COMPASS: Proactive Writing Guidance from Text Classifier Internals

Anonymous ACL submission

Abstract

When a text classifier misclassifies, the author typically receives only the label, with no guidance on how to revise. Existing explainability methods such as LIME, SHAP, and counterfactuals tell ML practitioners *why* a prediction was made but do not tell authors *how* to write. We introduce COMPASS, a framework that extracts vocabulary stable across architecturally diverse text classifiers and uses an LLM to translate this consensus into actionable per-class writing tips. On 7 benchmarks spanning 5 domains, we evaluate COMPASS by asking an LLM to either rewrite existing documents or generate new ones following the tips. COMPASS guidance improves Logistic Regression accuracy on 6 of 7 datasets, with large gains on aviation safety narratives; on specialised domains where generation from scratch underperforms, rewriting recovers. A zero-shot classifier trained on no COMPASS data confirms the gains on all 7 datasets, ruling out circular evaluation. An evaluation with 9 industrial practitioners finds novices rate the guidance higher than seniors with Hedges' g=1.47; given the small sample, we offer onboarding as a candidate application warranting further study.

1 Introduction

Automated text classifiers are gatekeepers in many high-stakes settings (email routing, resume screening, grant proposal triage, content moderation): when a document is misclassified, authors typically receive only the label—not *why* it was misclassified, and certainly not *how to fix it*. This is particularly acute in industrial and clinical domains, where time-pressured staff write free-text reports—maintenance work orders with abbreviations and non-standard grammar (Brundage et al., 2021), pilot safety reports of variable quality (NASA, 2024), clinical adverse-event reports that vary in specificity (U.S. Food and Drug Administration, 2024). Report quality directly affects downstream classification, yet receives no feedback at writing time, pushing the burden onto costly post-hoc cleaning (Hodkiewicz and Ho, 2016).

The vast majority of NLP research—including Technical Language Processing (Brundage et al., 2021)—focuses on processing *existing* text rather than improving text *before* it is written. Explainable AI (XAI) methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) offer post-hoc descriptions of model behaviour but are *descriptive*, not *prescriptive*; algorithmic recourse (Ustun et al., 2019; Karimi et al., 2022) formalises actionable change recommendations but almost exclusively for tabular data. Extending recourse to the text domain is fundamentally harder because the action space of natural language is generative, high-dimensional, and must remain coherent.

We propose COMPASS—**C**lassifier-**O**riented **M**ethod for **P**rescriptive **A**uthoring **S**upport & **S**caffolding—a framework that addresses this gap by automatically generating *proactive* writing guidance grounded in classifier internals.[1] Rather than explaining *why* a document was classified a certain way, COMPASS tells the author *how to write* so that their document will be correctly classified.

To assess COMPASS empirically, we evaluate on 7 datasets spanning 5 domains (general NLP, finance, medical, industrial maintenance, safety reporting) using 6 classifier architectures (classical ML and Transformers). We measure utility under two operating modes: an LLM rewrites existing test documents (*Rewritten*) or generates new ones from scratch (*Guided*), and we report per-classifier accuracy across all six architectures. COMPASS guidance improves accuracy on 6 of 7 datasets, with Logistic Regression (LR) gaining up to +27.9pt on ASRS; the lone regression is the 5-word MaintIE corpus, and a stratified human evaluation with

[1] Implementation is available at <URL upon acceptance>.

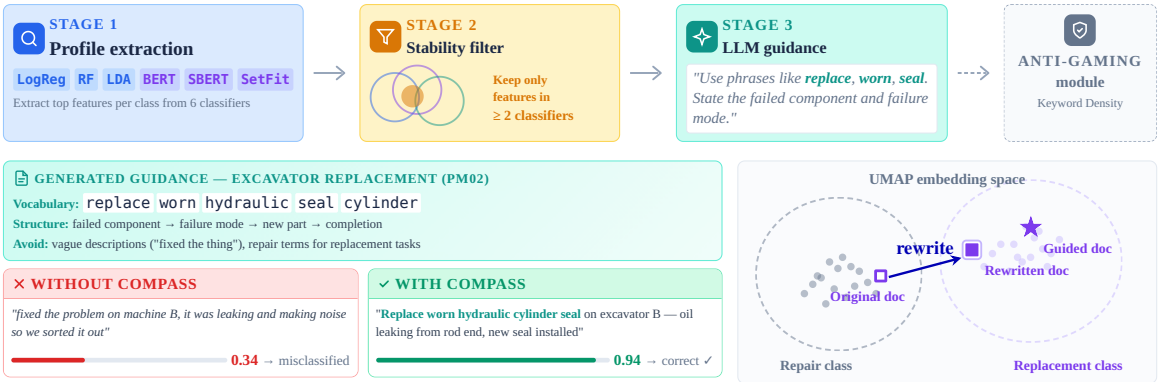


Figure 1: COMPASS pipeline overview (schematic; "Repair"/"Replacement" are illustrative). The dashed arrow and dashed border on the Anti-Gaming module indicate a post-hoc audit rather than a sequential pipeline stage. The embedding inset sketches the intuition: a misclassified Original moves into the correct region after Rewriting, while a Guided document is placed inside the correct cluster. Real UMAP projections in Figure 4.

industrial practitioners indicates the largest gains accrue to less-experienced writers.

Our contributions are:

1. **Task definition:** We formalise *classifier-guided writing guidance generation*: a prescriptive (vs. descriptive) feedback task complementing existing XAI methods and tabular recourse.
2. **Method:** A three-stage pipeline (Figure 1) that extracts feature profiles specific to each classifier architecture, filters them by stability across classifiers ($\geq$ 2 classifiers), and translates the result into actionable guidance via an LLM—guarded by an Anti-Gaming module (§5.4).
3. **Multi-domain evaluation:** Positive gains on 6 of 7 datasets across 5 domains and 6 classifier architectures, with LR improving by up to +27.9pt on ASRS aviation safety narratives.
4. **Human validation:** A stratified evaluation (n=9) where novices score the guidance higher than seniors (20.0/30 vs. 14.7/30; Hedges' g=1.47, p=0.062).

2 Related Work

COMPASS sits at the intersection of three areas: explaining classifier decisions, generating text under classifier constraints, and assisting authors with writing. None of these prior lines produce human-readable writing guidance grounded in classifier internals and aggregated over multiple classifiers.

2.1 Explaining Classifier Decisions

Feature attribution methods—LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), Anchors (Ribeiro et al., 2018), and rationale extraction (DeYoung et al., 2020)—highlight features driving a prediction, but remain *descriptive*: they do not synthesise findings into actionable writing guidance. Counterfactual methods generate minimal edits that flip a prediction: Polyjuice (Wu et al., 2021) fine-tunes GPT-2 for diverse edits, MiCE (Ross et al., 2021) optimises for minimal contrastive edits. Critically, these methods produce *per-document* edits conditioned on a source text; COMPASS instead produces a *class-level* guidance artefact—a set of writing tips for a target class—usable to author new documents from scratch, with no source-text dependency. Algorithmic recourse (Ustun et al., 2019; Karimi et al., 2021, 2022) formalises actionable change recommendations, distinguishing "why P and not Q?" from "what should I do?"—but the literature focuses on tabular features and has not been extended to text, where the action space is generative and high-dimensional. AReS (Rawal and Lakkaraju, 2020) aggregates individual recourses into global summaries; we adapt this aggregation idea for text via stability across classifiers.

2.2 Generating Text Under Classifier Constraints

Controllable text generation methods solve the inverse problem of producing text that satisfies a classifier: PPLM (Dathathri et al., 2020) steers generation via attribute classifiers, FUDGE (Yang and Klein, 2021) uses future discriminators, CTRL (Keskar et al., 2019) conditions on control codes. Text style transfer (Jin et al., 2022) modifies attributes while preserving content. Strategic classification (Hardt et al., 2016; Miller et al., 2020)

formalises agents who manipulate features to obtain favourable decisions. These methods produce *generated text*, not human-readable guidance for authors. COMPASS differs in that the target "style" is defined by classifier decision boundaries and the output is natural-language writing tips that remain robust across classifiers.

2.3 AI-Assisted Writing and Technical Language Processing

Writing assistance systems give feedback on general quality: SWIF^{2}T (Chamoun et al., 2024) for scientific papers, AcaWriter (Shibani et al., 2019) for rhetorical structure, CoAuthor (Lee et al., 2022) for human-AI collaborative writing. None target *classifiability*—alignment with a specific deployed classifier. Technical Language Processing (TLP) adapts NLP to specialised industrial text such as maintenance work orders (Brundage et al., 2021), supported by resources like MaintNet (Akhbardeh et al., 2020) and MaintIE (Bikaun et al., 2024). The TLP community has focused on cleaning and classifying *existing* text; we apply prescriptive guidance *before* writing, complementing this line of work.

3 Problem Formulation

3.1 Setup

Let $D = \{d_1, \ldots, d_n\}$ be a document corpus with labels $\{y_1, \ldots, y_n\} \in Y$, and $\{f_1, \ldots, f_K\}$ be K classifiers trained on D. Given a target class $y^* \in Y$, our goal is to generate natural-language writing guidance $G(y^*)$ such that a new document d' written following $G(y^*)$ satisfies:

$$P(f_k(d') = y^*) \gg P(f_k(d) = y^*) \quad \forall k \qquad (1)$$

where d is a baseline document on the same topic written without guidance. We operationalise this as classifier accuracy on documents generated under $G(y^*)$ relative to baseline documents; Tables 2 and 4 report this measurement.

3.2 Desiderata

The guidance $G(y^*)$ must satisfy:

- **Effectiveness**: d' is classified correctly with high confidence.
- **Actionability**: a human author can follow it.
- **Generality**: effective across multiple classifiers.
- **Improvement**: promotes genuine quality, not *gaming* (inflating classifier confidence through shortcuts such as keyword stuffing rather than authentic writing).

4 Method: COMPASS

4.1 Stage 1: Classifier Profile Extraction

For each classifier f_k and class y, we extract a *Classifier Profile* $\mathrm{CP}_k(y)$ of the top-N most important features using an architecture-tailored method: signed coefficients (logistic regression, LogReg), discriminative scores $s_i^{(y)} = \mathrm{FI}_i \cdot (\bar{x}_i^{(y)} - \bar{x}_i^{(\neg y)})$ (Random Forest (Breiman, 2001), RF), topic-word distributions weighted by SVC coefficients (Latent Dirichlet Allocation (Blei et al., 2003) + LinearSVC (Fan et al., 2008), LDA+SVC), occlusion (DistilBERT (Sanh et al., 2019)), and word-ablation perturbation (Sentence-BERT (Reimers and Gurevych, 2019), SBERT; SetFit (Tunstall et al., 2022)).

Profiles are then preprocessed to remove stopwords, subword fragments, email/URL patterns, and other artefacts (Appendix A).

4.2 Stage 2: Cross-Classifier Stability

Let $\mathrm{Top}_N(f_k, y)$ be the set of top-N features from classifier f_k for class y (we use N=15). We retain only features appearing in ≥ 2 classifiers; we call this property *cross-classifier stability*:

$$\mathrm{Stable}(y) = \{t : |\{k : t \in \mathrm{Top}_N(f_k, y)\}| \geq 2\} \qquad (2)$$

This set defines the stable profile used by Stage 3 and the Anti-Gaming filter (§6.3). To quantify cross-classifier agreement we report the pairwise Jaccard

$$J_{k,k'}(y) = \frac{|\mathrm{Top}_N(f_k, y) \cap \mathrm{Top}_N(f_{k'}, y)|}{|\mathrm{Top}_N(f_k, y) \cup \mathrm{Top}_N(f_{k'}, y)|}, \qquad (3)$$

which serves as a diagnostic of profile stability across classifier pairs. Features stable across architecturally different classifiers are more likely to reflect genuine domain characteristics than *model-specific artefacts*—features whose appearance in the profile reflects one classifier's architecture rather than the class itself. For Financial Sentiment *bearish*, the stable feature set is {*china, coronavirus, cut, cuts, down, downgraded, fall, falls, loss, lower, misses, oil, premarket, revenue, shares, tumbled*}; cleaned profile examples for additional classes appear in Appendix B.

4.3 Stage 3: LLM-Mediated Guidance Generation

An LLM converts the stable profile into the user-facing deliverable: a guidance document of 5–7

1. **Focus on Economic Indicators**: Mention how China's economic performance is affecting global markets, especially in the context of recent cutbacks and revenue misses. This helps readers understand the broader economic implications and why markets might be bearish.
2. **Highlight Market Performance**: Describe how oil prices have fallen sharply, leading to significant drops in share values. Use concrete examples like specific companies or sectors that have been particularly affected.
3. **Discuss Company Performance**: Explain how companies are missing their revenue targets, leading to downgrades and share price falls. Provide specific company names and the reasons behind their performance issues.

Figure 2: Guidance for Financial-Sentiment *bearish*.

Dataset	Domain	$\lvert C \rvert$	Train	Test
20 Newsgroups	NLP	5	2,914	1,940
AG News	NLP	4	10,000	7,600
Financial	Finance	3	9,543	2,388
PubMed RCT	Medical	5	5,000	1,000
ASRS	Safety	5	26,258	6,565
FDA MAUDE	Safety	2	3,999	1,000
MaintIE	Maint.	3	848	212

Table 1: Dataset overview.

actionable writing tips per class, framed as advice without mentioning classifiers or feature weights. Figure 2 shows the first three of six writing tips that COMPASS generated for this *bearish* class. The Stage 3 system prompt is fixed across all classes, datasets, and downstream PromptConfig modes (Appendix C); only the user message—carrying the class name and stable feature set—varies.

5 Experimental Setup

5.1 Datasets

We evaluate on seven datasets across five domains; see Table 1. The general-NLP datasets are 20 Newsgroups (Lang, 1995) and AG News (Zhang et al., 2015); the finance dataset is Financial Sentiment[2]; the medical dataset is PubMed RCT (Dernoncourt and Lee, 2017). The safety reporting datasets, NASA ASRS (NASA, 2024) and FDA MAUDE (U.S. Food and Drug Administration, 2024), comprise pilot-filed aviation safety narratives and medical device adverse event reports, respectively; both exemplify the time-pressured industrial-reporting scenario motivating this work (§1). The industrial maintenance dataset is MaintIE (Bikaun et al., 2024). AG News is subsampled from 120k to 10k, and PubMed RCT to ∼1k/class.

[2] https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment

5.2 Classifiers

We use six classifiers spanning classical ML and modern Transformers: TF-IDF+LogReg, TF-IDF+Random Forest, LDA+LinearSVC, DistilBERT (fine-tuned), Sentence Transformer (all-MiniLM-L6-v2)+LogReg, and SetFit (few-shot). Per-dataset baseline accuracies are reported in Table 2; the same models are used across all conditions for evaluation.

5.3 Document Generation and Evaluation Protocol

Two operating modes. COMPASS's guidance is designed to be read by human authors and used in two complementary ways: *prospectively*, writing a new document from scratch following the guidance; or *retrospectively*, revising an existing document while preserving factual content. For scalable evaluation we substitute an LLM for the human author, yielding three classification conditions:

- **Original**: the full held-out test set classified as-is, used as baseline.
- **Guided**: LLM-generated documents following the guidance, modelling prospective use.
- **Rewritten**: LLM-rewritten real test documents, modelling retrospective use; we report ROUGE-L (Lin, 2004) against the source as a surface proxy for content preservation.

Figure 3 shows the three conditions on the running Financial Sentiment *bearish* example, with the guidance from Figure 2 as the input to Guided/Rewritten generation.

PromptConfig. The Guided/Rewritten LLM is given the guidance (§4.3) together with an automatically derived *PromptConfig*: (1) **length specification** from training text word-count statistics; (2) **domain context** generated by an LLM from the Classifier Profile features, e.g., *down, falls, downgraded* → "financial news expressing market sentiment"; (3) **style instructions** generated by an LLM from sampled training texts, e.g., "use ALL UPPERCASE, telegraphic style"; (4) **few-shot examples** sampled from training data. An optional **dataset description** can be added manually but is omitted by default, as our ablation shows it can be counterproductive (§E.1). The complete LLM prompt appears in Appendix D.

```
$TAP: Barclays cuts to Equal Weight
```

(a) Original.

```
Barclays cuts to Equal Weight reflect China's
weakening economy and falling oil prices,
hitting companies like ExxonMobil and Chevron
hard. ...
```

(b) Rewritten.

```
China's economic slowdown and recent cutbacks
are affecting global markets, particularly in
energy and manufacturing. ...
```

(c) Guided.

Figure 3: Original/Rewritten/Guided documents on Financial-Sentiment *bearish*.

Experiment parameters. All text—training data and generated documents alike—is lowercased before classification to prevent case-mismatch artefacts between uppercase maintenance/clinical reports and mixed-case LLM output. For Guided we generate 30 documents per class; for Rewritten we take the first 30 available test documents of each class, fewer if a class has under 30 test documents, using two open-source LLMs—Mistral 7B (Jiang et al., 2023) and Qwen 2.5 7B (Qwen Team, 2024)—served via vLLM (Kwon et al., 2023), and we evaluate with all six classifiers. The three optional PromptConfig components style, few-shot, and description are ablated as all $2^3{=}8$ combinations (Appendix E.1).

5.4 Anti-Gaming Module

The Anti-Gaming Module sits at the end of the COMPASS pipeline (Figure 1) but operates as a post-hoc audit rather than a generation gate: it flags candidate outputs whose vocabulary distribution looks suspicious without rejecting or rewriting them. A post-hoc check audits generated documents for *keyword cramming*—the text-domain analog of strategic classification (Hardt et al., 2016), where an agent trivially inflates classifier confidence by repeating target-class keywords. We define the keyword density of document d for target class y as

$$\mathrm{KD}(d, y) = \frac{|\{w \in d : w \in \mathrm{Stable}(y)\}|}{|d|}, \quad (4)$$

where $\mathrm{Stable}(y)$ is the cross-classifier stable-feature set (Eq. 2) and $|d|$ is document length in tokens. The **Keyword Density Guard** flags any generated document with $\mathrm{KD}(d, y) > 1.5 \cdot \kappa_{95}(y)$, where $\kappa_{95}(y)$ is the 95th percentile of training-document keyword densities for class y. The flag rate is reported as a post-hoc audit on Guided/Rewritten outputs (§6.3).

6 Results

6.1 Guidance Evaluation

Table 2 reports per-classifier baseline accuracy and improvements (ΔG, ΔR), averaged over Mistral 7B and Qwen 2.5 7B; the document-generation prompt configuration and its ablation appear in Appendix E.1.

Naturalness gradient. ΔG tracks the surface form of the target text, ranging from long free-form prose at one end to short telegraphic entries at the other. All 6 classifiers improve on 20 Newsgroups, with gains from $+11.6$ to $+24.8$pt, and all 6 improve on ASRS with gains up to $+30.8$pt; LDA+SVC shows the smallest ASRS gain at $+8.1$pt. AG News headlines improve on 5 of 6 classifiers, with DstlB the lone regression at -1.2pt. MaintIE's 5-word entries regress uniformly on all 6 classifiers.

Classifier asymmetry. LR improves on 6 of 7 datasets, RF on 5 of 7, DstlB on 4 of 7. One interpretation is that vocabulary-level guidance aligns most strongly with classifiers whose decisions are dominated by lexical features; Transformer classifiers also attend to contextual and positional patterns that vocabulary-only guidance does not capture, making attention-aware guidance a natural future direction.

Rewritten as a robust fallback. When Guided regresses on a classifier, Rewritten often recovers the gap because the original text anchors vocabulary and structure. FDA MAUDE illustrates this asymmetry: ΔR is positive on 5 of 6 classifiers, reaching 97.8%$\rightarrow$100.0% on the best classifier, while ΔG is more volatile (SetFit -14.0pt, others $+1.8$ to $+7.8$pt); we cite this for the *pattern* since the binary task is near-saturated. PubMed RCT shows the same pattern: ΔR exceeds ΔG on 3 of 6 classifiers, locating the bottleneck in the LLM's domain fluency rather than COMPASS's feature selection. The fallback breaks down on MaintIE's 5-word entries, where the source is too short to anchor anything. Rewritten ROUGE-L ranges 0.22–0.42 across datasets (Appendix F), comparable to human paraphrases (Dolan and Brockett, 2005).

The patterns above reveal a decoupling between guidance content and LLM domain fluency: when

Dataset	LR			RF			SVC			DstlB			SBERT			SetFit		
	O	ΔG	ΔR	O	ΔG	ΔR	O	ΔG	ΔR	O	ΔG	ΔR	O	ΔG	ΔR	O	ΔG	ΔR
20 News	82.4	**+17.6**	**+16.3**	75.2	**+24.8**	**+24.1**	76.4	**+23.6**	**+21.6**	88.2	**+11.8**	**+11.5**	88.4	**+11.6**	**+11.6**	87.6	**+12.4**	**+12.1**
AG News	88.3	+8.8	+5.5	83.0	**+13.7**	+9.1	76.2	**+21.8**	+5.9	91.6	−1.2	+2.1	88.7	+8.8	+2.6	88.5	+8.1	+2.7
Financial	80.9	**+19.1**	**+16.9**	80.7	−2.3	−7.9	67.4	**−34.0**	**−30.2**	86.2	**+13.8**	+8.3	79.0	**+21.0**	**+14.9**	72.9	**+15.4**	+7.1
PubMed	70.7	+7.0	+9.6	66.1	**+15.2**	**+11.9**	43.9	**+24.1**	**+18.8**	74.9	−0.2	+7.8	57.5	**+11.5**	−4.8	69.1	**−18.4**	**−11.1**
ASRS	67.4	**+27.9**	**+13.3**	62.8	**+15.9**	−1.1	62.2	+8.1	−4.6	68.7	**+12.6**	**+10.6**	63.9	**+30.8**	+5.8	41.2	**+21.8**	**+16.1**
FDA MAUDE	96.1	+3.9	+2.2	97.4	+1.8	+0.9	92.4	+4.3	+0.9	97.8	+2.2	+2.2	92.2	+7.8	+5.3	91.5	**−14.0**	−6.5
MaintIE	78.3	−10.0	**−27.6**	78.3	**−18.3**	**−27.0**	73.1	**−36.4**	**−30.4**	70.8	**−37.4**	**−30.8**	80.2	**−16.9**	−9.5	75.5	−7.7	+5.2

Table 2: Per-classifier accuracy across 7 datasets and 6 classifiers, averaged over the two LLMs. For each classifier, O = baseline (Original) accuracy on the test set; ΔG = Guided improvement; ΔR = Rewritten improvement (both in percentage points). Green = positive Δ, red = negative Δ; bold marks $|\Delta| \geq 10$ pp.

Dataset	Class	$J_{\text{LR}\cap\text{RF}}$	Common features
20 News	sci.space	0.611	nasa, orbit, spacecraft
AG News	Sports	0.611	baseball, league, team
Financial	bearish	0.429	downgraded, falls, loss
PubMed	results	0.714	ci, group, significantly
ASRS	Aircraft	0.429	engine, emergency, fire
FDA	Injury	0.429	adverse, infection, patient
MaintIE	Activity	0.500	calibrate, repair, replace

Table 3: LR∩RF Jaccard for each dataset's top-Jaccard class, after preprocessing.

Dataset	XGBoost		Multinom. NB		Zero-shot NLI	
	Orig	ΔG	Orig	ΔG	Orig	ΔG
20 News	73.3	**+26.7**	86.0	**+14.0**	88.9	**+11.1**
AG News	82.5	**+13.3**	89.2	**+10.0**	86.7	**+13.3**
Financial	74.4	**+11.1**	66.7	+3.3	87.8	**+11.7**
PubMed	70.7	−5.3	62.0	**+18.0**	38.0	**+16.7**
ASRS	55.3	**+33.3**	34.7	+5.3	45.3	**+41.3**
FDA MAUDE	100.0	+0.0	91.7	**−41.7**	51.7	**+48.3**
MaintIE	93.3	**−22.8**	60.0	+6.7	46.7	**+49.4**

Table 4: Anti-circularity check using held-out classifiers. Format as in Table 2.

an LLM cannot generate text authentic to the target domain, Rewritten still recovers gains, showing that COMPASS's guidance content is sound and the limitation lies in the LLM evaluation proxy. COMPASS targets human authors, who write in their own domain idiom natively; we revisit this point in the human evaluation (§6.4) and Discussion (§7). Figure 4 confirms the pattern geometrically: Guided and Rewritten documents cluster near training centroids rather than forming keyword-heavy outliers, evidencing writing authentic to the target domain rather than gaming.

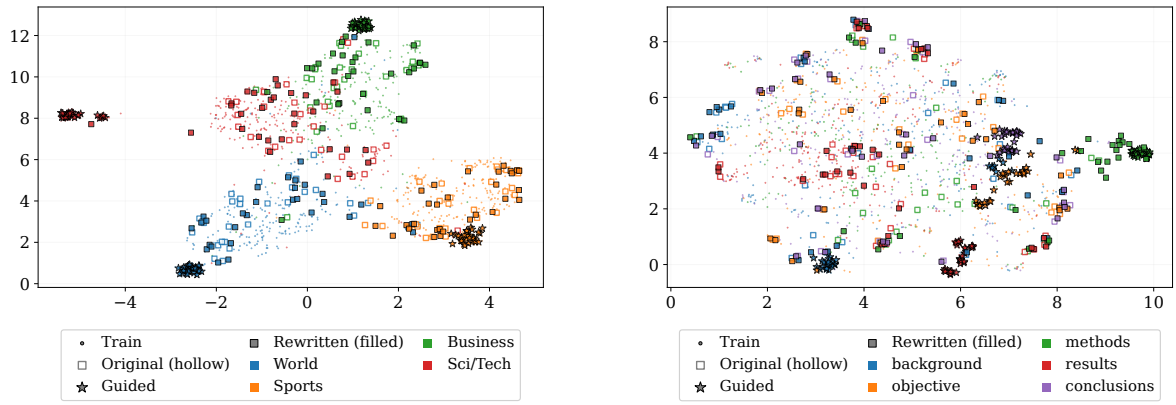
6.2 Cross-Classifier Robustness

We assess three forms of robustness: feature stability across the six classifiers used in profile extraction; accuracy under held-out classifiers not used in profile extraction; and consistency when the LLM differs between the guidance and document-generation steps.

Profile stability across the classifiers used in profile extraction. Table 3 shows that domain-specific terminology is stable across architecturally distinct classifiers (LR∩RF Jaccard 0.43–0.71 on common features after preprocessing), justifying the stability filter applied across classifiers (§4.2).

Held-out classifiers (anti-circularity check). A natural concern is that the gains in Table 2 are specific to the six classifiers whose features built the guidance. We re-evaluate the same Guided documents with three classifiers *not* used in COMPASS: XGBoost (Chen and Guestrin, 2016) on TF-IDF features, Multinomial Naive Bayes (McCallum and Nigam, 1998), and zero-shot NLI with DeBERTa-v3 (Laurer et al., 2024); the last has *no training on* COMPASS *data at all*. Across 7 datasets (Table 4), ΔG is positive on **4 of 7 datasets for XGBoost, 6 of 7 for Naive Bayes, and 7 of 7 for NLI**. NB shows only +5.3pt on ASRS, the smallest of the three. On four of the five broad-prose datasets (Financial, 20 Newsgroups, AG News, ASRS) *all three* held-out classifiers show positive ΔG; on the fifth (PubMed), XGBoost regresses (−5.3pt) while NB and NLI remain strongly positive (+18.0 and +16.7pt). These results substantially address the circular-evaluation concern: on broad-prose datasets gains hold across all three held-out families, including the zero-shot NLI classifier that was never trained on COMPASS data; specialised datasets show family-dependent behaviour, with NLI positive on both (+48.3, +49.4pt) while NB collapses on FDA (−41.7pt) and XGBoost on MaintIE (−22.8pt).

Cross-LLM consistency. If gains depend on a specific LLM, swapping the LLM between the guidance step (Stage 3) and the document-generation step should erode the pattern. The



(a) AG News.

(b) PubMed RCT.

Figure 4: UMAP projections for AG News and PubMed RCT. Remaining 5 datasets in Appendix H.

four pairings—same LLM throughout (Qwen for both, Mistral for both) or different LLMs (Qwen guidance + Mistral docs, and vice versa)—yield matching ΔG patterns across all six classifiers on all 7 datasets (Appendix E.2), supporting LLM-independence of the guidance content.

6.3 Pipeline Decomposition

We ablate four configurations: full COMPASS, −Stab. (no stability filter), Feat-list (raw feature list, no LLM mediation), and No Prof. (class name only) (Table 5). Four findings emerge.

Classifier Profiles are essential. The no-profile baseline drops to 50.0–74.6% on AG News, PubMed, and ASRS (vs. 89–100% for COMPASS): profiles supply domain vocabulary the LLM does not possess.

LLM mediation prevents keyword stuffing. Feat-list reaches or exceeds COMPASS accuracy on 6 of 7 datasets but raises KD flag rates to 39–95% on the natural-prose datasets (Financial, 20 News, AG News, PubMed); COMPASS mediates this to single-digit or low-double-digit flagging on most datasets.

Stability filter improves robustness. Removing it causes degradation on AG News (100%→85.4%), PubMed (89.3%→54.0%), ASRS (100%→70.0%), and MaintIE (79.4%→43.3%).

Anti-Gaming flags remain low under COMPASS. Under full COMPASS, KD flag rates remain $\leq 15\%$ on 6 of 7 datasets (4 datasets $\leq 5\%$: Financial, PubMed, FDA MAUDE, MaintIE); ASRS (39%) is the lone elevated case (see Limitations). Per-variant KD flag rates appear in the bottom panel of Table 5.

Dataset	COMPASS	−Stab.	Feat-list	No Prof.
	Guided Accuracy (%) ↑			
Financial	**100**	97.2	**100**	90.6
20 News	**100**	**100**	**100**	**100**
AG News	**100**	85.4	**100**	74.6
PubMed	89.3	54.0	**90.0**	50.0
ASRS	**100**	70.0	95.3	62.0
FDA MAUDE	**100**	**100**	**100**	72.5
MaintIE	79.4	43.3	**82.2**	64.4
	KD Flag Rate (%) ↓			
Financial	4	**0**	69	**0**
20 News	15	4	65	**0**
AG News	12	**0**	95	1
PubMed	1	**0**	39	**0**
ASRS	39	15	31	**6**
FDA MAUDE	**0**	**0**	8	**0**
MaintIE	**0**	**0**	9	**0**

Table 5: Component ablation. **Top**: Guided accuracy; **Bottom**: KD flag rate (red = severe stuffing). COMPASS = full pipeline; −Stab. = no Stage-2 filter; Feat-list = no Stage-3 LLM; No Prof. = no Stages 1–2.

6.4 Human Evaluation

We conducted a human evaluation with nine quality review practitioners from a regulated industrial manufacturer, both to evaluate COMPASS guidance quality beyond the LLM proxy and to test whether prescriptive guidance disproportionately helps less-experienced writers.

Setup and results. The corpus is 285 internal quality review minutes, labelled into 7 phenomenon-based incident categories; all entities are anonymised before classifier training (Appendix G). Nine practitioners rated COMPASS guidance on six Likert dimensions, with guidance gen-

Dimension	Mean	SD	Novice	Senior
Actionability	3.67	1.12	4.50	2.67
Adoption Intent	3.56	1.01	3.50	3.33
Usefulness	3.22	1.09	3.50	2.33
Generalizability	2.89	0.78	3.50	2.33
Completeness	2.78	0.83	3.00	2.33
Readability	2.22	0.97	2.00	1.67
Total (/30)	—	—	**20.0**	**14.7**

Table 6: Human evaluation on a 5-point Likert scale.

erated by Qwen 2.5 7B across all 7 categories (Table 6). The cohort comprised four novices with less than one year of experience, two mid-level practitioners with one to three years, and three seniors with five or more years. Actionability (3.67/5) and Adoption Intent (3.56/5) are strongest; Readability (2.22/5) is weakest. Open-text feedback indicates this reflects anonymisation placeholders substituted for sensitive entities such as customer names and equipment IDs; respondents indicated the readability cost is recoverable in practice.

Experience effect. Novices rated the guidance higher than seniors, scoring **20.0 vs. 14.7** out of 30. The difference yields Welch's t=2.40 at p=0.062 and Hedges' g=1.47 with 95% CI $[-0.55, 4.05]$. With n=9, this is a preliminary indication of a large effect under low statistical power. Internal consistency across the six dimensions is acceptable, with Cronbach's α=0.742. The largest gap is on Actionability (4.50 vs. 2.67), while Adoption Intent stays near-uniform (3.50 vs. 3.33). COMPASS answers "what should I write?" for newcomers; pending replication, this provides preliminary evidence that prescriptive guidance most benefits less-experienced writers.

7 Discussion

COMPASS reframes classifier internals from a debugging tool for ML practitioners into a writing tool for end-user authors. The no-profile baseline drops to 50.0–74.6% accuracy on AG News, PubMed, and ASRS, compared with 89–100% for full COMPASS (Table 5); this shows that prompting an LLM for a target class is insufficient on its own. The cross-classifier stable vocabulary supplies decision-boundary-aligned cues the LLM does not select from the class name alone.

LLM evaluation is a lower bound on real utility. Our results show a consistent asymmetry: Guided text generated by the LLM improves all 6 classifiers on natural-prose datasets such as 20 Newsgroups and ASRS, while on specialised domains such as PubMed RCT and FDA MAUDE, Rewritten exhibits tighter variance than Guided (Table 2): on FDA MAUDE, ΔG ranges -14.0 to $+7.8$pt across classifiers while ΔR is bounded within -6.5 to $+5.3$pt, halving the worst regression. Prompt engineering compounds this gap: more verbose document-generation prompts cause Guided accuracy to drop sharply on FDA MAUDE by approximately 23 percentage points (Appendix E.1), while Rewritten varies by only 3 pp. The bottleneck is the LLM's fluency in the target domain when over-constrained by the prompt, not COMPASS's feature selection. Crucially, COMPASS is designed for *human* authors, who write in their own domain idiom natively. Our stratified human evaluation supports this reading: novices rate the guidance 20.0/30 vs. seniors' 14.7/30 with Hedges' g=1.47, and the largest gap appears on Actionability at 4.50 vs. 2.67. MaintIE's uniform regression is consistent with this lower-bound interpretation (Table 2): short specialised entries leave little surface for an LLM to anchor on, but a domain practitioner can apply vocabulary guidance directly.

Practitioner implications. COMPASS is best suited to settings where classifier decisions have real consequences for non-ML-literate authors — examples include triage, screening, and regulatory reporting — and where the text is natural prose. The novice–senior gap (§6.4) suggests highest near-term impact in *onboarding* scenarios.

8 Conclusion

COMPASS turns classifier internals into *proactive* writing guidance for human authors. Across 7 benchmarks, Logistic Regression gains on 6 of 7 datasets, with gains up to $+27.9$pt on ASRS; MaintIE is the lone exception. Rewriting recovers specialised text where LLMs cannot generate content authentic to the target domain: on PubMed RCT, DistilBERT's -0.2pt Guided regression is recovered to $+7.8$pt under Rewriting; on FDA MAUDE the best classifier reaches 100.0% under both Guided and Rewritten. Anti-Gaming, UMAP, cross-LLM transfer, and held-out classifiers jointly confirm that the gains do not stem from gaming, and a stratified human study with Hedges' g=1.47 indicates larger gains for less-experienced writers; attention-aware guidance is a natural next step.

Limitations

- **Statistical power of the human evaluation.** The experience-stratified test approaches but does not reach $p<0.05$ ($p=0.062$, $n=9$). Post-hoc power is 35–46% (under Hedges' $g=1.47$ and Cohen's $d=1.75$ respectively), well below the conventional 80%; the result should be interpreted as a directional indication rather than confirmatory evidence.
- **Evaluation sampling and LLM stochasticity.** The Guided and Rewritten evaluation uses $n=30$ per class; MaintIE has some classes with fewer than 30 test documents, so several per-classifier Δ values on this dataset carry wide confidence intervals (Table 2). Documents are generated from a single LLM run per condition; reported magnitudes (but not directions) may vary across seeds, with multi-seed variance estimates deferred to future work.
- **LLM evaluation as a proxy.** LLM-based evaluation is scalable but underestimates COMPASS's utility on specialised text (§7). Our human evaluation (§6.4, $n=9$) provides initial validation on a single organisation; generalisation across multiple organisations and domains is beyond the scope of this study.
- **Confidentiality of the human evaluation corpus.** The human evaluation uses proprietary industrial data; we report aggregate statistics and the full protocol (§6.4, Appendix G) but cannot release raw data. The protocol is described in sufficient detail to enable replication on comparable corpora.
- **Pipeline assumptions.** Preprocessing removes noise but may also remove domain-relevant terms; the net impact is positive (Appendix A), but corpora dominated by numeric or symbol-laden vocabulary may suffer. Strict full-intersection across all six classifiers is near-empty due to vocabulary gaps between classical ML and Transformer profiles, so the stability filter requires ≥ 2 classifiers rather than full intersection (Eq. 2).
- **Anti-Gaming module limitations.** The Keyword Density Guard's $1.5\times$ training 95th-percentile multiplier is a reasonable default but not optimised; a systematic threshold-sensitivity analysis is left for future work. On ASRS, 39% of Guided documents are flagged (§6.3), driven by LLM-generated article-style explanations with one section per stable feature rather than incident narratives; the flagged documents still classify correctly into the target class, indicating over-literal compliance with feature-listing prompts rather than classifier-defeating keyword cramming.

Ethical Considerations

Human-subjects evaluation. The human evaluation (§6.4) involved nine voluntary, uncompensated adult employees with opt-out at any time. Before completing the survey, participants were briefed on the study purpose, the AI-derived origin of the guidance, the expected duration, and the aggregate-only use of their responses; submitting the form constituted informed consent. Reported statistics are aggregate, with experience level (<1 yr / 1–3 yr / 5+ yr) as the only personal covariate. The study followed the organisation's data-handling guidelines and did not require external IRB review under the applicable workplace-survey policy.

Confidential industrial data. The 285 internal records are proprietary. We anonymise all text before classifier fitting (regex + multilingual NER), report only aggregate statistics, and do not release the raw corpus; the protocol enables replication on comparable corpora.

Dual-use and adversarial use. COMPASS targets *cooperative* settings where authors and classifier deployers share the accuracy goal (industrial, medical, aviation reporting). The Anti-Gaming Module (§5.4) flags keyword stuffing as an explicit mitigation. We discourage deployment in adversarial settings (e.g. content-moderation evasion) without these safeguards.

Classifier-as-oracle and bias propagation. COMPASS treats the supervising classifiers as the authority on what constitutes good writing in a target class. This is a strong assumption: if a classifier has learned spurious correlations or absorbed biases from its training distribution, those signals enter the guidance the LLM produces, and authors who comply with it reproduce them in fresh text. The cross-classifier stability filter (§4.2) mitigates the spurious-feature case by retaining only features that ≥ 2 architecturally distinct classifiers agree on. It does not, however, detect biases that are systematic across all classifiers trained on the same biased distribution. In deployments where author-revised text feeds back into classifier training, this risks a feedback loop:

guidance pushes writing toward classifier-preferred patterns, the next training cycle reinforces those patterns, and the corpus narrows over time. Practitioners should audit COMPASS profiles for sensitive attributes before deployment, and avoid using author-revised text as classifier training data without independent quality control.

Use of AI assistants. For paper preparation we used a general-purpose LLM-based coding/writing assistant for code edits and surface prose tightening. All scientific claims and analyses are the authors' and were manually verified.

References

Farhad Akhbardeh, Travis Desell, and Marcos Zampieri. 2020. MaintNet: A collaborative open-source library for predictive maintenance language resources. In *Proceedings of COLING: System Demonstrations*.

Tyler K Bikaun, Tim French, Michael Stewart, Wei Liu, and Melinda Hodkiewicz. 2024. MaintIE: A fine-grained annotation schema and benchmark for information extraction from maintenance short texts. In *Proceedings of LREC-COLING*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Michael P Brundage, Thurston Sexton, Melinda Hodkiewicz, Alden Dima, and Sarah Lukens. 2021. Technical language processing: Unlocking maintenance knowledge. *Manufacturing Letters*, 27:42–46.

Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. 2024. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9742–9763, Bangkok, Thailand. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of KDD*, pages 785–794.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *Proceedings of ICLR*.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of IJCNLP*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of ACL*.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of ITCS*.

Melinda Hodkiewicz and Mark TW Ho. 2016. Cleaning historical maintenance work order data for reliability analysis. *Journal of Quality in Maintenance Engineering*, 22(2):146–163.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of FAccT*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*.

Ken Lang. 1995. NewsWeeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning (ICML)*, pages 331–339.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis*, 32(1):84–100.

Mina Lee and 1 others. 2022. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In *Proceedings of CHI*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in NeurIPS*.

Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48.

John Miller, Smitha Milli, and Moritz Hardt. 2020. Strategic classification is causal modeling in disguise. In *Proceedings of ICML*.

NASA. 2024. NASA aviation safety reporting system (ASRS). https://asrs.arc.nasa.gov/.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Kaivalya Rawal and Himabindu Lakkaraju. 2020. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. In *Advances in NeurIPS*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of KDD*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of AAAI*.

Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of ACL*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Antonette Shibani, Simon Knight, and Simon Buckingham Shum. 2019. Contextualizable learning analytics design: A generic model and writing analytics evaluations. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK '19)*, pages 210–219.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

U.S. Food and Drug Administration. 2024. FDA MAUDE: Manufacturer and user facility device experience database. https://open.fda.gov/.

Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of FAccT*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of ACL-IJCNLP*.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of NAACL*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Preprocessing Impact

Table 7 reports the fraction of top features remaining after removing stopwords, special characters, email/URL patterns, and subword fragments. Three patterns motivate the preprocessing pipeline. First, LDA+SVC carries the heaviest feature contamination on six of seven datasets (30–38%), MaintIE the lone exception (10%). Second, RF contamination is dataset-dependent: PubMed RCT (53%) and ASRS (44%) are highest while MaintIE (5%) is lowest; the remaining four datasets span 18–35%. Third, perturbation-based profiles from DistilBERT, SBERT, and SetFit are uniformly cleaner: DistilBERT is $\geq$92% clean on every dataset; SBERT and SetFit are $\geq$80% on six of seven, with Financial Sentiment the lone exception for both (77% and 68%).

Dataset	LR	RF	SVC	DstlB	SBERT	SetFit
20 News	86	65	64	97	93	92
AG News	91	79	70	92	98	98
Financial	75	68	67	92	77	68
PubMed	67	47	63	94	92	80
ASRS	93	56	69	96	92	90
FDA MAUDE	90	82	62	95	80	88
MaintIE	92	95	90	92	97	97

Table 7: Profile clean ratio (%) before preprocessing.

B Classifier Profile Examples

Table 8 shows cleaned top-5 features for selected classes across classifier types.

Two patterns motivate the Stage-2 stability filter. Classical ML classifiers (LR, RF) produce highly overlapping profiles—supporting LR$\cap$RF

Class	Clf	Top-5 Features
Financial bearish	LR	down, lower, misses, downgraded, falls
	RF	down, misses, lower, shares, downgraded
	SBERT	retreats, worst, downgraded, fall, plummeting
MaintIE State	LR	fault, error, broken, temperature, supply error
	RF	fault, error, broken, temperature
	SBERT	error, fault, damaged, faults, check

Table 8: Cleaned Classifier Profile examples. Classical ML classifiers (LR, RF) produce highly overlapping profiles. SBERT captures semantic synonyms (e.g. "plummeting" for bearish, "damaged" for State).

as a high-confidence subset (Table 3). Perturbation-based extraction from SBERT surfaces semantic synonyms (*plummeting* for *bearish*, *damaged* for *State*) that classical ML does not—useful when pooled, with the ≥2-classifier filter still removing SBERT-only artefacts.

C Guidance Generation Prompt

This appendix lists the prompt COMPASS sends to the LLM in Stage 3 to produce the per-class guidance (§4.3). The running example is the Financial-Sentiment *bearish* class.

Stage 3 prompt template. *System prompt:*

> You are a writing coach that helps authors write documents that will be correctly classified by text classifiers. Convert classifier feature profiles into actionable, natural-language writing guidance. Be specific and practical. Do not mention classifiers, features, or weights directly — frame advice as writing tips.

User prompt:

> Generate writing guidance for the category '[class_name]'.
>
> Cross-classifier stable features (appear in ≥2 classifiers): [stable features]
>
> By default, provide 5–7 specific, actionable writing tips that naturally incorporate these key vocabulary items. Each tip should explain WHAT to write and WHY it helps. Keep the tone professional and encouraging.

Example user prompt (Financial-Sentiment *bearish*):

> Generate writing guidance for the category 'bearish'.
>
> Cross-classifier stable features (appear in ≥2 classifiers): china, coronavirus, cut, cuts, down, downgraded, fall, falls, loss, lower, misses, oil, premarket, revenue, shares, tumbled
>
> By default, provide 5–7 specific, actionable writing tips . . .

Qwen 2.5 7B returns six writing tips under this prompt; the first three are shown in Figure 2 (the remaining three are omitted for space). A parallel Japanese system+user template is used when COMPASS runs on Japanese corpora (e.g., the industrial human evaluation of Appendix G). The prompts for the evaluation-time document generation step are in Appendix D.

D Document Generation Prompts and PromptConfig

This appendix specifies the prompts the evaluation LLM receives to produce Guided and Rewritten documents (§5.3), and shows a fully filled-in PromptConfig for the running Financial-Sentiment *bearish* example.

Guided prompt template (generate-from-scratch condition).

System prompt:

> You are a document author. Write a short document ([length_spec]) in English. Output ONLY the text. [style_hint, if active]

User prompt:

> [Context: [description], if active]
> Write a document about '[class_name]'. [context]
> Follow these guidelines:
> [guidance]
> [Examples: [few_shot_examples], if active]

Rewrite prompt template (revise-existing condition).

System prompt:

> You are an editor. Rewrite the given document to improve its domain specificity while preserving the original meaning. Keep approximately the same length ([length_spec]). Output ONLY the rewritten text. [style_hint, if active]

User prompt:

> Original document:
> [original_text]
>
> Rewrite using these guidelines:
> [guidance]
>
> Preserve the original meaning and facts. Improve domain vocabulary and specificity.

Example PromptConfig (Financial-Sentiment *bearish*, `full` prompt mode, all components active).

- `length_spec`: "9–15 words (median 11)"
- `context` (domain context, LLM-generated): "This classifier processes financial earnings announcements or news articles expressing market sentiment towards stocks."
- `style_hint` (LLM-generated): "1. Use all uppercase letters for ticker symbols and abbreviate

company names when possible. 2. Write in a formal, telegraphic style with minimal punctuation, focusing on brevity and clarity.. . . ”

- `few_shot_examples` (sampled from training data; 3 of 9 shown):
  - `$BYND - JPMorgan reels in expectations on Beyond Meat`
  - `$CX - Cemex cut at Credit Suisse, J.P. Morgan on weak building outlook`
  - `$CCL $RCL - Nomura points to bookings weakness at Carnival and Royal Caribbean`
- `description` (manual, opt-in): “Short financial news headlines and tweets expressing market sentiment toward stocks and companies.”

Under this `full` configuration, Rewritten preserves the source ticker ($TAP) while adopting guidance vocabulary, and Guided is fully synthesised from the stable-feature set. Body Figure 3 shows the parallel outputs under the headline-experiment `minimal` mode, where the ticker prefix is not retained.

E Ablation Studies

E.1 Prompt engineering ablation

We ablate three prompt components—style instructions (S), few-shot examples (F), and dataset description (D)—across all $2^3 = 8$ combinations (Table 9); guidance (G), length specification (L), and the auto-generated domain context (C) are kept constant in every mode. The manually-curated descriptions that fill the (D) slot for each dataset appear in Table 10.

Mode	G	L	C	S	F	D	Purpose
full	✓	✓	✓	✓	✓	✓	All components
no_desc	✓	✓	✓	✓	✓		Test desc. harm
no_fewshot	✓	✓	✓	✓		✓	Test few-shot value
no_style	✓	✓	✓		✓	✓	Test style value
style_only	✓	✓	✓	✓			Style in isolation
fewshot_only	✓	✓	✓		✓		Few-shot in isolation
desc_only	✓	✓	✓			✓	Desc. in isolation
minimal	✓	✓	✓				Guidance + length + context only

Table 9: Prompt ablation modes. **G** = Stage-3 guidance, **L** = length specification, **C** = auto-generated domain context (all three always on); **S** = style, **F** = few-shot, **D** = description (toggled across the $2^3 = 8$ combinations).

Table 11 shows the full 2^3=8 prompt ablation across all datasets. Values are mean-classifier Guided accuracy (%) averaged across the six classifiers, then averaged over Qwen and Mistral (n=30/class). Table 12 reports the corresponding cross-classifier standard deviation as a stability supplement.

Dataset	Description
20 Newsgroups	Informal discussion posts from internet newsgroups on topics including science, politics, religion, and technology.
AG News	News article headlines and lead sentences covering world affairs, sports, business, and science/technology.
Financial Sent.	Short financial news headlines and tweets expressing market sentiment toward stocks and companies.
PubMed RCT	Individual sentences from medical research abstracts describing study objectives, methods, results, and conclusions.
ASRS	Pilot-written aviation safety narratives reporting incidents, near-misses, and operational problems (100–500 words).
FDA MAUDE	Medical device adverse event reports describing malfunctions and patient injuries in clinical/regulatory language.
MaintIE	Short maintenance work order texts describing physical objects, their states, and activities performed.

Table 10: Dataset descriptions provided to the LLM as the `description` (D) field. These are manually curated to capture domain-specific text characteristics.

Key findings: (1) **Multi-component prompts over-constrain the LLM:** `full` (76.6%) and `no_desc` (76.4%) are the two weakest modes by mode-avg, separated from the next tier by 4–7 pp. (2) **Simpler prompts dominate the top:** `desc_only` (83.4%), `style_only` (83.4%), and `minimal` (83.0%) cluster within 0.4 pp; `minimal` is adopted for main experiments as competitive on accuracy and most stable across classifiers within the top tier. (3) **Description is helpful alone but hurts when combined with style and few-shot:** on PubMed, `desc_only` 74.9% vs. `full` 63.7%; on FDA MAUDE, `desc_only` 94.2% vs. `full` 72.8%. (4) **Classifier heterogeneity is dataset-determined:** cross-classifier std (Table 12) varies far more by dataset than by mode—Financial (∼24 pp) and MaintIE (∼16 pp) expose classifier-specific responses, while 20 News and AG News are nearly classifier-agnostic (<3 pp), motivating the multi-classifier evaluation used throughout the paper.

Dataset	min.	no_d	no_f	no_s	s_only	f_only	d_only	full
Financial	83.3	77.2	**85.3**	78.7	84.1	81.9	82.6	81.0
20 News	**100**	99.8	**100**	**100**	99.9	**100**	**100**	99.8
MaintIE	54.9	49.2	50.4	56.1	52.9	**62.9**	58.5	51.2
AG News	**96.0**	93.3	95.3	93.3	94.9	94.9	93.5	93.4
PubMed	70.2	59.6	**77.9**	64.1	77.6	60.6	74.9	63.7
ASRS	80.6	74.0	81.2	76.2	**84.3**	72.3	80.1	74.2
FDA MAUDE	**95.6**	81.9	89.6	91.3	89.9	92.8	94.2	72.8
mode-avg	82.9	76.4	82.8	79.9	**83.4**	80.8	**83.4**	76.6

Table 11: Full prompt ablation: mean-classifier Guided accuracy (%), 2 LLM avg. Columns: min.=minimal, no_d=no description, no_f=no few-shot, no_s=no style, s/f/d_only=single component, full=all. Bold = best per dataset (top section) or best across modes (bottom row). Simpler prompts (`minimal`, `style_only`, `desc_only`) cluster at the top; `full` and `no_desc` are the two weakest modes, suggesting that multi-component prompts over-constrain the LLM.

Dataset	min.	no_d	no_f	no_s	s_only	f_only	d_only	full	row avg
Financial	26.5	23.1	23.5	22.5	25.2	24.9	25.4	23.7	**24.3**
20 News	0.0	0.3	0.0	0.0	0.1	0.0	0.0	0.4	**0.1**
MaintIE	16.1	14.4	15.0	18.7	14.6	18.6	18.0	15.4	**16.3**
AG News	2.8	2.0	2.8	3.2	2.5	1.5	3.7	4.6	**2.9**
PubMed	12.2	20.3	15.5	15.0	11.3	17.7	12.6	17.9	**15.3**
ASRS	14.0	17.5	13.8	12.7	13.2	16.8	13.8	16.1	**14.7**
FDA MAUDE	8.9	11.6	14.0	10.5	13.3	8.2	10.8	12.1	**11.2**

Table 12: Stability supplement to Table 11: cross-classifier standard deviation (pp) of Guided accuracy across the six classifiers, 2 LLM avg. Lower = more classifier-agnostic Guided documents (every classifier agrees on the document's class). Variation is dominated by dataset (row) rather than prompt mode (column): 20 News and AG News are nearly classifier-agnostic (<3 pp), while Financial (∼24 pp) and MaintIE (∼16 pp) expose classifier-specific responses to Guided text.

E.2 LLM ablation: cross-LLM transfer

Tables 13 and 14 report per-classifier Original accuracy, ΔG, and ΔR under all four LLM pairings, one table per prompt mode. We write **A→B** for the pairing in which LLM A produced the guidance (Stage 3 tips, §4.3) and LLM B then produced the Guided and Rewritten documents under that guidance (§5.3). Q = Qwen 2.5 7B; M = Mistral 7B. Body Table 2 is the diagonal average $(Q{\to}Q + M{\to}M)/2$ under `minimal` mode; the two tables below provide the underlying per-cell view. Each follows the format of body Table 2 (n=30/class; green = positive, red = negative; bold $|\Delta| \geq 10$ pp).

Symmetry. Off-diagonal transfer (Q→M and M→Q) is essentially symmetric to the diagonal across all six classifiers on all 7 datasets: ΔG patterns under either off-diagonal pairing closely mirror Q→Q and M→M in both sign and magnitude. This supports the central claim of LLM independence: the discriminative signal lives in the Classifier Profile, not in any phrasing of the guidance specific to a particular LLM.

Surface-level drift between languages. A separate observation that does not affect classification accuracy: when Qwen generates documents from Mistral-authored guidance, the output occasionally includes non-ASCII (notably CJK) characters embedded in otherwise-English text. The effect is small in length-rich corpora (mean <0.1 non-ASCII characters per document on all 7 datasets) and is a property of Qwen's generation rather than the guidance content.

F Rewritten Content Preservation

Table 15 reports per-dataset ROUGE-L of Rewritten documents against their source test documents, averaged across the two LLMs (Qwen 2.5 7B and Mistral 7B) under the configuration of Table 2.

G Industrial Quality Review Study: Detailed Methodology

This appendix details the protocol behind the experience-stratified human evaluation reported in §6.4.

Dataset	LR			RF			SVC			DstlB			SBERT			SetFit		
	O	ΔG	ΔR	O	ΔG	ΔR	O	ΔG	ΔR	O	ΔG	ΔR	O	ΔG	ΔR	O	ΔG	ΔR
Q→Q																		
20 News	82.4	**+17.6**	**+16.3**	75.2	**+24.8**	**+23.5**	76.4	**+23.6**	**+21.6**	88.2	**+11.8**	**+11.1**	88.4	**+11.6**	**+11.6**	87.6	**+12.4**	**+11.8**
AG News	88.3	**+11.7**	+1.7	83.0	**+17.0**	+4.5	76.2	**+23.8**	−1.2	91.6	+8.4	+0.9	88.7	**+11.3**	−1.2	88.5	**+11.5**	−0.2
Financial	80.9	**+19.1**	**+14.7**	80.7	+4.9	−1.8	67.4	**−34.0**	**−27.4**	86.2	**+13.8**	+4.9	79.0	**+21.0**	**+11.0**	72.9	+7.1	+0.4
PubMed	70.7	**+26.6**	**+13.3**	66.1	**+33.2**	**+20.6**	43.9	**+30.8**	**+20.1**	74.9	**+17.1**	**+12.4**	57.5	**+35.2**	−5.5	69.1	−3.8	+4.9
ASRS	67.4	**+32.6**	**+21.3**	62.8	**+17.2**	+6.6	62.2	+3.8	+0.4	68.7	**+26.0**	**+14.0**	63.9	**+34.1**	**+12.8**	41.2	**+22.1**	**+22.1**
FDA MAUDE	96.1	+3.9	+3.9	97.4	+2.6	+0.9	92.4	+7.6	+2.6	97.8	+2.2	+2.2	92.2	+7.8	+6.1	91.5	+8.5	+5.2
MaintIE	78.3	**−11.6**	**−27.6**	78.3	**−11.6**	**−25.0**	73.1	**−38.7**	**−27.8**	70.8	**−37.4**	**−30.8**	80.2	**−13.5**	**−20.2**	75.5	−8.8	+5.9
M→M																		
20 News	82.4	**+17.6**	**+16.3**	75.2	**+24.8**	**+24.8**	76.4	**+23.6**	**+21.6**	88.2	**+11.8**	**+11.8**	88.4	**+11.6**	**+11.6**	87.6	**+12.4**	**+12.4**
AG News	88.3	+5.9	+9.2	83.0	**+10.4**	**+13.7**	76.2	**+19.7**	**+13.0**	91.6	**−10.8**	+3.4	88.7	+6.3	+6.3	88.5	+4.8	+5.6
Financial	80.9	**+19.1**	**+19.1**	80.7	−9.5	**−14.0**	67.4	**−34.0**	**−32.9**	86.2	**+13.8**	**+11.6**	79.0	**+21.0**	**+18.8**	72.9	**+23.8**	**+13.8**
PubMed	70.7	**−12.7**	+6.0	66.1	−2.8	+3.2	43.9	**+17.4**	**+17.4**	74.9	**−17.6**	+3.1	57.5	**−12.2**	−4.2	69.1	**−33.1**	**−27.1**
ASRS	67.4	**+23.3**	+5.3	62.8	**+14.6**	−8.8	62.2	**+12.4**	−9.6	68.7	−0.7	+7.3	63.9	**+27.5**	−1.2	41.2	**+21.5**	**+10.1**
FDA MAUDE	96.1	+3.9	+0.6	97.4	+0.9	+0.9	92.4	+0.9	−0.7	97.8	+2.2	+2.2	92.2	+7.8	+4.5	91.5	**−36.5**	**−18.2**
MaintIE	78.3	−8.3	**−27.6**	78.3	**−25.0**	**−29.0**	73.1	**−34.2**	**−33.1**	70.8	**−37.4**	**−30.8**	80.2	**−20.2**	+1.1	75.5	−6.6	+4.5
Q→M																		
20 News	82.4	**+17.6**	**+17.6**	75.2	**+24.8**	**+24.8**	76.4	**+23.6**	**+23.6**	88.2	**+11.8**	**+11.8**	88.4	**+11.6**	**+11.6**	87.6	**+12.4**	**+12.4**
AG News	88.3	**+11.7**	+6.7	83.0	**+17.0**	**+14.5**	76.2	**+23.8**	**+13.0**	91.6	+7.5	+4.2	88.7	**+11.3**	+8.8	88.5	**+11.5**	+7.3
Financial	80.9	**+18.0**	**+19.1**	80.7	−7.3	−1.8	67.4	**−34.0**	**−34.0**	86.2	**+11.6**	+6.0	79.0	**+21.0**	**+15.4**	72.9	**+10.4**	+3.8
PubMed	70.7	**+28.6**	**+28.6**	66.1	**+33.2**	**+27.9**	43.9	**+31.4**	**+26.8**	74.9	+7.8	**+15.1**	57.5	**+29.2**	−1.5	69.1	−3.8	−5.8
ASRS	67.4	**+32.6**	**+23.3**	62.8	**+20.6**	−0.8	62.2	+3.1	−1.6	68.7	**+19.3**	**+20.0**	63.9	**+34.8**	**+16.1**	41.2	**+17.5**	**+14.1**
FDA MAUDE	96.1	+3.9	+0.6	97.4	+2.6	−0.7	92.4	+4.3	−0.7	97.8	+2.2	+2.2	92.2	+7.8	+6.1	91.5	+8.5	−3.2
MaintIE	78.3	+3.9	−7.6	78.3	+0.6	−9.0	73.1	**−30.9**	**−27.8**	70.8	**−37.4**	**−30.8**	80.2	+3.1	**+11.8**	75.5	+2.3	+7.2
M→Q																		
20 News	82.4	**+17.6**	**+15.6**	75.2	**+24.8**	**+22.1**	76.4	**+23.6**	**+19.6**	88.2	**+11.8**	**+11.1**	88.4	**+11.6**	**+10.9**	87.6	**+12.4**	**+11.8**
AG News	88.3	+5.9	−0.8	83.0	**+15.4**	+8.7	76.2	−6.2	+2.2	91.6	**−29.1**	+1.7	88.7	+8.0	−1.2	88.5	**−29.4**	+0.6
Financial	80.9	**+16.9**	**+11.4**	80.7	**−11.8**	−6.2	67.4	**−34.0**	**−29.6**	86.2	**+13.8**	+0.5	79.0	**+21.0**	+8.8	72.9	**+11.5**	−0.7
PubMed	70.7	+6.0	**+17.3**	66.1	**+13.9**	**+11.9**	43.9	**+22.1**	**+13.4**	74.9	**−24.9**	**+13.8**	57.5	+9.8	−7.5	69.1	**−33.8**	+0.2
ASRS	67.4	**+26.6**	+4.6	62.8	**+15.2**	−8.1	62.2	+9.8	−0.2	68.7	**+12.6**	+7.3	63.9	**+18.8**	−0.5	41.2	**+17.5**	**+17.5**
FDA MAUDE	96.1	+3.9	+2.2	97.4	+2.6	+2.6	92.4	+7.6	−0.7	97.8	+2.2	−2.8	92.2	+7.8	−2.2	91.5	**−39.8**	+0.2
MaintIE	78.3	**−11.6**	**−26.3**	78.3	**−11.6**	**−22.3**	73.1	**−32.0**	**−29.1**	70.8	**−37.4**	**−30.8**	80.2	**−14.6**	**−10.9**	75.5	**+14.5**	+9.9

Table 13: Cross-LLM transfer under `minimal` prompt mode. Notation **A→B**: A produced the guidance (§4.3), B produced the documents (§5.3). Q = Qwen, M = Mistral. Format as in Table 2.

Corpus. The corpus comprises 285 internal quality review meeting minutes from a regulated manufacturer. Each record describes a production incident, the root-cause investigation, and corrective actions taken, written in the organisation's working language by report authors of varying experience. Document length is typically 100–500 tokens. The 7 supervision categories used by COMPASS are phenomenon-based, e.g. data evaluation errors, item registration, temperature control issues.

Pre-training anonymisation. Because the corpus contains commercially sensitive entities (customer names, equipment identifiers, product codes, lot numbers, locations, personal names, dates), we anonymise all training and held-out text *before* classifier fitting using a regex + NER pipeline (with optional multilingual GLiNER and Presidio backends, configured per language). Each entity class is replaced by a typed placeholder token (`<CUSTOMER>`, `<EQUIP>`, `<PRODUCT>`, `<LOT_ID>`, `<DATE>`, `<PERSON>`, `<SIZE>` for physical measurements, etc.). This ensures the classifier learns phenomenon-level vocabulary rather than spurious entity correlations, and that the extracted COMPASS profiles can be safely shared with evaluators.

Guidance generation. COMPASS guidance is produced for the 7 categories using Qwen 2.5 7B-Instruct (served via vLLM) on anonymised classifier profiles, in a 5–7 bullet "tips" format. The evaluator-facing document prepends two reference tables—a placeholder glossary (mapping each `<TOKEN>` to its semantic class) and a category-name table (mapping the English class identifier to a local-language label)—so that evaluators can interpret anonymised tokens fluently.

Evaluation protocol. Each of the 9 practitioners read the full guidance document (all 7 categories together) and rated the guidance *as a whole*—not per category—on the six Likert dimensions defined in Table 6, using a 1–5 scale (1=strongly disagree, 5=strongly agree). Open-text responses on strengths, weaknesses, and suggested improvements were collected and thematically coded. The form was administered via Microsoft Forms and took approximately 10 minutes on average per respondent. Participation was voluntary; no compensation was offered. Participants were greeted with

Dataset	LR			RF			SVC			DstlB			SBERT			SetFit		
	O	ΔG	ΔR	O	ΔG	ΔR	O	ΔG	ΔR	O	ΔG	ΔR	O	ΔG	ΔR	O	ΔG	ΔR
Q→Q																		
20 News	82.4	**+17.6**	**+17.6**	75.2	**+24.8**	**+24.8**	76.4	**+23.6**	**+22.2**	88.2	**+11.8**	**+11.8**	88.4	**+11.6**	**+11.6**	87.6	**+12.4**	**+12.4**
AG News	88.3	**+10.9**	+0.0	83.0	**+16.2**	−0.5	76.2	**+23.0**	+1.3	91.6	+0.9	−2.5	88.7	**+10.5**	+0.5	88.5	+9.8	−0.2
Financial	80.9	**+19.1**	**+12.5**	80.7	**+19.3**	+9.3	67.4	**−32.9**	**−28.5**	86.2	**+13.8**	−2.8	79.0	**+19.9**	+9.9	72.9	**+23.8**	−1.8
PubMed	70.7	+10.0	**+17.3**	66.1	**+14.6**	**+21.2**	43.9	**+20.8**	**+14.8**	74.9	+7.8	**+14.4**	57.5	**−12.2**	+0.5	69.1	**−12.4**	+2.9
ASRS	67.4	**+12.6**	**+18.6**	62.8	−6.1	+1.2	62.2	−6.2	−0.2	68.7	**+26.0**	**+16.6**	63.9	**+34.1**	+5.5	41.2	**+15.5**	**+19.5**
FDA MAUDE	96.1	**−19.4**	+3.9	97.4	**−35.7**	+2.6	92.4	**−27.4**	+0.9	97.8	**−26.1**	+2.2	92.2	**−38.9**	+1.1	91.5	**−36.5**	+5.2
MaintIE	78.3	**−11.6**	**−14.3**	78.3	**−27.2**	**−14.3**	73.1	**−38.7**	**−25.1**	70.8	**−37.4**	**−30.8**	80.2	**−32.4**	−9.5	75.5	**−16.6**	−0.8
M→M																		
20 News	82.4	**+17.6**	**+17.6**	75.2	**+24.8**	**+24.8**	76.4	**+21.6**	**+22.9**	88.2	**+11.8**	**+11.8**	88.4	**+11.6**	**+11.6**	87.6	**+12.4**	**+12.4**
AG News	88.3	+3.4	+2.5	83.0	+9.5	+8.7	76.2	**+20.5**	**+10.5**	91.6	**−13.3**	−2.5	88.7	+0.5	+6.3	88.5	−3.5	+4.0
Financial	80.9	**+10.2**	**+16.9**	80.7	−8.4	−8.4	67.4	**−31.8**	**−22.9**	86.2	+4.9	**+11.6**	79.0	+4.3	**+18.8**	72.9	−4.0	+8.2
PubMed	70.7	+2.0	+7.3	66.1	+7.2	+7.2	43.9	**+22.8**	**+25.4**	74.9	−0.2	+0.4	57.5	**−28.8**	−7.5	69.1	**−31.1**	**−29.8**
ASRS	67.4	**+15.9**	+6.6	62.8	**+11.9**	**−11.4**	62.2	+7.1	**−10.9**	68.7	+9.3	+5.3	63.9	**+25.5**	−5.2	41.2	**+12.8**	+0.1
FDA MAUDE	96.1	−9.4	+2.2	97.4	−9.1	+0.9	92.4	**−20.7**	−2.4	97.8	+0.5	+2.2	92.2	−3.9	+4.5	91.5	**−34.8**	**−29.8**
MaintIE	78.3	**−13.9**	−6.3	78.3	**−30.5**	**−17.0**	73.1	**−38.7**	**−31.8**	70.8	**−37.4**	**−30.8**	80.2	**−11.3**	**−16.2**	75.5	−2.1	**−10.1**
Q→M																		
20 News	82.4	**+17.6**	**+17.0**	75.2	**+24.8**	**+24.8**	76.4	**+20.9**	**+21.6**	88.2	**+11.8**	**+11.8**	88.4	**+11.6**	**+11.6**	87.6	**+12.4**	**+11.8**
AG News	88.3	**+10.9**	+9.2	83.0	**+15.4**	**+15.4**	76.2	**+22.2**	**+16.3**	91.6	+0.9	+4.2	88.7	+9.6	+8.0	88.5	+7.3	+9.0
Financial	80.9	−7.5	**+10.2**	80.7	−7.3	−5.1	67.4	**−31.8**	**−34.0**	86.2	**−14.0**	−0.6	79.0	−4.6	**+12.1**	72.9	−7.4	−2.9
PubMed	70.7	**+16.6**	**+22.6**	66.1	**+21.9**	**+26.6**	43.9	**+25.4**	**+29.4**	74.9	**+11.8**	**+12.4**	57.5	−8.2	−1.5	69.1	−8.4	**−10.4**
ASRS	67.4	**+29.9**	**+21.9**	62.8	**+15.2**	−6.1	62.2	+3.1	−3.6	68.7	**+10.6**	**+14.6**	63.9	**+35.5**	+9.5	41.2	**+15.5**	**+11.5**
FDA MAUDE	96.1	**−14.4**	+2.2	97.4	**−10.7**	−0.7	92.4	**−20.7**	−4.1	97.8	−2.8	+2.2	92.2	−8.9	+2.8	91.5	**−13.2**	**−16.5**
MaintIE	78.3	+0.6	+4.4	78.3	**−31.6**	−1.0	73.1	**−39.8**	**−29.1**	70.8	**−37.4**	**−30.8**	80.2	**+12.0**	+1.1	75.5	−4.4	+1.9
M→Q																		
20 News	82.4	**+17.6**	**+17.0**	75.2	**+24.8**	**+21.5**	76.4	**+23.6**	**+20.9**	88.2	**+11.8**	**+11.8**	88.4	**+11.6**	**+10.9**	87.6	**+12.4**	**+11.1**
AG News	88.3	+2.5	+5.0	83.0	**+13.7**	+9.5	76.2	**+23.0**	+8.8	91.6	**−16.6**	+3.4	88.7	+3.0	+5.5	88.5	−8.5	+6.5
Financial	80.9	**+19.1**	**+10.2**	80.7	−2.9	−1.8	67.4	**−20.7**	**−19.6**	86.2	**+13.8**	+6.0	79.0	**+17.6**	+7.6	72.9	+2.6	+1.5
PubMed	70.7	−4.0	+2.6	66.1	−1.4	−6.1	43.9	+5.4	**+14.8**	74.9	+7.1	+7.8	57.5	**−35.5**	**−14.8**	69.1	**−25.1**	**−11.1**
ASRS	67.4	**+16.6**	+5.3	62.8	+1.2	−8.1	62.2	+9.1	−3.6	68.7	**+16.6**	+3.3	63.9	**+15.5**	+4.8	41.2	**+26.1**	**+10.8**
FDA MAUDE	96.1	+2.2	+3.9	97.4	+0.9	+2.6	92.4	+5.9	−2.4	97.8	+0.5	+0.5	92.2	+6.1	−2.2	91.5	**−41.5**	−8.2
MaintIE	78.3	**−20.5**	**−23.6**	78.3	**−40.5**	**−19.6**	73.1	**−39.8**	**−25.1**	70.8	**−37.4**	**−30.8**	80.2	**−16.9**	**−13.5**	75.5	−8.8	+5.9

Table 14: Cross-LLM transfer under `full` prompt mode. Notation **A→B**: A produced the guidance (§4.3), B produced the documents (§5.3). Q = Qwen, M = Mistral. Format as in Table 2.

Dataset	ROUGE-L
20 Newsgroups	0.222
AG News	0.362
Financial	0.279
PubMed RCT	0.354
ASRS	0.369
FDA MAUDE	0.420
MaintIE	0.341

Table 15: Rewritten ROUGE-L by dataset, two-LLM average.

the following preamble before the Likert items, translated and neutralised here: "Thank you for your cooperation. In this survey, you will evaluate the writing guidance for internal quality review minutes from several perspectives. The guidance was produced by AI analysis of past internal quality review minutes. Estimated time: 5–10 minutes. Responses are used only for aggregation; individual responses will not be published."

Statistical analysis. We report per-dimension means, standard deviations, and the share of positive responses (rating 4 or 5). For experience-stratified comparison we use Welch's t-test (unequal variance) and Hedges' g (small-sample-corrected Cohen's d) between the novice (n=4, <1 yr) and senior (n=3, 5+ yr) groups; the intermediate group (n=2, 1–3 yr; Total 20.5/30, per-dimension means 2.50–4.00) is reported descriptively but excluded from the inferential test due to its small size. Internal consistency across the six dimensions is measured by Cronbach's α.

Qualitative findings. Thematic coding of open-text responses surfaced three recurring positive themes and three negative themes. **Positive:** (i) the placeholder-glossary + per-category example structure was rated easy to follow; (ii) respondents endorsed the underlying initiative of standardising report writing through guidance; (iii) the guidance was seen as a useful scaffold for future LLM-assisted authoring. **Negative:** (i) anonymisation placeholders disrupt narrative flow, especially in numeric contexts—open-text responses noted that "post-editing recovers fluency without semantic loss"; (ii) two of the seven categories (those with the smallest cluster sizes) had structurally heterogeneous guidance, reducing perceived within-document consistency; (iii) one senior respondent

reported a perceived gap between anonymised content and lived experience of real incidents.

Threats to validity. Five caveats apply: (i) the sample is small (n=9) and recruited from a single organisation; (ii) the experience distribution is unbalanced (4/2/3); (iii) placeholder tokens degrade surface readability and likely depress the Readability score, in line with the open-text feedback; (iv) two of seven categories had structurally heterogeneous guidance, which we treat as a methodological limit on COMPASS when consensus clusters are very small; (v) we evaluate one length variant of the guidance (`tips`) and one LLM (Qwen 2.5 7B), so effects across LLMs and guidance lengths are not isolated in this study.

H Additional UMAP Visualizations

All UMAP projections use Sentence Transformer embeddings (all-MiniLM-L6-v2) with n=30 documents per class per condition. Markers: training data (small dots), Original (□ hollow), Guided (☆ star), Rewritten (■ filled). Class identity is encoded by colour (see legend in each subfigure).

I Computational Budget and Infrastructure

Hardware. All experiments were run on a single Linux server with two NVIDIA A100 80 GB PCIe GPUs, an AMD EPYC 7763 64-core CPU, and 123 GB of system RAM. Long-running LLM workloads use vLLM (Kwon et al., 2023) for batched inference.

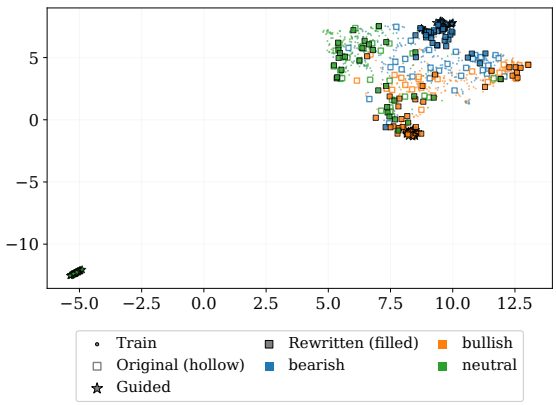
Model sizes. Parameter counts for every model in the COMPASS pipeline and held-out evaluation:

- **Classical ML classifiers** (TF-IDF + LogReg, TF-IDF + Random Forest, LDA + LinearSVC): each $<$1 M parameters.
- **DistilBERT base-uncased** (fine-tuned per dataset): 66 M parameters.
- **Sentence-Transformer all-MiniLM-L6-v2 + LogReg head**: 22 M (encoder) + head.
- **SetFit** (all-MiniLM-L6-v2 base + LogReg head): 22 M + head, few-shot fine-tuned.
- **Stage 3 / document-generation LLMs**: Qwen 2.5 7B Instruct (7.6 B) and Mistral 7B Instruct v0.3 (7.3 B), served via vLLM.
- **Held-out classifiers**: XGBoost on TF-IDF (a few hundred trees); Multinomial Naive Bayes (a few thousand features); zero-shot NLI with DeBERTa-v3 large (435 M).

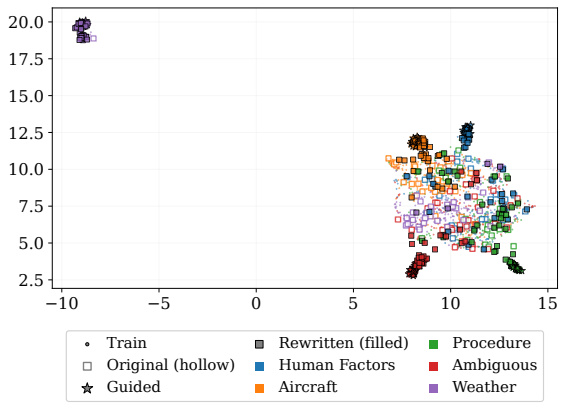
Compute budget. Approximate wall-clock costs on the hardware above:

- **Phase A** (6 classifiers $\times$ 7 datasets, dominated by DistilBERT and SetFit fine-tuning): $\sim$6–8 GPU hours.
- **Phase B** (Stage 3 guidance generation, 7 classes $\times$ 2 LLMs per dataset): $<$1 GPU hour.
- **Phase C** (document generation + classification across 7 datasets; 8 prompt modes for each diagonal LLM pairing plus 2 modes (`full`, `minimal`) for each off-diagonal pairing; n=30 documents/class): per-evaluation wall-clock ranges from 2–5 min on MaintIE to 30–40 min on ASRS. Aggregate $\sim$60–80 GPU hours, including stable-only re-runs.
- **Held-out classifier inference**: XGBoost / Naive Bayes are sub-minute; zero-shot NLI inference with DeBERTa-v3 large totals $\sim$1–2 GPU hours.
- **Anti-Gaming KD analysis, Cross-LLM tables, UMAP visualisations**: $<$2 GPU hours combined.

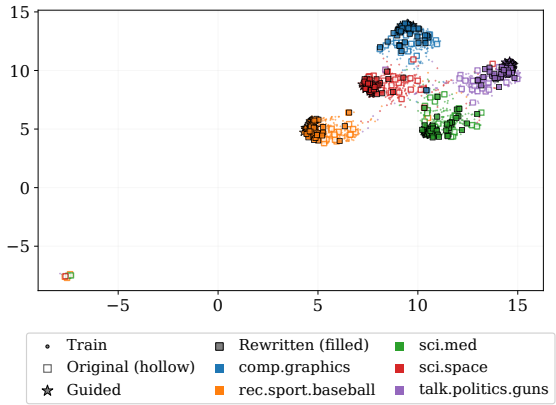
Total reported budget: approximately 70–90 GPU hours on the A100 80 GB hardware. Exploratory and deprecated runs not used in the paper are not included.



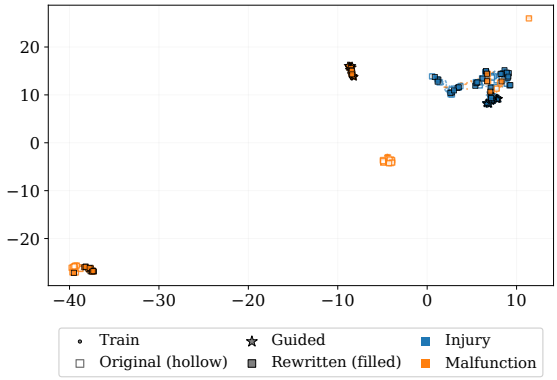
(a) Financial Sentiment.



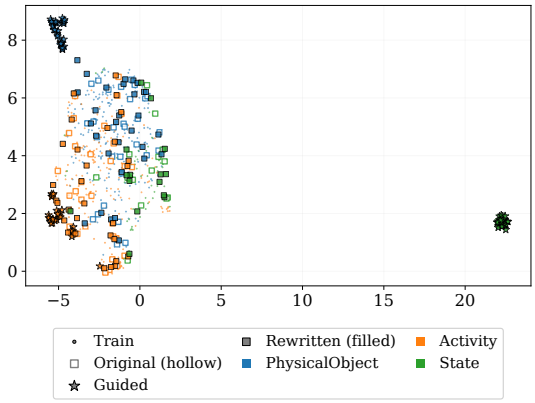
(b) ASRS.



(c) 20 Newsgroups.



(d) FDA MAUDE.



(e) MaintIE.

Figure 5: UMAP projections for the remaining five datasets, under Sentence Transformer embeddings (all-MiniLM-L6-v2), $n = 30$ per class per condition. Markers as in Figure 4.