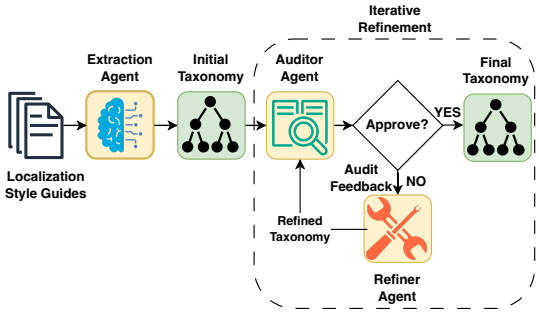


Figure 2: Style-guide-grounded taxonomy induction. An *Extraction Agent* turns style guides into an initial taxonomy; an *Auditor Agent* then checks its coverage, looping through a *Refiner Agent* until it approves or an iteration limit is reached.

conventions. The taxonomy is partitioned into *Deterministic* categories C_D, covering high-precision rule-matchable phenomena, and *Semantic* categories C_S, covering context-dependent phenomena such as tone, register, or implicit meaning.

We construct T with an agentic extract–audit–refine loop (Figure 2; Appendix I). An Extraction Agent identifies candidate rules from the style guide and organizes them into an initial hierarchy, an Auditor Agent checks coverage against the source, and a Refiner Agent incorporates missing or under-specified rules. The loop terminates when no major gaps remain or an iteration limit is reached, yielding the prescriptive category space used for annotation, coverage constraints, and distributional alignment in later stages.

3.2.2 Hybrid Classification (Stage 2)

Given taxonomy T, we define a labeling function $\ell(s) \subseteq C$ that maps each segment $s \in D$ to one or more taxonomy categories, embedding the corpus into the category space from Stage 1. Classification is hybrid. Deterministic categories C_D are assigned by rule-based matchers, such as regular expressions for terminology or formatting patterns, producing high-precision labels with confidence 1.0. Semantic categories C_S are assigned by a context-aware classifier whose outputs are constrained to valid taxonomy paths. Each segment receives a primary category and optional secondary categories to capture co-occurring phenomena. The resulting labeled dataset (D, ℓ) supports the coverage constraint and defines the empirical distributions used by the fidelity objective.

3.2.3 Stratified Sampling (Stage 3)

Given labeled data (D, ℓ), we greedily select a budgeted subset $S \subset D$, $|S| \leq B$, to approximate the constrained optimization problem of Section 3.1. At each step, we add the segment that most improves the objective while preserving satisfied constraints.

Level 1: Coverage enforcement. We first build a seed set that satisfies the minimum coverage constraint (Eq. 1) for each feasible category $c \in C$. Categories are processed from rare to frequent, and segments are selected using a fixed mix of confidence levels. Each selected segment is augmented with a local context window from the same document to preserve translation context. This stage ensures explicit coverage of long-tail categories and initializes S for later optimization.

Level 2: Coherence optimization. We next improve coherence by adding documents that contribute diverse taxonomy signals. Directly maximizing $\mathcal{J}_{\text{eff}}(S)$ (Eq. 2) is hard, since each document's contribution depends on its size and the evolving sample. We instead use a category-aware document score analogous to TF–IDF, with taxonomy categories playing the role of terms. For category c and document j,

$$\mathrm{CF}(c, j) = \frac{\#\{s \in J_j : c \in \ell(s)\}}{|J_j|}, \tag{4}$$

$$\mathrm{IDF}(c) = \log \frac{|\mathcal{J}|}{\mathrm{df}(c)}, \tag{5}$$

where $\mathrm{df}(c)$ is the number of documents containing c, CF is the category's frequency within a document, and IDF downweights categories common across documents. To stop a category repeated within one document from dominating, we saturate CF with hyperparameter k_1,

$$\mathrm{CF}_{\text{sat}}(c, j) = \frac{\mathrm{CF}(c, j)(k_1 + 1)}{\mathrm{CF}(c, j) + k_1}, \tag{6}$$

and score each document as

$$\mathrm{score}(j) = w_{\text{conf}}(j)\, w_{\text{size}}(j) \times \sum_{c \in C} \mathrm{CF}_{\text{sat}}(c, j)\, \mathrm{IDF}(c), \tag{7}$$

where w_{conf} favors reliable assignments and w_{size} downweights documents that are too small or too large (Appendix A, B). At each step we add the highest-scoring document's remaining segments to S.

Pair	Segments	Docs	Categories	Edit rate (%)
en → fr	57,524	13,211	149	56.6
en → hi	64,932	18,897	80	53.1
ja → en	83,522	9,379	126	60.4
ja → zh	82,578	9,659	127	68.0

Table 1: Locale pairs used for baseline comparison analysis. Segments and documents from the classified evaluation pool; categories from the merged taxonomy; edit rate from post-edit annotations.

Level 3: Fidelity alignment. With the remaining budget, we reduce the fidelity objective (Eq. 3) to align the sample with the corpus in category distribution and edit-rate profile. Estimating target statistics from D, we add segments that close the largest gaps, selecting edited segments to match the target difficulty profile and unedited segments from underrepresented categories. This corrects the distributional drift introduced by coverage and coherence optimization, yielding a sample that faithfully reflects the corpus in both composition and difficulty.

4 Experiments

4.1 Data and Baselines

We evaluate on four localization translation directions, en→fr, en→hi, ja→en, and ja→zh. Each direction uses a proprietary, in-house localization post-editing corpus in which source segments are machine-translated and then post-edited by professional linguists. Each segment thus carries a source, an MT translation, and a post-edited target, and is linked to its source document. Its post-edit score is the edit rate between the MT output and the post-edit, approximated by TER (Section 3.1). Table 1 summarizes the pools, which range from 57K–84K segments, 9K–19K documents, 80–149 taxonomy categories, and edit rates from 53.1% to 68.0%. We also use target-locale style guides, which specify orthographic, lexical, and formatting constraints not observable from the corpus, to induce the taxonomy (Section 3.2.1, example in Appendix C).

All methods sample $B = 5{,}000$ segments per locale from the same classified evaluation pool. We compare TACTICS against Random uniform sampling, DSIR (Xie et al., 2023) (importance resampling on n-gram statistics), and ClusterCLIP (Shao et al., 2024) (embedding clustering with balanced sampling).

Locale ($\lvert C\rvert$)	Method	$\text{Cov}_{20}\lvert\text{Cov}_1$	Gini	Pearson r	$\text{MSE}_C \downarrow$	$\text{MSE}_{\text{TER}} \downarrow$	Edit (%)	S/D
en→fr (149)	Random	56|106	0.421	**0.999**	0.43	5.26	56.7	2.0
	DSIR	13|70	0.133	0.983	**0.38**	**1.46**	64.8	1.2
	ClusterCLIP	56|101	0.426	0.999	1.23	24.6	58.2	2.0
	TACTICS	**70|133**	**0.497**	0.982	6.90	95.9	**56.6**	**3.1**
en→hi (80)	Random	64|79	0.388	**0.999**	3.66	**7.98**	53.4	1.8
	DSIR	9|70	0.164	0.952	49.6	66.4	56.8	1.2
	ClusterCLIP	66|80	0.393	0.999	**1.24**	14.9	55.5	1.8
	TACTICS	**79|80**	**0.468**	0.969	46.5	133	**53.1**	**2.4**
ja→en (126)	Random	53|85	0.349	**1.000**	**0.09**	**11.5**	59.9	1.9
	DSIR	11|57	0.151	0.892	49.8	843	81.0	1.2
	ClusterCLIP	54|87	0.331	0.942	26.8	1090	74.0	1.8
	TACTICS	**85|102**	**0.491**	0.972	11.8	111	**60.4**	**3.4**
ja→zh (127)	Random	52|96	0.346	**0.999**	**0.08**	**6.02**	67.8	1.9
	DSIR	7|57	0.114	0.893	50.2	1510	88.1	1.1
	ClusterCLIP	51|94	0.308	0.929	25.1	1690	85.7	1.7
	TACTICS	**59|113**	**0.550**	0.949	9.17	71.6	**68.0**	**4.0**

Table 2: Intrinsic sampling quality at $B = 5{,}000$. $\text{Cov}_{20}\lvert\text{Cov}_1$ counts categories with ≥ 20 and ≥ 1 samples; r is Pearson correlation with the full-corpus category distribution; MSE_C and MSE_{TER} ($\times 10^{-6}$) are mean squared errors of the sampled vs. full-corpus category and TER distributions; Edit is post-edit rate; S/D is segments per document. Random matches the corpus distributions most closely, while TACTICS trades fidelity to cover rare categories and harder segments. Baselines: DSIR (Xie et al., 2023), ClusterCLIP (Shao et al., 2024).

4.2 Evaluation Setup

We select pipeline models by role, as the method is agnostic to the specific LLMs. Taxonomy induction runs once per locale, so we use **Claude Sonnet 4.5** to generate the taxonomy and the more capable **Claude Opus 4.5** to audit it for missing or under-specified rules (Figure 2, Section 3.2.1). Classification runs over the whole corpus, so we label segments with **Kimi K2.5**, constraining its outputs to valid taxonomy paths. From a different family than the taxonomy models, it labels cheaply at scale and independently of the generator. For downstream evaluation we compare NMT_{base}, a dedicated neural MT system, with LLM_{MT}, a **Gemma 4-26B-A4B** model prompted to translate, conditioned on summarized style guides.

We report intrinsic metrics aligned with the three objectives of TACTICS. **Coverage** is Cov_{20} and Cov_1, the number of categories with at least 20 and at least 1 sampled segments. **Coherence** combines the Gini coefficient over per-document segment counts with segments per document (S/D), where higher values mean more multi-segment context. **Fidelity** compares the sample to the full corpus via Pearson correlation (r) and category- and TER-distribution MSE (higher r, lower MSE better), with edit rate as a difficulty proxy. Downstream, we report TER (Snover et al., 2006), BLEU (Papineni et al., 2002), and chrF (Popović, 2015) against the human post-edited target, together with withholding rate (WH), the fraction of segments passing an automated QA pipeline. The pipeline applies deterministic checks (a translation-memory match against known-good references, and length heuristics that flag implausible target-to-source length ratios or name lengths) together with a learned classifier that predicts translation quality from the source and target sentence. These checks use no style guides, so they are independent of the taxonomy and the LLM_{MT} prompt and favor neither system.

4.3 Results

We provide taxonomy consistency checks, including inter-model agreement and manual spot checks, in Appendix J.

Intrinsic sampling quality (Table 2). TACTICS improves coverage and coherence across all directions. Cov_{20} rises over Random from 56 to 70, 64 to 79, 53 to 85, and 52 to 59 for en→fr, en→hi, ja→en, and ja→zh. It also attains the highest Gini and segments per document (S/D) everywhere, selecting multiple segments from the same document rather than isolated ones. This matters because translation quality is context-dependent, and a segment like "Click here" is only translatable alongside its neighbors. But covering the tail pulls the sample away from the corpus. Random most closely matches the full-corpus category and TER distributions (lowest MSE), since it simply reproduces the corpus proportions, whereas TACTICS deviates more to represent rare categories

Locale	Metric	NMT_{base}				LLM_{MT}			
		Rand.	ClusterCLIP	DSIR	TACTICS	Rand.	ClusterCLIP	DSIR	TACTICS
en→fr	TER↓	0.309	**0.305**	0.309	0.315	0.267	0.252	0.257	**0.231**
	BLEU↑	0.655	0.662	0.655	**0.668**	0.686	0.699	0.690	**0.717**
	chrF↑	0.795	0.800	0.798	**0.804**	0.799	0.812	0.807	**0.833**
	WH↑	**6.50**	5.48	6.24	5.64	7.80	5.90	7.18	**11.84**
en→hi	TER↓	0.349	0.357	0.409	**0.343**	0.322	0.328	0.341	**0.320**
	BLEU↑	0.600	0.606	0.608	**0.625**	0.636	0.633	0.612	**0.641**
	chrF↑	0.716	0.718	0.700	**0.730**	0.699	0.702	0.686	**0.760**
	WH↑	1.55	1.70	1.45	**2.50**	3.40	3.65	3.20	**15.38**
ja→en	TER↓	**0.410**	0.422	0.422	0.560	0.323	0.301	**0.299**	0.328
	BLEU↑	0.619	0.616	**0.639**	0.634	0.728	0.753	0.750	**0.795**
	chrF↑	0.718	0.710	0.737	**0.745**	0.808	0.827	0.821	**0.847**
	WH↑	5.25	**6.39**	6.00	4.93	6.34	6.78	6.12	**31.79**
ja→zh	TER↓	0.306	0.319	**0.302**	0.319	0.146	0.116	0.145	**0.090**
	BLEU↑	0.576	0.561	0.585	**0.590**	0.788	0.833	0.790	**0.865**
	chrF↑	0.613	0.610	0.620	**0.713**	0.814	0.854	0.815	**0.907**
	WH↑	4.75	4.60	**4.90**	3.98	5.85	5.70	5.90	**27.38**

Table 3: Downstream translation performance across sampling strategies. Best value per direction, metric, and system is in **bold**. Lower TER is better, and higher BLEU, chrF, and WH are better. WH is the withholding rate, the percentage of segments passing automated QA checks. Baselines include DSIR (Xie et al., 2023) and ClusterCLIP (Shao et al., 2024).

Locale	Systems	Segs	True Δ	Verdict
en→fr	NMT / LLM	3959	$+0.016$	not sig.
en→hi	NMT / LLM	4334	$+0.314$	significant
ja→en	LLM^{-SG} / NMT	3168	$+1.121$	significant
ja→zh	LLM / NMT	3062	-0.086	significant

Table 4: Full-pool human-audit reference verdicts, the ground truth each subsample aims to reproduce. Systems: NMT (NMT_{base}), LLM (LLM_{MT}, style-guided), and LLM^{-SG} (LLM_{MT} without style guides); ja→en uses LLM^{-SG}, the only direction it was annotated. Each segment is judged by one linguist with a second-expert QA pass. MQM penalty $= 25\,\text{critical} + 5\,\text{major} + \text{minor}$ (lower is better), so $\Delta = A - B > 0$ means the first-listed system A is worse. Full t and p values in Appendix F.

and harder segments. Even so, TACTICS keeps the corpus category ordering (Pearson $r = 0.949$–0.982) and stays far more stable than DSIR and ClusterCLIP, whose MSE swings by orders of magnitude across directions.

Downstream translation performance (Table 3). On a random sample, NMT_{base} and LLM_{MT} score very close across TER, BLEU, chrF, and WH, since it rarely includes the segments that separate them. The TACTICS set covers these rarer and harder cases, and the gap becomes clear. On ja→zh the LLM_{MT}–NMT_{base} WH gap widens from about one point to over twenty, BLEU from 0.788 to 0.865 and chrF from 0.814 to 0.907, and the mean WH gap across directions grows from 1.3 to 17.3 points. Human evaluation shows the same pattern by category. Auditing one NMT system on comparably sized TACTICS and random samples (∼3,000 vs. 2,874 segments), the TACTICS audit covered 18 error categories against 15 for random, including three the random audit missed entirely, `number-format`, `text-truncation`, and `offensive-content`. Several of these carry major errors, three for text-truncation and one for offensive-content, so a random-only audit leaves real, category-specific failures unmeasured. We therefore use the two together, TACTICS to check per-category quality including the long tail, and random to estimate overall quality at the corpus proportion.

Sample efficiency (Table 4, Figure 3). Table 4 gives the true system ranking each subsample must recover, and Figure 3 measures how efficiently the samplers reach it. The four selectors are Random (uniform, unstratified), Equal-allocation (flat per-category quota), Low-confidence-first (lowest classifier-confidence segments, TACTICS's hard-example heuristic alone), and TACTICS (the full three-level pipeline). The three category-aware selectors differ only in how they spend the budget, isolating TACTICS's allocation strategy rather than the corpus-level baselines of Section 4.1. The top row is wasted budget, the fraction of a batch that both systems translate cleanly and that reveals nothing about their difference, which TACTICS keeps lowest in most settings. The bottom row is the paired t-statistic magnitude $|t|$ per batch, larger when a batch separates the systems more clearly and significant above the dashed line. On en→hi

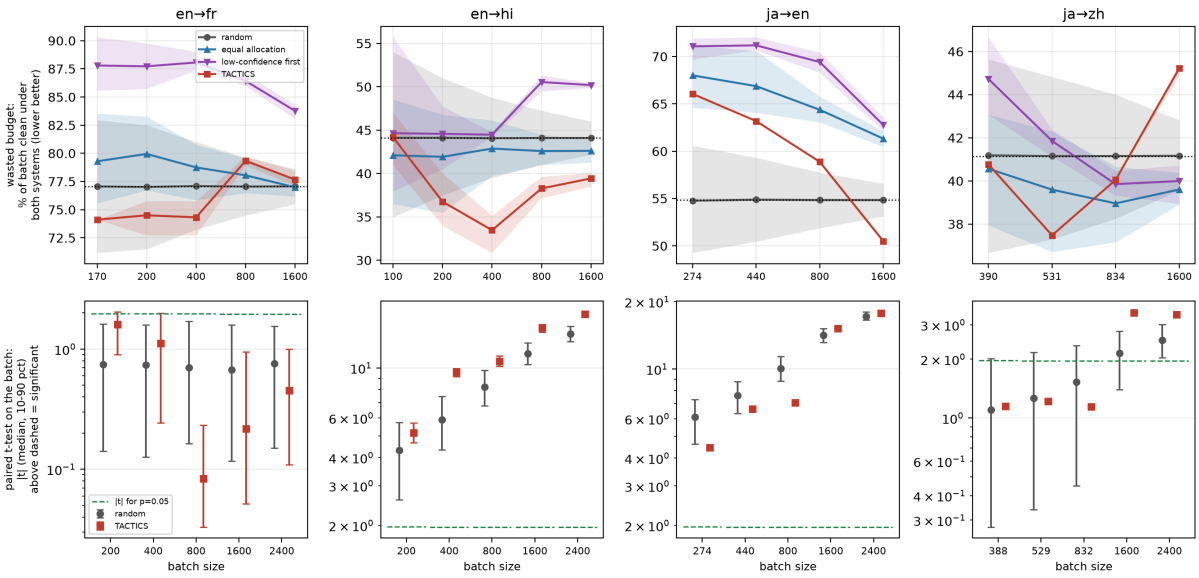


Figure 3: Sample efficiency across four directions. **Top**, wasted budget, the percentage of a batch that is clean under both systems and so cannot separate them (lower is better). **Bottom**, the paired $|t|$ on the batch (median over draws, 10–90 percentile bars). Points above the dashed line are significant at p=0.05. TACTICS is compared against random, equal-allocation, and low-confidence-first selection.

TACTICS exceeds random at every budget (ratio 1.19–1.63), and on ja→zh it clears the threshold at 1600 and 2400 segments where random does not. en→fr is a control, where the systems are tied ($\Delta = 0.016$) and both samplers stay below the line, so TACTICS signals a difference only when one is real. Although it estimates the corpus average less accurately (Appendix F), TACTICS detects a real gap from far fewer segments than random.

5 Discussion

A representative sample is not a diagnostic one. Random sampling reproduces the corpus almost exactly, which is why it struggles to compare systems. Most of a corpus is routine text both systems handle, so a faithful sample is dominated by segments that carry no signal about their difference, while the cases that separate systems, such as rare terminology, date and number formats, or long-context documents, are easily left out. A sample can thus match the corpus yet fail to reveal which system is better.

Generic selectors miss prescriptive rules. DSIR and ClusterCLIP rank segments by lexical or embedding similarity, capturing frequent surface patterns but not the prescriptive rules localization depends on, such as terminology, punctuation, and date, currency, and locale-specific formats. They therefore cover fewer taxonomy categories than TACTICS, and their sample statistics vary widely across directions.

Coverage trades against fidelity by design. By guaranteeing rare categories and harder segments, TACTICS deviates from the corpus more than random and so has higher distributional error. This is intentional. TACTICS keeps the corpus category ordering while shifting weight to the long tail, making it a weaker estimator of the corpus average but a stronger detector of system differences. Fidelity is a constraint to stay close enough, not the objective to maximize.

Use both samplers together. The two roles are complementary. Human evaluation is expensive, so the limited annotation budget is best spent on a TACTICS set, which guarantees that every content type is judged, including the rare and difficult cases random evaluation can miss. A cheaper random sample then estimates overall quality at the corpus proportion. Together they answer whether a system is good everywhere and how it performs on average.

6 Conclusion and Future Work

We presented TACTICS, a taxonomy-aware method for building compact, diagnostic MT evaluation sets. By inducing categories from local-

ization style guides and jointly optimizing coverage of rare phenomena, document-level coherence, and distributional fidelity, TACTICS guarantees that evaluation covers the content types that distinguish systems, including the long tail that random sampling misses. Across four directions it improves category coverage over random, lexical, and embedding-based selection, and recovers the true system ranking from fewer segments where a real quality gap exists. It does so by trading corpus-average fidelity for discriminative power, so we recommend pairing a TACTICS set, where costly human evaluation is best spent, with a random sample that estimates overall quality at the corpus proportion.

The core mechanism, taxonomy-aware constrained sampling over coverage, coherence, and fidelity, is not specific to machine translation. Any evaluation setting with a prescriptive taxonomy, long-tail skew over that taxonomy, and a fixed annotation budget could benefit. Future work includes validating TACTICS on public benchmarks and style guides to support reproducibility, extending it to other domains such as content moderation and clinical text, expanding human validation of taxonomy boundaries, and studying adaptive sampling budgets.

Limitations

Our primary evaluation uses a proprietary localization corpus that we cannot release, which limits direct reproducibility. We therefore plan a public companion study on open style guides and benchmarks. The evaluation is also confined to localization, and generalization to other domains such as news, biomedical, or legal text requires further validation. The taxonomy induction pipeline depends on the availability and quality of prescriptive style guides, which may not exist or may vary in granularity across locales, and the LLM-based classifier reaches only moderate inter-annotator agreement (Cohen's $\kappa = 0.51$), so misclassifications can affect coverage. By design, TACTICS trades fidelity to the corpus distribution for coverage and discriminative power, making it a deliberately biased estimator of corpus-average quality that should be paired with a random sample when an aggregate estimate is needed.

Acknowledgments

We thank Elijah Smith for automating and containerizing the sampling pipeline, enabling large-scale sampling jobs to be scheduled and run reliably with minimal manual intervention. This infrastructure was instrumental in carrying out the experiments presented in this paper.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Loïc Barrault et al. 2020. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of WMT*.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Arun Tejasvi Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evaluation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2022. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics (TACL)*, 9:1460–1474.

Leo Gao, Stella Biderman, Sid Black, et al. 2021. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, et al. 2022. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics (TACL)*, 10:522–538.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation (WMT)*.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. 2021. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. In *Proceedings of LREC*.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*.

Akash Kumar Mohankumar and Mitesh M. Khapra. 2022. Active evaluation: Efficient NLG evaluation with few pairwise comparisons. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

NLLB Team, Marta R. Costa-jussà, James Cross, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311318, USA. Association for Computational Linguistics.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2024. Efficient benchmarking of language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys*, 55(2).

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. *Preprint*, arXiv:2004.04696.

Yunfan Shao, Linyang Li, Zhaoye Fei, Hang Yan, Dahua Lin, and Xipeng Qiu. 2024. Balanced data sampling for language model training with clustering. *Preprint*, arXiv:2402.14526.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari S. Morcos. 2023. D4: Improving llm pretraining via document de-duplication and diversification. *arXiv preprint arXiv:2308.12284*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. Comparing test sets with item response theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. 2024. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W White, Longqi Yang, Reid Andersen, Georg Buscher, Dhruv Joshi, and Nagu Rangan. 2024. Tnt-llm: Text mining at scale with large language models. *Preprint*, arXiv:2403.12173.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data selection for language models via importance resampling. *Preprint*, arXiv:2302.03169.

Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Shangbin Feng, Zhenwen Liang, Zhihan Zhang, and Meng Jiang. 2024. Chain-of-layer: Iteratively prompting large language models for taxonomy induction from limited examples. *Preprint*, arXiv:2402.07386.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025. How to select datapoints for efficient human evaluation of NLG models? *Transactions of the Association for Computational Linguistics (TACL)*.

Config	Docs	Segs	Gini	Eff. D	Top-10	Cov_{30}
CF–IDF (wt.)	17	5,194	0.113	16.6	66.3%	37/76
CF–IDF (pure)	1,622	5,150	**0.491**	**959.6**	**8.0%**	**49/76**
CF–IDF (sat., wt.)	16	5,032	0.094	15.8	68.5%	37/76
CF–IDF (sat., pure)	**815**	**5,007**	0.441	551.3	9.2%	48/76

Table 5: Level-2 scoring ablation under fixed budget ($B = 5{,}000$) with Level-1 disabled. Cov_{30}: categories with $\geq$30 segments.

Appendix

A Level-2 Scoring Ablation

Level-2 scores each document by how much rare-category evidence it concentrates (Eq. (7)). Table 5 ablates the two design choices in that function, *saturation* and *aggregation*, under a fixed budget $B = 5{,}000$ (seed 42) with Level-1 disabled, so the differences reflect Level-2 scoring alone. The score combines two per-category signals. *Category frequency* $\text{CF}(c, j)$ is the fraction of segments in document j assigned to category c and is high when the document gives concentrated evidence for a phenomenon. *Inverse document frequency* $\text{IDF}(c)$ is large for rare categories, so the product $\text{CF} \cdot \text{IDF}$ favors documents rich in rare phenomena rather than documents that only repeat common ones. *Saturation* applies a concave transform CF_{sat} so that a category repeated many times within one document cannot dominate its score. *Aggregation* is either *pure*, where each category contributes once as $\text{CF}_{\text{sat}}(c, j)\,\text{IDF}(c)$ (Eq. (7)), or *weighted*, which additionally scales each term by the segment count $n(c, j) = |\{s \in J_j : c \in \ell(s)\}|$ and rewards repeated occurrences. All four variants share the same confidence weight $w_{\text{conf}} \in \{1.2, 1.0, 0.8\}$ (high, medium, low) and size factor w_{size} (Appendix B). The table shows a sharp split between the two aggregation choices. Weighted variants collapse the budget into 16–17 documents (Gini $\approx$ 0.1, Top-10 $\approx$ 66–69%) and cover only 37/76 categories. Pure variants instead spread selection over hundreds to more than 1,600 documents, raising effective diversity by more than an order of magnitude and lifting coverage to 48–49/76. **Saturated pure** trades a small amount of raw diversity for more balanced within-document evidence, selecting 815 documents at Gini 0.441 and $\text{Cov}_{30} = 48/76$, and is our default Level-2 configuration.

B Level-2 Weighting Details

The Level-2 document score (Eq. 7) applies two multiplicative weights that favor reliable, moderately sized documents. The confidence weight $w_{\text{conf}}(j)$ is 1.2 for high-confidence documents, 1.0 for medium, and 0.8 for low. For the size weight, let $n_j = |J_j|$ be the number of segments in document j, let Q_1 be the first quartile of document sizes, and let $n_{\max}$ be a segment cap, and define $r_j = \min\big(\frac{n_j - n_{\max}}{n_{\max}}, 1\big)$. The size weight is

$$w_{\text{size}}(j) = \begin{cases} 0.8 + 0.2 \cdot \frac{n_j}{Q_1}, & n_j \leq Q_1, \\ 1.0, & Q_1 < n_j \leq n_{\max}, \\ \max\big(0.5,\, 1 - 0.3\, r_j\big), & n_j > n_{\max}. \end{cases} \tag{8}$$

This down-weights both highly fragmented samples drawn from very small documents and budget domination by very large ones.

C Example Style Guide Snippet

Style Guide Snippet (Illustrative)

Currency Format. The currency symbol should be inserted after the number and preceded by a space. For Canadian dollars, there should be a space between the dollar sign and the currency symbol.
Examples: 50,99 / 4 500 EUR / 50 $ CA
As a general rule, use ISO currency codes.

Common currencies:

Country/Region	Currency name	Symbol	ISO code
EU	euro		EUR
Canada	Canadian dollar	CA $	CAD
USA	US dollar	$	USD

If currency symbols or ISO codes (e.g., $, , GBP, EUR) appear in the source text, they should be preserved and formatted according to locale-specific rules.

Date Format. Dates must follow the French day–month format rather than the US month–day format.

US English	French	Comments
7/14/2013	14/07/2013	N/A
July 14, 2013	14 juillet 2013	Extended

This snippet is a representative portion of a larger localization style guide. It illustrates the type of constraints used to induce the taxonomy in Section 3.2.1.

D Per-Level Pipeline Breakdown

Each level of TACTICS contributes a distinct sampling property. Table 6 reports per-stage statistics to illustrate how coverage, coherence, and fidelity are introduced incrementally. L1 constructs

the initial coverage-oriented seed set, L2 adds coherent document-level context, and L3 fills the remaining budget to match the full-corpus edit-rate statistics.

Level contributions. L1 selects the initial coverage-oriented seed set, ranging from roughly 1.5K to 2.9K segments across locales. L2 then adds coherent document-level context to reach approximately 3.5K segments. At this point coverage is effectively fixed, for example fr-FR reaches 98/105 categories and hi-IN reaches 79/80. L2 also introduces coherent multi-segment documents, with the number of selected documents varying by locale and document-size distribution.

L3 uses the remaining budget to reach $N = 5{,}000$ and align the sample with the full-corpus edit-rate profile. Since L3 only adds segments, coverage does not decrease from L2 to L3. The edit-rate correction can be substantial, with ja-JP decreasing from 75.7% to 61.5% and hi-IN from 74.2% to 53.1%. This happens because L1 and L2 over-sample edited segments, which are more category-diverse, while L3 rebalances the final sample by adding segments that restore distributional fidelity.

E Cross-Locale Generalization

We evaluate TACTICS across 22 locales spanning seven language families to assess whether the sampling objectives generalize beyond the main baseline-comparison directions. Table 7 reports intrinsic sampling quality for the test split at $N = 5{,}000$. Across locales, TACTICS maintains strong distributional fidelity, with mean Pearson correlation $r = 0.992$ and mean edit-rate deviation of 0.5 percentage points. L1 coverage varies with taxonomy size, where compact taxonomies such as en-AE (48 categories) achieve higher coverage at the ≥ 20 threshold, while larger taxonomies such as en-AU (177 categories) are harder to cover under the same fixed budget. L2 Gini averages 0.54 across locales, indicating consistent document-level diversity.

F Corpus-Mean Estimation Error

Table 8 reports the RMSE of the estimated system delta against the full-pool truth for random and TACTICS batches. TACTICS has higher RMSE in most cells, and the error is bias-dominated (bias^2/MSE near 100%). This is expected of a deliberately unrepresentative diagnostic batch, where concentrating rare and harder segments makes the sample a poor estimator of the corpus average while making it a stronger detector of the system ranking. This trade-off motivates pairing TACTICS with a separate random sample for aggregate estimation.

G System-Ranking Recovery

Table 9 reports how reliably a sampled batch recovers the correct system ranking, complementing the corpus-average estimation error of Appendix F. Using the full-pool audit verdict of Table 4 as reference, for each budget we draw many batches and run a paired t-test on the per-segment MQM deltas, reporting the median $|t|$ and the *recovery rate* (draws that are significant at $p < 0.05$ and match the verdict sign).

The pattern mirrors the estimation trade-off. Where the true effect is clear (en→hi, ja→en) both samplers recover it almost always. Where it is small (ja→zh, $\Delta = -0.086$) TACTICS recovers the verdict at 1600–2400 segments (100% vs. 62–93%), and on the null control (en→fr, $\Delta = +0.016$) it raises no false verdict (0% vs. random's 4–6%).

H Fidelity Diagnostics

Figure 4 shows representative Level-3 fidelity diagnostics for the hi-IN test split. The sampled set closely matches the full evaluation pool across taxonomy-category distribution, post-edit distribution, and clean-segment/lock-rate statistics.

I Taxonomy Generation Pipeline

The taxonomy generation pipeline consists of three LLM-based agents orchestrated in an iterative loop, **Extract** → **Audit** → **Refine** → **Audit** → …, running for up to 5 iterations or until the auditor marks the taxonomy as complete. Each agent can use a different LLM instance. Below we reproduce the full system prompts (DSPy signatures) for each agent.

I.1 Taxonomy Extraction Agent

The extraction agent operates in two stages, *Feature Extraction* followed by *Synthesis*.

I.1.1 Stage 1: Feature Extraction

```
Analyze a translation style guide and
identify SOURCE-TEXT TRIGGERS.

IGNORE THE FOLLOWING:
- Rules that only apply to the target
```

Locale	Dir	$\|C\|$	Stage	Seg.	Cov_{20}	Edit (%)	L2 Docs
fr-FR	en→fr	105	L1	2,882	–	–	–
			L1+L2	3,503	98/105	47.6	169
			L1+L2+L3	5,000	98/105	**45.7**	169
ja-JP	en→ja	76	L1	1,897	–	–	–
			L1+L2	3,524	75/76	75.7	302
			L1+L2+L3	5,000	75/76	**61.5**	302
ko-KR	en→ko	86	L1	2,255	–	–	–
			L1+L2	3,504	83/86	68.2	331
			L1+L2+L3	5,000	83/86	**55.4**	331
hi-IN	en→hi	80	L1	1,510	–	–	–
			L1+L2	3,506	79/80	74.2	483
			L1+L2+L3	5,000	79/80	**53.1**	483
en-US	ja→en	126	L1	2,913	–	–	–
			L1+L2	3,510	86/126	64.6	53
			L1+L2+L3	5,000	86/126	**60.4**	53
zh-CN	ja→zh	127	L1	2,530	–	–	–
			L1+L2	3,507	93/127	73.3	105
			L1+L2+L3	5,000	93/127	**68.0**	105

Table 6: Per-stage pipeline statistics at $N = 5{,}000$. Cov_{20} denotes the number of taxonomy categories with at least 20 sampled segments. Bold edit rates indicate the final L3-aligned edit rate, matching the full evaluation pool.

(e.g., 'no double spaces in target')
- Team-specific or project-specific instructions (e.g., "Note for X projects", "Exception for Y projects", "Specific instructions for Z")
- Rules from other organizations/clients (only extract rules for the specified organization_name)

FOCUS ON:
- Universal source attributes that force a specific translation decision
- Linguistic markers (e.g., Second-person pronouns)
- Content types (e.g., UI labels, Legal names)
- Technical constraints (e.g., ICU placeholders)
- General rules applicable across all projects for this organization

Inputs:
markdown_content: The style guide content
organization_name: Organization/Client name of the style guide

Output:
source_features: Structured list of features including categories, names, rules, and pattern hints.

I.1.2 Stage 2: Taxonomy Synthesis

Consolidates raw features into a source-centric hierarchical TaxonomyStructure.

Logic:
1. PERSPECTIVE SHIFT: Transform target-evaluation rules into source-text triggers. (e.g., Change 'Check for natural translation' to 'Idiomatic/Complex Source').
2. BUCKETING: Group into 'Deterministic' (Regex-ready) vs. 'Semantic' (LLM-required).
3. LABELING: Ensure Semantic labels describe what to LOOK FOR in the Source-Text.
4. DEDUPLICATION: Merge identical source-triggers from different guide sections.

IGNORE DURING SYNTHESIS:
- Team-specific or project-specific instructions
- Rules that only apply to specific teams, projects, or contexts

Input:
raw_features: Aggregated features from guide extraction

Output:
organized_taxonomy: The final nested hierarchy (Deterministic/Semantic) ready for the Classifier.

I.2 Taxonomy Auditor Agent

Audit taxonomy extraction for structural completeness.

SCOPE CONSTRAINTS:
- Be pragmatic, not exhaustive.
- Focus on STRUCTURAL completeness, not exhaustive examples or word lists.
- A category is complete if it captures the CONCEPT and RULE, not every instance.
- Do NOT request specific word pairs, complete tables, or dictionary references.
- Maximum 3–5 high-priority structural gaps per audit.

Locale	Family	$\|C\|$	Cov_{20}	Gini	r / Edit Δ
es-MX	Romance	76	69/76	0.441	0.980 / 0.1
zh-TW	Sinitic	87	48/87	0.490	0.987 / 0.0
en-AE	Germanic	48	35/48	0.576	0.991 / 1.7
en-AU	Germanic	177	36/177	0.626	0.997 / 0.3
en-CA	Germanic	82	38/82	0.571	0.995 / 1.5
en-GB	Germanic	95	32/95	0.747	0.951 / 0.0
en-IE	Germanic	128	34/128	0.610	0.994 / 1.9
en-IN	Germanic	112	26/112	0.584	0.999 / 0.6
en-SG	Germanic	72	40/72	0.537	0.996 / 1.5
en-US	Germanic	126	34/126	0.491	0.997 / 0.4
en-ZA	Germanic	136	34/136	0.605	0.992 / 0.8
nl-BE	Germanic	89	44/89	0.517	0.994 / 0.2
nl-NL	Germanic	100	47/100	0.525	0.996 / 0.0
hi-IN	Indic	80	51/80	0.468	0.996 / 0.0
ja-JP	Japonic	76	48/76	0.526	0.992 / 0.0
ko-KR	Koreanic	86	42/86	0.500	0.997 / 0.0
fr-BE	Romance	77	45/77	0.511	0.994 / 0.0
fr-FR	Romance	105	44/105	0.541	0.996 / 0.3
it-IT	Romance	96	45/96	0.510	0.995 / 0.2
pt-BR	Romance	150	42/150	0.474	0.997 / 0.1
pt-PT	Romance	92	44/92	0.516	0.991 / 0.2
pl-PL	Slavic	87	37/87	0.482	0.992 / 0.0
Mean	–	**99**	**44.0%**	**0.543**	**0.992 / 0.5**

Table 7: Cross-locale intrinsic sampling quality at $N = 5{,}000$. Cov_{20} reports categories with at least 20 samples. Edit Δ is the absolute edit-rate deviation from the full evaluation pool, in percentage points.

```
- Mark is_complete=True if >80% of major
rule categories are represented.

IGNORE DURING AUDIT:
- Team-specific or project-specific
instructions
- Rules from other organizations/clients
- DO NOT flag team/project-specific
content as "missing" from the taxonomy

MARK COMPLETE (is_complete=True) IF:
- All major GENERAL category TYPES are
represented
- Patterns capture general structure
adequately
- Descriptions explain the rule/concept
sufficiently
- Source document lacks extractable
rules (intro/purpose text only)
- You cannot meaningfully validate
against source content

MARK INCOMPLETE (is_complete=False)
ONLY IF:
- Entire GENERAL rule categories are
clearly missing
- Critical patterns have no
representation at all
- AND source document contains actual
rules to validate against

EDGE CASE - AUTOMATIC PASS:
If source document contains only
introductory/purpose statements
with no actual rules, patterns, or
guidelines to extract, you MUST return
is_complete=True. Do not flag items
```

Locale	Batch	$\text{RMSE}_{\text{rand}}$	RMSE_{TAC}	DEFF	bias^2/MSE
en→fr	200	0.0763	0.1010	1.75	93%
	400	0.0532	0.0627	1.39	79%
	800	0.0355	0.0138	0.15	88%
	1600	0.0218	0.0272	1.55	85%
	2400	0.0144	0.0280	3.79	87%
en→hi	200	0.0747	0.0860	1.32	88%
	400	0.0513	0.2269	19.61	99%
	800	0.0346	0.0928	7.20	97%
	1600	0.0212	0.0750	12.54	98%
	2400	0.0148	0.0679	20.93	99%
ja→en	274	0.1846	0.6644	12.95	100%
	440	0.1428	0.5706	15.97	100%
	800	0.0947	0.3493	13.60	100%
	1600	0.0556	0.4533	66.50	100%
	2400	0.0320	0.1628	25.92	100%
ja→zh	388	0.0809	0.1052	1.69	100%
	529	0.0667	0.0657	0.97	100%
	832	0.0492	0.0082	0.03	100%
	1600	0.0292	0.0874	8.97	100%
	2400	0.0162	0.0375	5.32	100%

Table 8: Corpus-mean estimation error. RMSE of the estimated system delta against the full-pool truth, for random and TACTICS batches. DEFF $= (\text{RMSE}_{\text{TAC}}/\text{RMSE}_{\text{rand}})^2$, and values below 1 favor TACTICS. TACTICS has higher RMSE in most cells, and the error is bias-dominated (bias^2/MSE near 100%), as expected of a deliberately unrepresentative diagnostic batch, where concentrating hard segments makes the batch a poor estimator of the corpus average but a stronger detector of the system ranking (Table 4, Figure 3).

```
as "missing" when the source itself
contains nothing to extract.

LOGICAL CONSTRAINT:
"Cannot assess completeness" +
"is_complete=False" is INVALID. If
validation is impossible due to
source content → is_complete=True
automatically. Do NOT waste iterations
on unvalidatable content.

DO NOT REQUEST:
- Exhaustive word lists or substitution
tables
- Specific external reference links or
dictionary citations
- Every example from source document
- Granular details requiring multiple
extraction passes
- Items not present in the source
document
- Team/project-specific rules or
exceptions

Inputs:
markdown_content: The original style
guide text
organization_name: Organization/Client
name
current_taxonomy: Current extracted
TaxonomyStructure

Outputs:
missing_rules: List of specific
```

Locale	Batch	$\lvert t\rvert_{\text{rand}}$	$\lvert t\rvert_{\text{TAC}}$	ratio	Rec_{rand}	Rec_{TAC}
en→fr	200	0.75	1.60	2.14	4%	0%
	400	0.74	1.11	1.50	5%	0%
	800	0.70	0.08	0.12	6%	0%
	1600	0.67	0.22	0.33	4%	0%
	2400	0.76	0.45	0.59	4%	0%
en→hi	200	4.32	5.15	1.19	95%	100%
	400	5.89	9.60	1.63	100%	100%
	800	8.25	10.70	1.30	100%	100%
	1600	11.55	15.11	1.31	100%	100%
	2400	14.13	17.31	1.22	100%	100%
ja→en	274	6.11	4.45	0.73	100%	100%
	440	7.64	6.64	0.87	100%	100%
	800	10.08	7.08	0.70	100%	100%
	1600	14.15	15.18	1.07	100%	100%
	2400	17.23	17.76	1.03	100%	100%
ja→zh	388	1.10	1.15	1.04	11%	0%
	529	1.27	1.21	0.96	16%	0%
	832	1.53	1.14	0.75	26%	0%
	1600	2.15	3.45	1.60	62%	100%
	2400	2.51	3.38	1.35	93%	100%

Table 9: System-ranking recovery on the audited pool. For each direction and budget, each sampler draws batches and we run a paired t-test on the per-segment MQM deltas of the two systems in Table 4. The $|t|$ columns give the median over draws, ratio $= |t|_{\text{TAC}}/|t|_{\text{rand}}$ (> 1 favors TACTICS), and Rec is the fraction of draws that are significant ($p < 0.05$) *and* match the full-pool verdict sign. en→fr is a null control, where lower Rec is better.

```
rules/structural gaps identified
is_complete: Boolean - True if >80%
covered, False only if major structural
categories are missing
```

I.3 Taxonomy Refiner Agent

```
Update the existing category taxonomy
structure based on the feedback from the
auditor agent.

INSTRUCTIONS:
1. Review the 'current_taxonomy'.
2. Incorporate the missing rules
identified in 'audit_report'.
3. Return the FULL, updated JSON
structure.
4. Maintain the existing Group →
Subgroup Hierarchy.

IGNORE DURING REFINEMENT:
- Team-specific or project-specific
instructions
- Rules that only apply to specific
teams, projects, or contexts
- DO NOT add project-specific rules
when incorporating missing rules from
audit_report

Inputs:
current_taxonomy: The existing
TaxonomyStructure JSON
audit_report: Specific missing rules or
errors found by the auditor

Output:
revised_taxonomy: The updated version of
the taxonomy
```

I.4 Orchestration

The three agents are composed in a LangGraph workflow with conditional routing:

1. **Extraction node** (iteration 0): Runs the two-stage Taxonomy Extraction Agent (Feature Extraction → Synthesis) to produce an initial `TaxonomyStructure`.

2. **Audit node**: Runs the Taxonomy Auditor Agent against each style guide document. If >80% of major rule categories are covered, marks `is_complete=True`.

3. **Routing decision**: If complete or `iterations > max_iterations` (default 5), the workflow terminates. Otherwise, it proceeds to refinement.

4. **Refinement node** (iterations ≥ 1): Runs the Taxonomy Refiner Agent to incorporate the auditor's missing rules into the existing taxonomy, then loops back to the audit node.

Each stage supports a separate LLM instance (`extraction_lm`, `audit_lm`, `refinement_lm`), enabling heterogeneous model configurations (e.g., a stronger model for extraction, a faster model for auditing). All modules use DSPy's `ChainOfThought` wrapper, which elicits step-by-step reasoning before producing the structured output.

J Taxonomy Validation

We assess taxonomy consistency via cross-family agreement between two independent classifiers from different model families, Kimi K2.5 and Claude Haiku 4.5. Agreement improves after refinement (Cohen's κ from 0.27 to 0.51 on a 100-segment pilot), indicating clearer category boundaries. A taxonomy spot checker additionally samples induced categories and verifies that each is grounded in a prescriptive style-guide rule, clearly defined, and distinguishable from its neighbors, and Table 10 shows this mapping for fr-FR.

Taxonomy sizes range from 48 to 177 categories across locales, reflecting variation in style-guide complexity. English variants (e.g., en-AU,

en-IE, en-ZA) are largest (128–177), East Asian locales (ja-JP, ko-KR, zh-TW) most compact (76–87), and Romance locales fall in between (77–150).

K Example Taxonomy for French (fr-FR)

Figure 5 shows a condensed view of the taxonomy induced for French (fr-FR), separating *Deterministic* categories (detectable with regex-style or rule-based patterns) from *Semantic* categories (requiring context-aware LLM reasoning). The full taxonomy contains 27 deterministic categories (72 sub-categories) and 17 semantic categories (42 sub-categories).

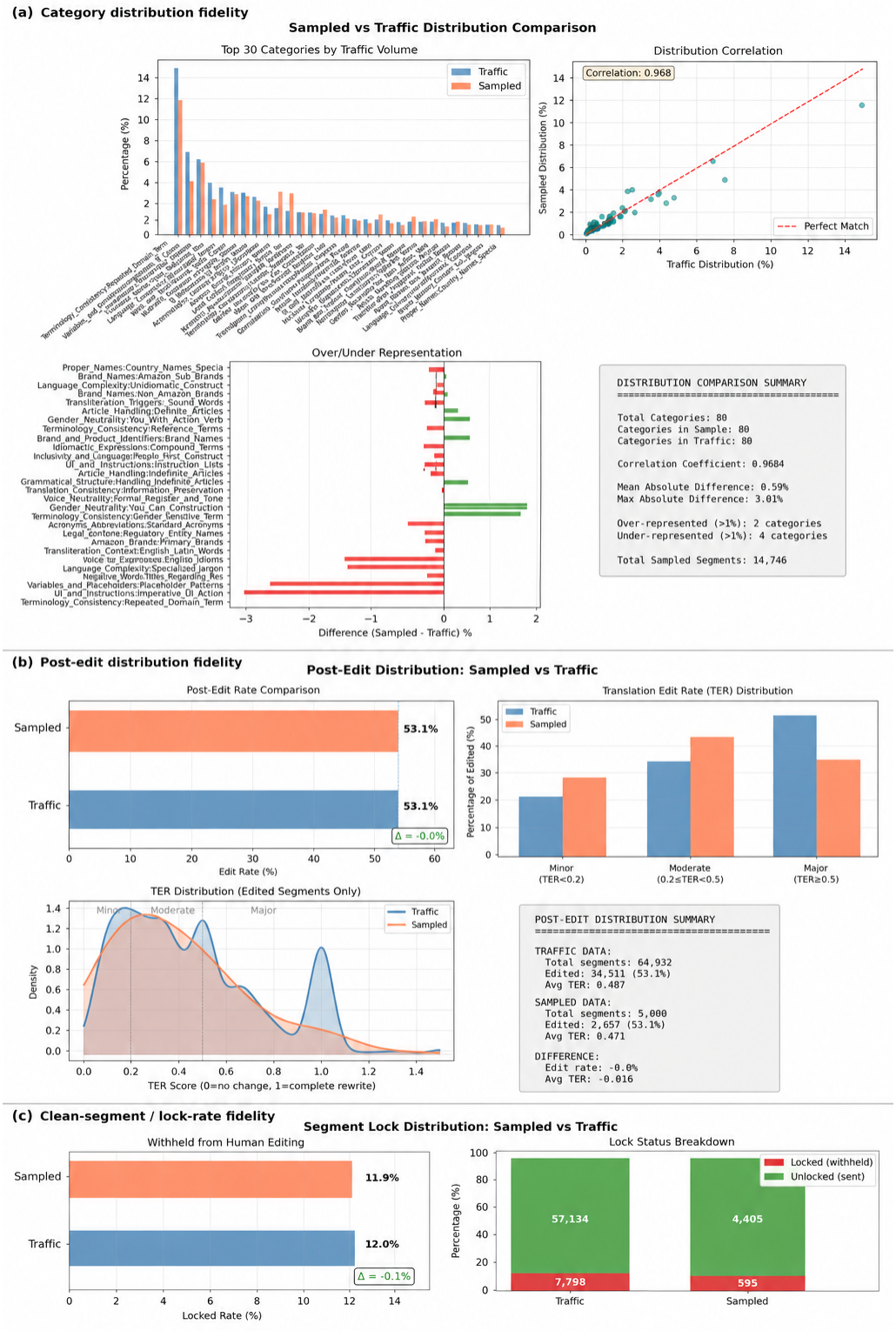


Figure 4: Fidelity diagnostics for the hi-IN test split. (a) Category distribution fidelity compares taxonomy-category frequencies between sampled data and the full evaluation pool. (b) Post-edit distribution fidelity compares edit rate and TER severity distributions. (c) Clean-segment / lock-rate fidelity compares the percentage of segments withheld from human editing. These diagnostics illustrate how Level 3 aligns the final sampled set with the full pool.

Style Guide Rule	Taxonomy Category	Type	Spot-Check Example
	Fluency – Grammar & Spelling		
No plural on acronyms (PDFs → PDF)	Acronyms_Abbrev. / Pluralized_Acronyms	Det.	`\b[A-Z]{2,}s\b` catches "PDFs", "DVDs"
Country abbrev. (UK → Royaume-Uni)	Acronyms_Abbrev. / Country_Abbreviations	Det.	`UK|U.K.` → "Royaume-Uni"
Dotted abbrev. (U.S. → États-Unis)	Acronyms_Abbrev. / Dotted_Abbreviations	Det.	`U.S.` matched by `[A-Z]\.([A-Z]\.)+`
Tech spelling (Wi-Fi, e-mail)	Technology_Terms / Tech_Spelling	Det.	"WiFi" → "Wi-Fi"; "email" → "e-mail"
Web capitalization	Technology_Terms / Web_Capitalization	Det.	"web" → "Web"
Proper names invariable (iPads)	Proper_Names / Apple_Products	Det.	"iPads" → "les iPad" (no plural)
	Fluency – Typography		
Ampersand → "et"	Punctuation_Typo. / Ampersand	Det.	"Shoes & Bags" → "Chaussures et sacs"
Hashtag → n^{o}	Punctuation_Typo. / Hashtag_Number	Det.	"#1" → "n^{o} 1"
No exclamation in UI	Punctuation_Typo. / Exclamation_in_UI	Det.	"Welcome!" → remove "!" in UI strings
	Locale Convention		
Currency format (symbol after number)	Currency_Pricing / Currency_Symbols	Det.	"$19.99" → "19,99 $" or "19,99 USD"
Date format (MM/DD → DD/MM)	Date_Formats / US_Numeric_Date	Det.	"7/14/2013" → "14/07/2013"
Written date reorder	Date_Formats / Written_Month_Date	Det.	"July 14, 2013" → "14 juillet 2013"
12h → 24h time	Time_Formats / Twelve_Hour_Time	Det.	"8:30 pm" → "20 h 30"
Thousands sep. (comma → space)	Number_Formats / Thousands_Separator	Det.	"2,935,000" → "2 935 000"
Decimal sep. (period → comma)	Number_Formats / Decimal_Numbers	Det.	"11.5" → "11,5"
Percent spacing	Number_Formats / Percentage_Values	Det.	"10%" → "10 %"
Discount format	Number_Formats / Discount_Expressions	Det.	"40% off" → "- 40 %"
Fahrenheit → Celsius	Measurements / Temperature_Fahrenheit	Det.	"72°F" → "22 °C"
Imperial → metric	Measurements / Imperial_Units	Det.	"12 inches" → "30,48 cm"
Clothing sizes	Measurements / Clothing_Sizes	Det.	"Small" → "Taille S"
	Style		
Polite "vous" form	Voice_Tone / Second_Person_Address	Sem.	"you" → always "vous" (formal)
Imperative in instructions	Instruction_Style / Imperative	Sem.	"Print label" → "Imprimez l'étiquette"
Infinitive for buttons	Instruction_Style / Button_Link_Labels	Sem.	"Find a package" → "Trouver un paquet"
Gender-neutral roles	Gender_Neutrality / Role_Terms	Sem.	"specialist" → "Spécialiste" (not "Expert(e)")
Predicate adj. with "you"	Gender_Neutrality / Predicate_Adj.	Sem.	"you're ready" → rephrase to avoid gendered agreement
Long sentences	Language_Complexity / Long_Complex	Sem.	Break up multi-clause sentences for readability

Table 10: Spot-check mapping between French (fr-FR) localization style guide rules and taxonomy categories. **Type**: Det. = Deterministic (regex-based), Sem. = Semantic (LLM-based). Each row shows a style guide rule, its corresponding taxonomy category, detection type, and a concrete example of how the rule is operationalized.

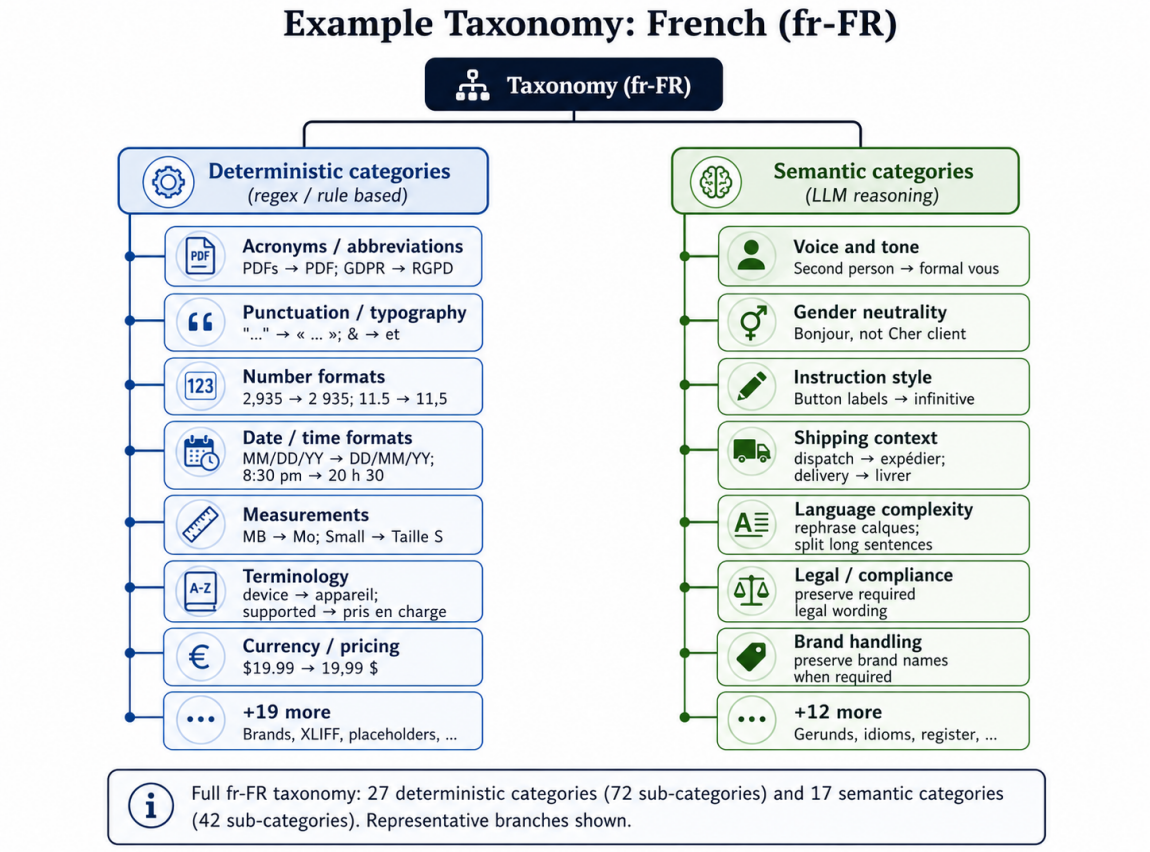


Figure 5: Condensed example taxonomy for French (fr-FR). Deterministic categories capture rule-matchable phenomena such as punctuation, number formats, terminology, and currency conventions. Semantic categories capture context-dependent phenomena such as tone, gender neutrality, instruction style, and shipping context.