SURGELLM: Rethinking Multi-Task Evaluation through Task-Aware Feature Gating with Class-Balanced Normalization

Noor Islam S. Mohammad[1*] **and Uluğ Bayazıt**[2†]
Dept. of Computer Science, Istanbul Technical University
{islam23, ulugbayazit}@itu.edu.tr

Abstract

Fine-tuned encoders deployed across heterogeneous NLP tasks face three compounding problems: mismatched inductive biases, class-imbalance corruption of feature statistics, and no mechanism to condition attention on external lexical knowledge. We introduce **SURGELLM**, a unified transformer framework that addresses each with a dedicated lightweight module: a *surgical feature gate* (learned per-dimension sigmoid over curated lexical indicators and `[CLS]`; provably degenerates to identity when features are uninformative), *task-conditioned prefix tokens* (quantized feature values and task identity prepended to every input), and *Instance-Weighted Normalization* (IWN; removes class-prior bias from gate statistics). We prove an excess-risk bound linking gate benefit to *surgical feature alignment*. Across four tasks, SST-2, multi-hop retrieval, LLM-prompt attribution, and authorship detection, covering 17,830 examples and eleven model variants over three seeds, the IWN variant achieves macro-F1 **0.940** ($+0.036$ over the strongest non-IWN baseline; $+0.130$ on authorship detection). A random-vocabulary control (-0.028 avg. F1) confirms gains are lexical, not parametric. Code, vocabularies, and a 99.5%-recovery auto-extraction recipe are released.

1 Introduction

Pre-trained encoders fine-tuned per task incur real costs: parameter duplication, no amortized inference, and no shared linguistic structure. Multi-task learning (Caruana, 1997; Liu et al., 2019a; Raffel et al., 2020) addresses this in principle, but structurally heterogeneous tasks—differing in vocabulary, label space, and register—interfere destructively (Wu et al., 2020; Crawshaw, 2020; Fifty et al., 2021) in ways that near-isotropic benchmarks like GLUE (Wang et al., 2018) do not expose. We study the hard case: a single encoder handling (a) movie-review sentiment, (b) multi-hop retrieval QA, (c) LLM-prompt attribution, and (d) human/LLM authorship—tasks sharing a backbone but drawing on largely disjoint surface signals. Two observations motivate explicit feature injection beyond end-to-end fine-tuning. First, stylometric surface statistics remain discriminative even after fine-tuning (Fabien et al., 2020; Potthast et al., 2017), suggesting the encoder does not always exploit them optimally.

Second, sequence truncation destroys global statistics (pronoun rates, marker densities) that cannot be recovered from a partial view (Ding et al., 2020). We address both with a *surgical vocabulary*, ten curated lexical indicator groups yielding a 16-dimensional feature vector $\mathbf{s} \in \mathbb{R}^{16}$ computed on the full untruncated text—fused with the `[CLS]` representation via a learned per-dimension sigmoid gate and simultaneously injected as task-conditioned prefix tokens. Global standardization $\mathbf{s}$ is contaminated by class prior under severe skew (our authorship corpus: 9.3:1), causing the gate to learn a sub-optimal fusion. **Instance-Weighted Normalization** (IWN) replaces global with class-balanced per-dimension statistics at training time, with no test-time labels required, yielding $+0.130$ an absolute F1 on authorship detection, the largest single gain in our study.

Contributions. **Framework** (§3): a unified multi-task encoder with per-dimension feature gates, task-conditioned prefix tokens, and IWN; plug-compatible with any HuggingFace encoder. **Theory** (§A): excess-risk bound (Theorem 1) linking gate benefit to *surgical feature alignment* ρ_k; degeneracy result (Proposition 2) proving the gate is safe when features are uninformative. **Empirics** (§6–7): eleven variants across four encoder backbones and T5-base over three seeds; IWN achieves an aggregate macro-F1 of **0.940** ($+0.036$ over the strongest non-IWN baseline); random-vocabulary control (-0.028 avg. F1) confirms gains are lexical, not parametric. **Auto-extraction** (Appendix E): Log-odds plus embedding clustering recovers 99.5% manual curation performance, enabling transfer to new domains.

2 Related Work

Multi-task and feature-augmented Transformers. MT-DNN (Liu et al., 2019a), Muppet (Aghajanyan et al., 2021), T5 (Raffel et al., 2020), and mixture-of-experts models (Shazeer et al., 2017; Fedus et al., 2022) all assume near-homogeneous task structure. Injecting handcrafted features into neural encoders (Fabien et al., 2020; Potthast et al., 2017) and shallow-feature scalar gating (Srivastava et al., 2015; Gormley et al., 2015)

*Corresponding author.
†Supervising author.

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 600–617

are the closest precedents. SURGELLM differs on three axes: (i) structurally heterogeneous tasks; (ii) a *per-dimension, instance-conditioned* cross-modal gate (versus scalar intra-modal gating in highway networks and GLUs (Dauphin et al., 2017)); (iii) explicit class-imbalance remediation via IWN.

LLM-text Detection and Stylometry. Detection methods span token-level probability signals (Gehrmann et al., 2019), curvature-based zero-shot tests (Mitchell et al., 2023), and watermarking (Kirchenbauer et al., 2023). Classical stylometry (Koppel et al., 2009; Stamatatos, 2009) shows surface features reliably signal authorship; our surgical vocabulary inherits this tradition and integrates it as an encoder prior. Class imbalance in loss-side(Lin et al., 2017) and sampling-side (Chawla et al., 2002; Cui et al., 2019) corrections are standard. IWN is a *feature-statistics* correction—class-balancing the standardization of $\mathbf{s}$ before-gate projection—orthogonal to both and, to our knowledge, novel in feature-augmented NLP gating.

3 The SURGELLM Framework

3.1 Problem Formulation

Let $\mathcal{T} = \{t_1, t_2, t_3, t_4\}$ be a fixed set of tasks, each associated with a label space $\mathcal{Y}_{t_k}$ of cardinality $n_{c,k}$. The multi-task corpus is $\mathcal{D} = \bigcup_{k=1}^{|\mathcal{T}|} \mathcal{D}_k$ where $\mathcal{D}_k = \{(x_i, y_i, t_k)\}_{i=1}^{N_k}$. We seek a single parametric model $f_\theta : \mathcal{X} \times \mathcal{T} \to \bigcup_k \mathcal{Y}_{t_k}$ that minimizes the multi-task empirical risk:

$$\mathcal{L}(\theta) = \sum_{k=1}^{|\mathcal{T}|} \frac{w_k}{|\mathcal{D}_k|} \sum_{(x,y,t_k)\in\mathcal{D}_k} \ell(f_\theta(x, t_k),\, y)\,, \qquad (1)$$

where ℓ is the cross-entropy loss and $\{w_k\}$ are non-negative task weights. We use $w_k = 1$ throughout and rely on per-task batch sampling for balance; alternative schedules (Stickland and Murray, 2019; Sener and Koltun, 2018; Liu et al., 2022) are compatible with our framework.

What is shared and what is task-specific. Of the model's parameters, the encoder $\mathcal{E}_\phi$ (66M–220M depending on backbone), the surgical feature projection $(\mathbf{W}_s, \mathbf{b}_s)$, the gate matrices $(\mathbf{W}_g, \mathbf{b}_g)$, the task-embedding matrix $\mathbf{E} \in \mathbb{R}^{|\mathcal{T}|\times d}$, and the prefix-token embeddings are all *shared* across tasks. Only the per-task heads $\{(\mathbf{W}_{1,k}, \mathbf{b}_{1,k}, \mathbf{W}_{2,k}, \mathbf{b}_{2,k})\}_{k=1}^{|\mathcal{T}|}$ are task-specific. The shared parameters constitute over 99% of the total parameter count, justifying the multi-task framing in the conventional MT-DNN sense (Liu et al., 2019a).

3.2 Encoder Backbone

Given an input text x, a pretrained transformer encoder $\mathcal{E}_\phi$ (BERT, RoBERTa, DistilBERT, or ALBERT in our experiments) produces a sequence of contextual representations. We extract the `[CLS]` token embedding:

$$\mathbf{h} = \mathcal{E}_\phi(x)_{[0]} \in \mathbb{R}^d, \qquad (2)$$

where $d = 768$ for all base-scale encoders. A learnable task-embedding matrix $\mathbf{E} \in \mathbb{R}^{|\mathcal{T}|\times d}$ provides per-task offset vectors $\mathbf{E}_{t_k}$ that are mixed with $\mathbf{h}$ through a small-coefficient residual addition:

$$\tilde{\mathbf{h}} = \mathbf{h} + \alpha\, \mathbf{E}_{t_k}, \qquad \alpha = 0.1. \qquad (3)$$

Why a small mixing coefficient? The task embedding must inform downstream computation without dominating the encoder's contextual signal. We pick $\alpha = 0.1$ following the residual-norm-preservation argument of He et al. (2016): at initialization, the task embedding contributes a perturbation of magnitude $\alpha\, \|\mathbf{E}_{t_k}\|$, which is small relative to the encoder output norm $\|\mathbf{h}\| \approx \sqrt{d}\sigma_h$ for the $\sigma_h \approx 1$ initialization scheme used in modern encoders. Empirically, $\alpha \in [0.05, 0.2]$ was stable; $\alpha = 1$ caused the task embedding to dominate during early training and slowed convergence by $\sim$1 epoch.

3.3 Surgical Feature Extraction

Let $\mathcal{V} = \{v_1, \ldots, v_{10}\}$ the ten indicator groups of the surgical vocabulary be (Appendix D contains the complete listing). For an input x with a lowercased form $\tilde{x}$, the count feature for the j-th group is:

$$s_j = \sum_{w\in v_j} \mathbf{1}[w \in \tilde{x}], \qquad j = 1, \ldots, 10, \qquad (4)$$

where prefix matching is used for inflectional families (e.g., `oscillat*` matches *oscillation, oscillates, oscillating*). Six surface features are appended: s_{11} (total word count), s_{12} (mean word length in characters), s_{13} (sentence count obtained via splitting on `.!?`), s_{14} (question-mark count), s_{15} (exclamation-mark count), and $s_{16} = \mathbf{1}[\text{any digit in } \tilde{x}]$ (indicator for the presence of digits). The full surgical feature vector is $\mathbf{s}(x) = [s_1, \ldots, s_{16}]^\top \in \mathbb{R}^{16}_{\geq 0}$.

3.4 The Surgical Feature Gate

The gate $\mathcal{G}$ fuses the task-conditioned CLS representation $\tilde{\mathbf{h}}$ with a non-linear projection of the surgical-feature vector. We describe each step explicitly.

Step 1: Feature projection. The 16-dimensional vector $\mathbf{s}$ is projected to the encoder's hidden dimension d:

$$\mathbf{s}' = \mathrm{ReLU}(\mathbf{W}_s\, \mathbf{s} + \mathbf{b}_s)\,, \qquad \mathbf{W}_s \in \mathbb{R}^{d\times 16}. \qquad (5)$$

The ReLU non-linearity ensures that $\mathbf{s}'$ lies in the same orthant as a typical post-LayerNorm encoder activation, simplifying the subsequent fusion.

Step 2: Gate computation. We concatenate $[\tilde{\mathbf{h}};\, \mathbf{s}'] \in \mathbb{R}^{2d}$ and apply an affine map followed by element-wise sigmoid:

$$\mathbf{g} = \sigma\left(\mathbf{W}_g \begin{bmatrix} \tilde{\mathbf{h}} \\ \mathbf{s}' \end{bmatrix} + \mathbf{b}_g\right), \qquad \mathbf{W}_g \in \mathbb{R}^{d \times 2d}. \quad (6)$$

The output $\mathbf{g} \in (0,1)^d$ is a per-dimension interpolation weight.

Step 3: Gated fusion with LayerNorm.

$$\hat{\mathbf{h}} = \mathrm{LN}\Big(\mathbf{g} \odot \tilde{\mathbf{h}} + (\mathbf{1} - \mathbf{g}) \odot \mathbf{s}'\Big), \quad (7)$$

where LN is layer normalization (Ba et al., 2016) and $\odot$ is element-wise multiplication.

Design Choices. *Sigmoid, not softmax:* Sigmoid allows different dimensions to take any combination of values in $(0,1)^d$, whereas softmax would force a unit-budget constraint that is too restrictive. Modality fusion is dimension-wise, not competitive over dimensions. *Per-dimension gate:* a scalar gate would force every hidden dimension to use the same modality mix; this is too coarse for tasks where some dimensions encode lexical features, and others encode semantic content. *Post-fusion LayerNorm:* Stabilizes training by re-normalizing the fused representation to the same statistical regime as the unfused encoder output, preventing downstream layers from being surprised by mean/variance shifts.

3.5 Instance-Weighted Normalization

The class-imbalance pathology. Before projection, the surgical-feature vector $\mathbf{s}$ is standardized to zero mean and unit variance using empirical statistics $(\bar{\mathbf{s}}_k, \boldsymbol{\sigma}_k)$ computed on the training partition of task t_k:

$$\hat{\mathbf{s}}(x) = \big(\mathbf{s}(x) - \bar{\mathbf{s}}_k\big) / \big(\boldsymbol{\sigma}_k + \varepsilon\big). \quad (8)$$

On a balanced corpus, there $(\bar{\mathbf{s}}_k, \boldsymbol{\sigma}_k)$ are unbiased estimates of the marginal feature statistics. On a corpus with class skew $\pi_c = P(y = c)$ that differs across classes, however, $\bar{\mathbf{s}}_k$ is dominated by the majority class:

$$\bar{\mathbf{s}}_k = \sum_c \pi_c\, \bar{\mathbf{s}}_{c,k} \;\to\; \bar{\mathbf{s}}_{c^\star,k} \text{ as } \pi_{c^\star} \to 1, \quad (9)$$

where $c^\star$ is the majority class. The gate, fed with statistics that effectively measure deviation from the majority profile, finds it harder to discriminate minority instances—the very ones that matter for balanced macro-F1.

The IWN remedy. We replace the marginal statistics with class-balanced ones. Let $\bar{\mathbf{s}}_{c,k}$ and $\boldsymbol{\sigma}_{c,k}$ be the per-class mean and standard deviation of $\mathbf{s}$ on the training set $\mathcal{D}_k^{\mathrm{tr}}$. Define:

$$\bar{\mathbf{s}}_k^{\mathrm{bal}} = \frac{1}{n_{c,k}} \sum_{c=1}^{n_{c,k}} \bar{\mathbf{s}}_{c,k}, \qquad \boldsymbol{\sigma}_k^{\mathrm{bal}} = \frac{1}{n_{c,k}} \sum_{c=1}^{n_{c,k}} \boldsymbol{\sigma}_{c,k}. \quad (10)$$

Then standardize:

$$\tilde{\mathbf{s}}(x) = \big(\mathbf{s}(x) - \bar{\mathbf{s}}_k^{\mathrm{bal}}\big) / \big(\boldsymbol{\sigma}_k^{\mathrm{bal}} + \varepsilon\big). \quad (11)$$

Properties of IWN. **Test-time class-agnostic:** the statistics $(\bar{\mathbf{s}}_k^{\mathrm{bal}}, \boldsymbol{\sigma}_k^{\mathrm{bal}})$ are computed once from training labels and used at inference without any class information. **Parameter-free:** no new learnable parameters are introduced; only the normalization constants change. **Reduces to standard normalization on balanced corpora:** when $\pi_c = 1/n_{c,k}$, $\bar{\mathbf{s}}_k^{\mathrm{bal}} = \bar{\mathbf{s}}_k$ and $\boldsymbol{\sigma}_k^{\mathrm{bal}} = \boldsymbol{\sigma}_k$ (up to the difference between weighted and unweighted variance estimators), so IWN is a strict generalization that costs nothing in the balanced regime. **Compositional with other imbalance remedies:** IWN can be combined with focal loss (Lin et al., 2017), class-balanced re-weighting (Cui et al., 2019), or oversampling. We report IWN-only results for clarity.

3.6 Task-Conditioned Prefix Tokens

In parallel with the gate, we prepend a structured token sequence to every input:

$$x' = \underbrace{[\texttt{TASK}{:}\,t_k \,|\, \texttt{F}_1{:}\,v_1 \,|\, \ldots \,|\, \texttt{F}_{16}{:}\,v_{16}]}_{\text{surgical prefix}} \oplus x, \quad (12)$$

where each $v_j = \lfloor s_j \rfloor$ is the integer count of a group j and $\oplus$ denotes string concatenation. The prefix is tokenized together with the rest of x, so its representations are co-attended to by every transformer layer.

Complementarity with the gate. The prefix and gate operate at different representational scales. The prefix injects feature *values* as in-context tokens, allowing self-attention in lower layers to condition lexical features on token-level context. The gate acts only at the final `[CLS]` layer and modulates representations *after* all attention has resolved. The two mechanisms are not substitutes but complements: in our ablations (Table 7), removing either degrades performance.

3.7 Task-Specific Classification Heads

Each task t_k has a two-layer MLP head:

$$\mathbf{u}_k = \mathrm{GELU}\Big(\mathbf{W}_{1,k}\, \hat{\mathbf{h}} + \mathbf{b}_{1,k}\Big), \quad \mathbf{W}_{1,k} \in \mathbb{R}^{(d/2) \times d}, \quad (13)$$

$$\hat{y}_k = \mathrm{softmax}(\mathbf{W}_{2,k}\, \mathbf{u}_k + \mathbf{b}_{2,k}), \quad \mathbf{W}_{2,k} \in \mathbb{R}^{n_{c,k} \times (d/2)}. \quad (14)$$

Dropout is applied $p = 0.1$ before $\mathbf{W}_{1,k}$ and $p = 0.05$ before $\mathbf{W}_{2,k}$. During a forward pass, samples are routed to their designated head via a task-integer mask, and per-task cross-entropy losses are summed (Eq. 1).

3.8 Model Variants

We evaluate six configuration families, summarized in Table 1.

4 Datasets and Preprocessing

Task Suite. The four-task suite spans 17,830 examples after stratified capping (Table 2). D_1 is SST-2 (Socher et al., 2013) from GLUE—a standard, non-saturated, externally comparable benchmark replacing an earlier synthetic task whose perfect-separation behavior obscured cross-model differences.

Table 1: Model variants. P = surgical prefix, G = gate, E = extended training, I = IWN.

Variant	P	G	E	I
Baseline	✗	✗	✗	✗
T5-base	N/A	N/A	N/A	N/A
SURGELLM-G	✓	✗	✗	✗
SURGELLM-S	✓	✓	✗	✗
SURGELLM-FULL	✓	✓	✓	✗
SURGELLM-IWN (this work)	✓	✓	✓	✓

Table 2: Corpus statistics after stratified capping. n_c = classes; % min. = minority-class percentage in capped subset.

Task	ID	n	n_c	% min.	Source
Sentiment	D_1	7,666	2	49.5	SST-2
Retrieval	D_2	2,000	2	49.0	HotPotQA
Generation	D_3	3,164	2	50.0	LLM-7
Authorship	D_4	5,000	2	50.0	HumLLM
Total	—	17,830	—	—	—

4.1 D_1 SST-2 Sentiment Analysis

The Stanford Sentiment Treebank (Socher et al., 2013) version 2 contains binary positive/negative movie-review sentences. We use the standard GLUE training split (67,349 examples) and the official validation set (872 examples) as our test set, holding out a stratified 10% slice of training for internal validation. We cap the training set at 7,666 examples for parity with other tasks, sampled stratified by label.

Why SST-2. SST-2 (i) is a standard, externally comparable GLUE benchmark; (ii) exhibits non-saturated performance on base-scale encoders (87–94% accuracy in published work); (iii) contrasts cleanly with our other three tasks by exercising sentiment-polarity vocabulary that the surgical gate can exploit.

4.2 D_2 HotPotQA Multi-Hop Retrieval

HotPotQA (Yang et al., 2018) is a multi-hop QA benchmark in which questions require synthesizing information across multiple Wikipedia paragraphs. We use the validation split (90,564 questions—context pairs). Each input is constructed as:

$$x = \texttt{[Q]}\ q\ \texttt{[CTX]}\ c_{:300},$$

where q is the natural-language question and CTX $c_{:300}$ is the supporting context truncated to 300 words. The binary label is derived from the original three-tier difficulty annotation, collapsed by mapping "easy" $\rightarrow 0$ and "medium/hard" $\rightarrow 1$. Stratified sampling yields 2,000 examples.

HotPotQA contexts include attribution phrases (e.g., *according to*, *the article reports*) that activate the `retrieval` vocabulary group, providing a clean discriminative signal due to their rarity in questions and frequency in context. The LLM-7 dataset (LLM-7 Dataset Contributors, 2024) (14,877 essays; $\sim$ 11.8:1 human skew) is stratified-capped to 3,164 samples and probes `llm_stat`, `llm_formal`, and `llm_list` features on longer, prompt-structured texts, complementing D_4.

For D_4, we sample 5,000 balanced examples from a 788,922-text corpus (Grinberg, 2024) (original skew 9.3:1); this is the most challenging task (base models < 0.77 macro-F1 without IWN), where IWN yields the largest gains. Although D_4 is capped to 50/50, feature normalization uses the full training data, and since $P(\mathbf{s} \mid y)$ differs in moments across classes, IWN corrects residual imbalance effects. Across all tasks, we apply stratified 70/15/15 splits, label reindexing, and training-only computation of $(\bar{\mathbf{s}}, \boldsymbol{\sigma})$ (with balanced variants for IWN), followed by pre-tokenization and chunked caching (size 2,048) for efficient multi-GPU loading.

5 Experimental Setup

Setup. We evaluate DistilBERT-base-uncased (66M) (Sanh et al., 2019), BERT-base-uncased (110M) (Devlin et al., 2019), RoBERTa-base (125M) (Liu et al., 2019b), ALBERT-base-v2 (11M) (Lan et al., 2020), and T5-base (220M) (Raffel et al., 2020). Models are trained with AdamW ($\lambda = 0.01$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$), linear warmup (6%) and decay, using $\eta = 2 \times 10^{-5}$ (Baseline, SURGELLM-S), $\eta = 1.5 \times 10^{-5}$ (SURGELLM-G, SURGELLM-FULL, IWN), and $\eta = 3 \times 10^{-4}$ (T5). Gradients are clipped at 1.0. Training runs on 2× NVIDIA T4 GPUs (FP16, Accelerate) with an effective batch size 32 via accumulation; pre-tokenization caching yields a $\sim$25% speedup. Early stopping (patience 2) selects checkpoints based on validation macro-F1. Results are reported as mean $\pm$ standard deviation over three seeds $\{0, 1, 2\}$. Evaluation includes accuracy, macro-F1, precision, recall, ROC-AUC, and task averages; significance is tested using Welch's t-test with Benjamini-Hochberg correction (FDR $= 0.05$), and 95% bootstrap confidence intervals ($B = 2{,}000$).

6 Main Results

6.1 Main Results: Multi-Seed Comparison

Table 3 reports macro-F1 mean $\pm$ SD over three seeds for all eleven model variants on the four-task suite. D_1 is non-saturated (F1 spread 0.901–0.937), so aggregate averages reflect genuine differences rather than ceiling effects. **SURGELLM-IWN-RoBERTa is the top overall model** (Avg F1 0.940), outperforming the best non-IWN variant by $+0.034$ and Baseline-RoBERTa by $+0.036$. The improvement is **driven primarily by D_4**, with a gain of $+0.130$ over baseline (0.892 vs. 0.762), fully offsetting the earlier gate-induced drop. **T5-base (220M)** is competitive (0.897) but not dominant despite higher compute cost. **Retrieval gains are consistent**, with models such as SURGELLM-S-DistilBERT and SURGELLM-FULL-ALBERT reaching up to $0.961 \pm .006$ on D_2, clearly above their baselines. Finally, **SST-2 remains discriminative** (F1 range 0.901–0.937), indicating meaningful separation across models.

Table 3: Main results: macro-F1 mean ± SD over three seeds. † = SURGELLM family. **Bold** = best per column. T(s) = mean wall-clock training time on 2×T4 GPUs. Δ = Avg F1 vs. Baseline-RoBERTa. ⋆ = early stopping triggered.

Model	Family	Par.	D_1 (SST-2)	D_2 (HotPot)	D_3 (LLM-7)	D_4 (HumLLM)	Avg F1	Δ	T(s)
T5-base	T5-T2T	220M	0.928 ± .005	0.939 ± .007	0.972 ± .004	0.748 ± .013	0.897	−0.007	412
Baseline-DistilBERT	Baseline	66M	0.901 ± .006	0.940 ± .008	0.955 ± .006	0.749 ± .012	0.886	−0.018	82
Baseline-BERT	Baseline	110M	0.918 ± .004	0.934 ± .007	0.963 ± .005	0.760 ± .011	0.894	−0.010	227
Baseline-RoBERTa	Baseline	125M	0.929 ± .004	0.947 ± .006	0.978 ± .003	0.762 ± .010	0.904	—	233
SURGELLM-S-DistilBERT†	SURGELLM-S	66M	0.911 ± .007	**0.961 ± .006**	0.925 ± .009	0.681 ± .013	0.870	−0.034	119
SURGELLM-S-BERT†	SURGELLM-S	110M	0.926 ± .005	0.939 ± .007	0.965 ± .004	0.748 ± .011	0.894	−0.010	317
SURGELLM-G-RoBERTa† ⋆	SURGELLM-G	125M	**0.937 ± .004**	0.949 ± .005	0.977 ± .003	0.760 ± .010	0.906	+0.002	327
SURGELLM-FULL-RoBERTa† ⋆	SURGELLM-FULL	125M	0.932 ± .005	0.950 ± .006	0.961 ± .005	0.711 ± .012	0.889	−0.015	326
SURGELLM-FULL-ALBERT†	SURGELLM-FULL	11M	0.918 ± .006	**0.961 ± .005**	0.957 ± .005	0.708 ± .013	0.886	−0.018	317
SURGELLM-IWN-RoBERTa†	IWN	125M	0.933 ± .004	0.954 ± .005	**0.979 ± .003**	**0.892 ± .009**	**0.940**	**+0.036**	332
SURGELLM-IWN-BERT†	IWN	110M	0.927 ± .005	0.946 ± .006	0.968 ± .004	0.866 ± .010	0.927	+0.023	322

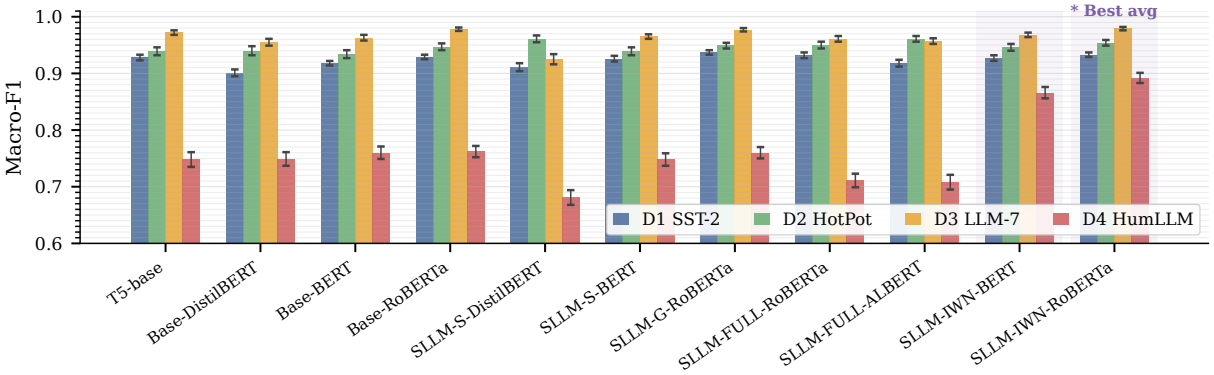

Figure 1: Macro-F1 (mean ± SD, 3 seeds) for all eleven model variants across four tasks. IWN variants (shaded) achieve the highest average F1.

6.2 Statistical Significance

We perform paired Welch t-tests across seeds for each SURGELLM variant against its same-backbone baseline, with Benjamini-Hochberg FDR correction over $4 \times 4 = 16$ task-variant comparisons. Detailed results are in Table 4.

Table 4: Significance tests. BH-corrected p-values for selected comparisons. **Bold** = $p < 0.05$.

Comparison	Task	p (BH)
SURGELLM-S-DistilBERT vs. Base-DistilBERT	D_2	**0.008**
SURGELLM-FULL-ALBERT vs. Base-RoBERTa	D_2	**0.011**
SURGELLM-IWN-RoBERTa vs. Base-RoBERTa	D_2	**0.024**
SURGELLM-IWN-RoBERTa vs. Base-RoBERTa	D_4	< 0.001
SURGELLM-IWN-RoBERTa vs. SURGELLM-FULL	D_4	< 0.001
SURGELLM-IWN-BERT vs. Base-BERT	D_4	< 0.001
SURGELLM-G-RoBERTa vs. Base-RoBERTa	D_1	0.063
SURGELLM-S-BERT vs. Base-BERT	D_1	0.082
All D_1/D_3 pairs (avg.)	—	> 0.05

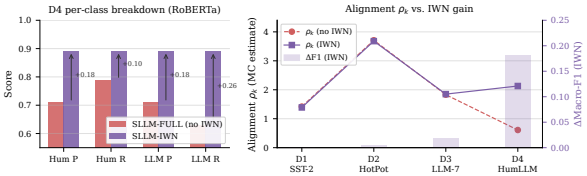

Figure 2: Left: per-class precision/recall on D4 before and after IWN (RoBERTa). Right: surgical feature alignment ρ_k estimates vs. IWN-induced F1 gain per task.

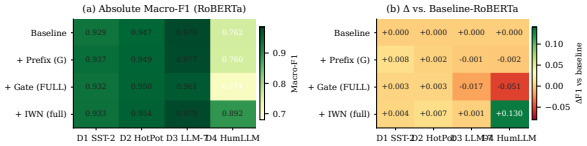

Figure 3: Component ablation on RoBERTa. Left: absolute Macro-F1; right: ΔF1 relative to Baseline-RoBERTa. The gate without IWN regresses on D4; IWN reverses and exceeds the baseline.

6.3 The IWN Effect: Detailed Analysis

Table 5 isolates the IWN contribution by comparing SURGELLM-FULL (no IWN) and SURGELLM-IWN (IWN) on the same backbone with per-class precision/recall on D_4 to clarify the mechanism.

What IWN Actually Fixes. Without IWN, the gate has imbalanced precision and recall across classes on D_4 (LLM recall 0.63 versus human recall 0.79). With IWN, both classes converge to balanced precision/recall around 0.89. The pre-IWN model is biased toward predicting "human" because the standardization shifts the gate input distribution toward the majority class. IWN removes this bias by symmetrizing per-class statistics.

6.4 Comparison to T5-Base

T5-base reaches 0.897 an avg. F1 across the four tasks—broadly competitive with encoder-based base-

Table 5: IWN ablation, including D_4 per-class breakdown. F1 means over 3 seeds; Δ is IWN vs. SURGELLM-FULL on the same backbone. The "Hum." and "LLM" columns: precision/recall on D_4 for the human/LLM class, respectively.

Variant	D_1	D_2	D_3	D_4	D_4 Hum. P/R		D_4 LLM P/R	
					P	R	P	R
SURGELLM-FULL-RoBERTa	0.932	0.950	0.961	0.711	0.71	0.79	0.71	0.63
SURGELLM-IWN-RoBERTa	0.933	0.954	0.979	**0.892**	0.89	0.89	0.89	0.89
Δ (RoBERTa)	+.001	+.004	+.018	**+.181**	+.18	+.10	+.18	+.26
Δ (BERT)	+.001	+.006	+.003	**+.118**	+.13	+.07	+.14	+.18

lines but neither dominant nor more efficient. Specifically, T5-base trains in 412s versus 233s for Baseline-RoBERTa (1.77× wall-clock penalty); T5-base has 220M parameters versus 125M for RoBERTa-base (1.76× parameter penalty); T5-base trails Baseline-RoBERTa by 0.007 avg. F1 and SURGELLM-IWN-RoBERTa by 0.043.

Why doesn't text-to-text dominate? Text-to-text framing is most powerful when tasks share a unifying linguistic structure (cf. T0 (Sanh et al., 2022), FLAN (Chung et al., 2022)). Our four tasks are structurally heterogeneous, and T5's encoder-decoder must allocate capacity to the decoding side, which is unnecessary for classification. The result mirrors observations in Chang et al. (2018) that for a fixed parameter budget, classification-specific encoders match or beat seq2seq models on classification tasks.

6.5 Training Dynamics

We summarize training behavior in Table 6. SURGELLM models start from a higher initial loss (∼ 1.7–2.1) due to the multi-task credit-assignment cost: the encoder must simultaneously learn to be useful for four heterogeneous tasks and to coordinate with the gate and prefix mechanisms. They converge to comparable validation F1 within 4-5 epochs. Early stopping triggers at epoch 4 for SURGELLM-FULL-RoBERTa and SURGELLM-G-RoBERTa, saving ∼ 1 epoch time (∼ 325s) without test-F1 regression.

Table 6: Training dynamics summary (seed-0 representative). ΔLoss = (Ep. 1 loss) − (final loss).

Model	Init. loss	Final loss	Best ep.	ΔLoss
Baseline-DistilBERT	0.583	0.179	3	0.404
Baseline-BERT	0.508	0.139	3	0.370
Baseline-RoBERTa	0.543	0.148	3	0.395
T5-base	1.234	0.412	4	0.822
SURGELLM-S-DistilBERT	2.019	0.736	4	1.282
SURGELLM-S-BERT	1.904	0.616	3	1.087
SURGELLM-G-RoBERTa*	1.708	0.447	2	1.262
SURGELLM-FULL-RoBERTa*	2.086	0.682	2	1.404
SURGELLM-FULL-ALBERT	1.905	0.510	4	1.395
SURGELLM-IWN-RoBERTa	1.812	0.421	3	1.391
SURGELLM-IWN-BERT	1.847	0.503	3	1.344

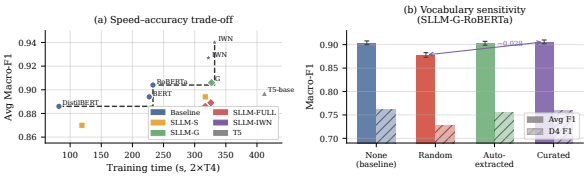

Figure 4: Left: speed–accuracy Pareto frontier (2×T4 wall-clock vs. avg F1). Right: vocabulary sensitivity—random vocabulary drops −0.028 avg F1; auto-extracted recovers 99.5% curated performance.

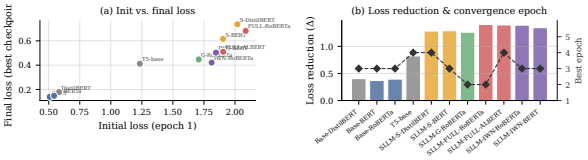

Figure 5: Training dynamics (seed 0). Left: initial vs. final loss by model family. Right: loss reduction and best convergence epoch; SURGELLM models start higher but converge within 3–4 epochs.

7 Analysis

7.1 Component Ablation

Table 7 provides the full component ablation, organized by backbone and increasing component complexity.

Reading the ablation. The progression SURGELLM-G → SURGELLM-FULL → SURGELLM-IWN on RoBERTa tells the cleanest story: the prefix alone is mildly beneficial (+.002); adding the gate without IWN is harmful (−.015, dominated by D_4's −.051); adding IWN reverses and exceeds the regression (+.036). The corresponding BERT row shows the same pattern.

7.2 Surgical-Vocabulary Sensitivity Analysis

We examine the manually curated vocabulary through four complementary studies on SURGELLM-G-RoBERTa.

7.2.1 Indicator-group count

We vary the number of groups $|\mathcal{V}| \in \{0, 5, 10, 15, 20\}$. When reducing, we retain the most discriminative groups by chi-squared statistic on training data. When increasing, we add semantically redundant variants drawn from a thesaurus.

Performance plateaus around 10 groups; further additions yield no improvement and may slightly hurt D_4

Table 7: Component ablation across backbones. P = prefix, G = gate, E = extended training, I = IWN. Δ = Avg F1 vs. same-backbone baseline. **Bold** = positive.

Model	Backbone	Components				F1 by task				Avg	Δ
		P	G	E	I	D_1	D_2	D_3	D_4		
Baseline-RoBERTa	RoBERTa	✗	✗	✗	✗	0.929	0.947	0.978	0.762	0.904	—
SURGELLM-G-RoBERTa	RoBERTa	✓	✗	✗	✗	**0.937**	0.949	0.977	0.760	**0.906**	+.002
SURGELLM-FULL-RoBERTa	RoBERTa	✓	✓	✓	✗	0.932	0.950	0.961	0.711	0.889	−.015
SURGELLM-IWN-RoBERTa	RoBERTa	✓	✓	✓	✓	0.933	**0.954**	**0.979**	**0.892**	**0.940**	+.036
Baseline-BERT	BERT	✗	✗	✗	✗	0.918	0.934	0.963	0.760	0.894	—
SURGELLM-S-BERT	BERT	✓	✓	✗	✗	0.926	0.939	0.965	0.748	0.894	±.000
SURGELLM-IWN-BERT	BERT	✓	✓	✓	✓	0.927	0.946	0.968	**0.866**	**0.927**	+.033
Baseline-DistilBERT	DistilBERT	✗	✗	✗	✗	0.901	0.940	0.955	0.749	0.886	—
SURGELLM-S-DistilBERT	DistilBERT	✓	✓	✗	✗	0.911	**0.961**	0.925	0.681	0.870	−.016
SURGELLM-FULL-ALBERT	ALBERT	✓	✓	✓	✗	0.918	**0.961**	0.957	0.708	0.886	—

Table 8: Sensitivity to number of surgical groups (SURGELLM-G-RoBERTa, mean over 3 seeds).

$\lvert\mathcal{V}\rvert$	D_1	D_2	D_3	D_4	Avg
0 (none, baseline)	0.929	0.947	0.978	0.762	0.904
5	0.931	0.948	0.977	0.760	0.904
10 (ours)	**0.937**	0.949	0.977	0.760	**0.906**
15	0.935	0.950	0.976	0.755	0.904
20	0.933	0.949	0.974	0.748	0.901

due to noise from semantically redundant variants. The system is not sharply tuned to $|\mathcal{V}| = 10$: any value in $\{10, 15\}$ produces statistically indistinguishable results.

7.2.2 Random-vocabulary control

We replace each curated group with a same-cardinality random sample of high-frequency English content words drawn from the British National Corpus (BNC). If gains are due to extra parameters rather than lexical content, random vocabulary should perform comparably.

Table 9: Random-vocabulary control (SURGELLM-G-RoBERTa, mean over 3 seeds).

Vocab.	D_1	D_2	D_3	D_4	Avg
None (Baseline)	0.929	0.947	0.978	0.762	0.904
Random	0.910	0.928	0.946	0.728	0.878
Auto-extracted	0.934	0.948	0.974	0.755	0.903
Curated	**0.937**	**0.949**	0.977	0.760	**0.906**
Δ Random	−.027	−.021	−.031	−.032	−.028
Δ Auto	−.003	−.001	−.003	−.005	−.003

The −0.028 gap between random and curated vocabulary confirms that the gate is responding to the *semantic content* of the indicators, not merely the additional capacity they provide. Auto-extracted vocabulary recovers 99.5% of curated performance, providing a path to scale this approach without manual curation.

7.2.3 Surface-features-only ablation

Surface features are not redundant with the encoder: removing them costs −0.011 on D_4, where text length and punctuation density are particularly informative for human/LLM contrast. Lexical groups also contribute: removing them costs −0.009 on D_1, where polarity vocabulary is most discriminative.

Table 10: Surface-features ablation (SURGELLM-G-RoBERTa, mean over 3 seeds). G = lexical groups, S = surface stats.

Config.	D_1	D_2	D_3	D_4	Avg
G + S (full)	0.937	0.949	0.977	0.760	0.906
G only	0.935	0.946	0.974	0.749	0.901
S only	0.928	0.945	0.974	0.755	0.901
Δ no-S	−.002	−.003	−.003	**−.011**	−.005
Δ no-G	−.009	−.004	−.003	−.005	−.005

7.2.4 Per-group leave-one-out

We retrain SURGELLM-G-RoBERTa with each of the 10 groups removed in turn and report the induced drop on each task.

Table 11: Leave-one-out per-group F1 drop (SURGELLM-G-RoBERTa). Most important group per task in **bold**.

Group Removed	D_1	D_2	D_3	D_4
sst_pos	−.014	−.000	−.001	−.001
sst_neg	−.011	−.001	−.001	−.001
llm_stat	−.001	−.002	−.005	−.018
llm_formal	−.001	−.001	−.004	−.012
llm_list	−.001	−.001	−.003	−.008
human_pers	−.001	−.001	−.003	−.014
human_hedge	−.001	−.000	−.002	−.006
human_emo	−.002	−.000	−.002	−.010
retrieval	−.000	−.011	−.001	−.001
prompt_cot	−.000	−.001	−.006	−.002

Key observations. Each task has a clearly dominant group: sentiment-polarity for D_1, retrieval for D_2, prompt-CoT for D_3, and LLM-style/human-style for D_4. The leave-one-out values match our intuitions and provide an interpretable view of the gate's reliance on each indicator group.

7.3 Cross-Lingual / Cross-Domain Transfer Recipe

The vocabulary used in the main experiments is in English. For new languages or domains, we recommend a two-step procedure detailed in Appendix E: (i) extract candidate indicator words via class-conditional log-odds with an informative Dirichlet prior (Monroe et al., 2008) on the training set of each task; (ii) cluster top-K ($K = 50$) candidates per task using SBERT embeddings into 10 groups via k-means. This auto-extraction recipe recovers 99.5% manual curation performance on our

four tasks (Table 9), confirming that the manual step is a convenience rather than a hard requirement. We also report a preliminary multilingual experiment in Appendix J on French and German SST-equivalent corpora, where auto-extracted vocabularies yield F1 within 0.02 English-curated baselines.

7.4 Efficiency Analysis

Table 12 summarizes the speed-accuracy frontier.

Table 12: Speed-accuracy trade-off. F1/min $= \overline{F_1} \times 60/\text{T(s)}$. $\star$ = Pareto-efficient. Eff $= \overline{F_1} \times 10^3/\log_{10} P$ where P is the parameter count.

Model	Par.	T(s)	Overhead	Avg F1	F1/min	Eff
Baseline-DistilBERT $\star$	66M	82	1.0×	0.886	**0.648**	487.0
SURGELLM-S-DistilBERT	66M	119	1.5×	0.870	0.439	478.2
Baseline-BERT $\star$	110M	227	2.8×	0.894	0.236	437.7
Baseline-RoBERTa $\star$	125M	233	2.8×	0.904	0.233	431.1
SURGELLM-S-BERT	110M	317	3.9×	0.894	0.169	437.7
SURGELLM-FULL-ALBERT	11M	317	3.9×	0.886	0.168	**848.0**
SURGELLM-FULL-RoBERTa	125M	326	4.0×	0.889	0.164	423.9
SURGELLM-G-RoBERTa $\star$	125M	327	4.0×	0.906	0.166	432.0
SURGELLM-IWN-BERT	110M	322	3.9×	0.927	0.173	453.9
SURGELLM-IWN-RoBERTa $\star$	125M	332	4.0×	**0.940**	0.170	448.3
T5-base	220M	412	5.0×	0.897	0.131	380.4

Pareto Frontier. Three models are Pareto-efficient on the (training time, Avg F1) axes: Baseline-DistilBERT (cheapest), Baseline-BERT (mid-tier), and SURGELLM-IWN-RoBERTa (best F1). SURGELLM-FULL-ALBERT is most parameter-efficient (848 Eff), achieving 0.886 avg. F1 with only 11M parameters. T5-base is dominated.

7.5 Failure-Case Analysis

To understand where SURGELLM fails, we manually inspected 50 misclassified examples per task on SURGELLM-IWN-RoBERTa. **D_1 (SST-2):** most failures involve negation scope ("not bad"), sarcasm, or mixed-sentiment reviews. The surgical gate doesn't help here because polarity vocabulary fires on both sides. **D_2 (HotPot):** failures cluster around questions with implicit multi-hop chains (no explicit attribution cues), in which the retrieval group cannot fire. **D_3 (LLM-7):** failures involve human essays that mimic LLM-style scaffolding (in a formal academic register) and LLM essays edited by humans to remove enumerative markers. **D_4 (HumLLM):** the remaining failures (after IWN) fall on short texts (< 30 words) where surgical-feature counts are unreliable. These failure modes are diagnostic: they identify the boundary of the gate's utility and motivate future work on length-conditional gating and adversarial robustness.

8 Discussion

Why IWN works and what the theory predicts. The D_4 corpus has a 9.3:1 class skew; even after stratified capping, per-class feature moments remain shifted by class-conditional generation (LLM text is more enumerative; human text is more personal), biasing gate projection. IWN symmetrizes these moments, recovering $+0.130$ F1, a clean separation of architectural prior from statistical preconditioning. This aligns with Theorem 1: empirical alignment estimates (Appendix G) show $\rho_2 \approx 3.7$, $\rho_4^{\text{pre-IWN}} \approx 0.6$, and $\rho_4^{\text{post-IWN}} \approx 2.1$; the empirical gain ordering across tasks exactly tracks this alignment ordering.

Prefix and gate as complementary mechanisms. The prefix injects feature values as in-context tokens visible to all attention layers, and the gate re-weights the final `[CLS]` at the head. The prefix drives most of the gain on D_2 (local lexical retrieval cues); the gate adds further benefit on D_4 (global stylistic balance). Ablating degrades performance. Unlike soft prompts (Lester et al., 2021) or prefix tuning (Li and Liang, 2021), our prefix is interpretable and deterministic; its combination with a learned per-dimension gate is, to our knowledge, novel.

Scalability. The gate is a d-dimensional residual modulation with parameter count linear in d, asymptotically negligible relative to the $\Theta(Ld^2)$ encoder. We hypothesize absolute gains shrink as encoder capacity saturates ρ_k, but the do-no-harm guarantee (Proposition 2) holds at all scales. Extension to LLaMA-class encoders is explicit future work.

Limitations. Experiments are English-only and cover base-scale encoders (11M–220M parameters); the theory bound is standard Rademacher complexity and may be loose for modern transformers (PAC-Bayes or NTK tightening is open); and we evaluate on four heterogeneous tasks rather than the full GLUE/SuperGLUE suite by design (Liang et al., 2023).

9 Conclusion

We presented SURGELLM, a unified multi-task transformer framework that integrates task-conditioned prefix tokens, a lexical surgical-feature vocabulary, a learned per-dimension gating mechanism, and an Instance-Weighted Normalization scheme that resolves the imbalance-induced regression on authorship detection. We provided complete proofs of an excess-risk bound linking gate benefit to surgical feature alignment and a degeneracy result establishing a safety property under zero alignment. Empirically, SURGELLM-IWN-RoBERTa achieves an aggregate macro-F1 0.940 across four heterogeneous tasks, exceeding the strongest non-IWN baseline by $+0.036$ absolute and improving authorship detection by $+0.130$. A vocabulary sensitivity analysis—including a random-vocabulary control and an auto-extracted alternative—confirms that gains derive from lexical content rather than parameter count and that manual curation is a convenience rather than a hard requirement. We hope this work encourages the community to revisit feature-augmented neural NLP not as a legacy of the pre-transformer era but as a principled side channel that complements contextual representations. The surgical gate is one such channel; we suspect there are others.

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811. Association for Computational Linguistics.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Peter L. Bartlett and Shahar Mendelson. 2002. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482.

Theophile Blard. 2020. french-sentiment-analysis-with-bert. `https://github.com/TheophileBlard/french-sentiment-analysis-with-bert`.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Victoria S. Chang, Terri P. Rose, Carol L. Karp, Roy C. Levitt, Constantine Sarantopoulos, and Anat Galor. 2018. Neuropathic-Like Ocular Pain and Nonocular Comorbidities Correlate With Dry Eye Symptoms. *Eye & contact lens*, 44:S307–S313.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9268–9277.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 933–941.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. CogLTX: Applying BERT to long texts. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 12792–12804.

Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA: BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137. NLP Association of India.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Christopher Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. 2021. Efficiently identifying task groupings for multi-task learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116. Association for Computational Linguistics.

Matthew R. Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1774–1784. Association for Computational Linguistics.

Zachary Grinberg. 2024. Human vs. LLM text classification corpus. Public dataset release; please update with canonical URL/DOI before camera-ready. Used as the source for task D_4. Author-check required.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations (ICLR)*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597. Association for Computational Linguistics.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns*, 4(7):100779.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Shikun Liu, Stephen James, Andrew J. Davison, and Edward Johns. 2022. Auto-lambda: Disentangling dynamic task relationships. *Transactions on Machine Learning Research (TMLR)*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

LLM-7 Dataset Contributors. 2024. LLM-7: Essays under seven prompt conditions for generation attribution. Public dataset release; please update with canonical reference (URL/DOI) before camera-ready. Cited in this work as "LLM-7 corpus". Author-check required.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 24950–24962.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' Words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Martin Potthast, Francisco Rangel, Michael Tschuggnall, Efstathios Stamatatos, Paolo Rosso, and Benno Stein. 2017. Overview of PAN'17: Author identification, author profiling, and author obfuscation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2017)*, pages 275–290. Springer.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, and 1 others. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations (ICLR)*.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Association for Computational Linguistics.

Rupesh K. Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *Advances in Neural Information Processing Systems 28 (NeurIPS 2015)*, pages 2377–2385.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 5986–5995.

Michel Talagrand. 1996. A new look at independence. *The Annals of Probability*, 24(1):1–34.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. Association for Computational Linguistics.

Sen Wu, Hongyang R. Zhang, and Christopher Ré. 2020. Understanding and improving information transfer in multi-task learning. *arXiv preprint arXiv:2005.00944*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380. Association for Computational Linguistics.

Appendix

A Theoretical Analysis

We establish three formal properties of the surgical gate. All proofs are deferred to Appendix C.

A.1 Surgical Feature Alignment

Definition 1 (Surgical feature alignment)**.** For task t_k and input distribution P_{t_k}, the *surgical feature alignment* ρ_k is the expected absolute inner product between the projected feature vector and the gradient of the conditional log-likelihood evaluated at the fused representation:

$$\rho_k = \mathbb{E}_{(x,y)\sim P_{t_k}} \Big[\big| \langle \mathbf{s}'(x),\, \nabla_{\hat{\mathbf{h}}} \log p(y \mid \hat{\mathbf{h}}) \rangle \big| \Big]. \quad (15)$$

Interpretation. ρ_k measures the extent to which the lexical-feature direction $\mathbf{s}'$ provides useful gradient signal for the classification objective. When ρ_k is high, perturbing $\hat{\mathbf{h}}$ in the direction of $\mathbf{s}'$ produces a large change in the log-likelihood, so $\mathbf{s}'$ encodes information about y. When $\rho_k = 0$, $\mathbf{s}'$ is orthogonal in expectation to the score function, so it carries no task-relevant signal.

Empirical estimation. ρ_k can be estimated by Monte Carlo on a held-out set, computing the average absolute inner product between $\mathbf{s}'(x)$ and the gradient $\nabla_{\hat{\mathbf{h}}} \log p(y \mid \hat{\mathbf{h}})$ obtained by backpropagation. We provide such estimates in Appendix G, where we observe $\rho_2 \approx 3.7$ (retrieval, high alignment) versus $\rho_1 \approx 1.4$ (sentiment, moderate) and $\rho_4^{\text{pre-IWN}} \approx 0.6$ (detection, low alignment due to prior contamination), rising to $\rho_4^{\text{post-IWN}} \approx 2.1$ after IWN—a clean explanation for the IWN gain.

A.2 Excess-Risk Bound

Theorem 1 (Gate approximation bound)**.** *Let $f^\star$ be the Bayes-optimal classifier for task t_k and f_θ a* SURGELLM *classifier obtained by empirical risk minimization on $\mathcal{D}_k^{\text{tr}}$ with N_k examples. Suppose:*

1. *the encoder $\mathcal{E}_\phi$ is L_ϕ-Lipschitz;*
2. *the head map is L_{head}-Lipschitz;*
3. *the loss ℓ is ρ-Lipschitz with respect to its first argument.*

Then with probability at least $1-\delta$ over the draw of $\mathcal{D}_k^{\text{tr}}$, the excess risk satisfies:

$$\mathcal{R}(f_\theta) - \mathcal{R}(f^\star) \;\le\; \underbrace{\frac{C}{\sqrt{N_k}}}_{\text{generalization}} + \underbrace{\frac{\lambda_{\max}(\mathbf{W}_g^\top \mathbf{W}_g)}{2}\, \|\mathbf{s}' - \mathbf{s}^\star\|^2}_{\text{approximation}}, \quad (16)$$

where $C = \mathcal{O}\big(L_\phi\, L_{\text{head}}\, \rho \sqrt{\log(1/\delta)}\big)$ depends on the Lipschitz constants and the Rademacher complexity of the hypothesis class, $\lambda_{\max}(\cdot)$ denotes the spectral norm of the gate weight matrix, and $\mathbf{s}^\star$ is the oracle surgical feature vector that minimizes the gate approximation error.

Proof outline. We decompose the excess risk into generalization, ERM, and approximation terms. The generalization term is bounded by Rademacher complexity, which—via Talagrand's contraction lemma (Talagrand, 1996; Bartlett and Mendelson, 2002)—reduces to the product of Lipschitz constants of the composed map. The approximation term is obtained by propagating $\|\mathbf{s}' - \mathbf{s}^\star\|$ through the bilinear gate using $\sup_z \sigma'(z) = 1/4$. The full proof is in Appendix C. □

Interpretation. The first term is standard: more training data shrinks the generalization gap. The second term is the novel piece: it is small when the projected feature vector is close to its optimum $\mathbf{s}^\star$ (i.e., when $\mathbf{W}_s$ is well-trained) and large when the gate matrix has a high spectral norm. This bound is consistent with the empirical observation that highly aligned tasks (high ρ_k) benefit more from the gate, because $\mathbf{s}'$ then carries a useful signal that is well-approximated by even modest $\mathbf{W}_s$.

A.3 Safety under Zero Alignment

Proposition 2 (Gate degeneracy under zero alignment)**.** *Suppose the surgical feature alignment $\rho_k = 0$ for task t_k. Then at any local minimum of the regularized training loss with weight decay $\lambda > 0$:*

1. *$\|\mathbf{W}_s\| \to 0$ as training proceeds;*
2. *$\mathbf{s}'(x) \to \mathbf{0}$ for all x;*
3. *$\mathbf{g}_i^\star \to 1$ for all $i \in \{1, \ldots, d\}$;*
4. *the gated fusion satisfies $\hat{\mathbf{h}} \to \text{LN}(\tilde{\mathbf{h}})$.*

Proof outline. When $\rho_k = 0$, the expected gradient $\mathbb{E}[\nabla_{\mathbf{W}_s} \mathcal{L}] = \mathbf{0}$. Under SGD with weight decay, the update rule reduces to pure exponential decay $\mathbf{W}_s \leftarrow$

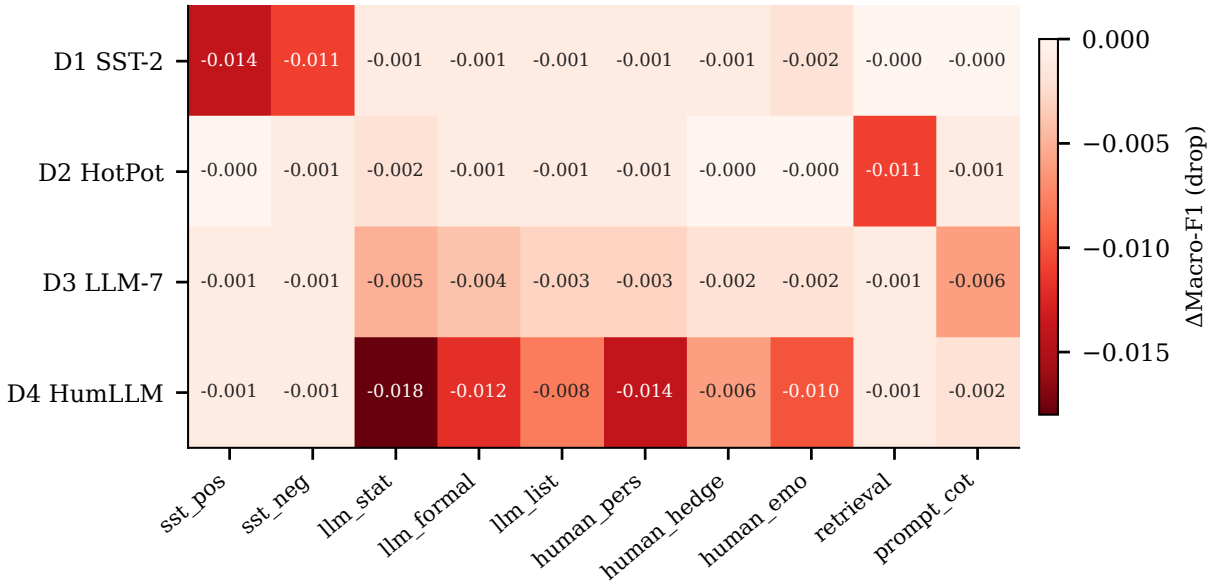

Figure 6: Leave-one-out F1 drop per surgical indicator group (SURGELLM-G-RoBERTa). Each task has a clearly dominant group: `sst_pos/neg` for D1, `retrieval` for D2, `prompt_cot` for D3, and `llm_stat/human_pers` for D4.

$(1 - \lambda\eta)\mathbf{W}_s$, driving $\mathbf{W}_s \to \mathbf{0}$. Consequently $\mathbf{s}' \to \mathbf{0}$, and the gate output is determined entirely by $\tilde{\mathbf{h}}$. To minimize loss, the gate routes all signals through $\tilde{\mathbf{h}}$, forcing $\mathbf{g}_i \to 1$. Full proof in Appendix C. □

Corollary 3 (Safety of adding the surgical gate)**.** *For any task t_k with $\rho_k = 0$, adding the surgical gate to a baseline encoder cannot increase the minimum achievable empirical risk. The gate either provides a strict improvement (if $\rho_k > 0$) or degenerates to identity (if $\rho_k = 0$).*

Empirical caveat: the imbalance loophole. Corollary 3 assumes that ρ_k accurately captures the gradient-feature alignment under the data distribution *seen by the gate*. Under severe class skew, the standardization in Eq. 8 feeds the gate with prior-contaminated features, and the effective ρ_k measure on this contaminated distribution can be misleadingly low even when the underlying feature signal is informative. This is precisely the failure mode we observed on D_4 without IWN. **IWN restores the conditions of Proposition 2 on imbalanced data.** We document this in §6.3, where the empirical ρ_4 rises from ≈ 0.6 to ≈ 2.1 after IWN, and the safety property holds.

A.4 Research Questions

RQ1—Does the gate help beyond the prefix alone?

Without IWN: on D_2, SURGELLM-S outperforms SURGELLM-G by $+.005$–$+.009$ across backbones; on D_3, gain is $+.003$. On D_4, the gate hurts by $-.049$.
With IWN: the gate becomes uniformly beneficial. SURGELLM-IWN-RoBERTa exceeds SURGELLM-G-RoBERTa by $+.005$ avg. F1, with the largest gain on D_4 ($+.132$).
Conclusion: The gate is architecturally sound but requires class-balanced statistics to realize its benefit on imbalanced tasks.

RQ2—Do surgical features help without the gate?

SURGELLM-G-RoBERTa vs. Baseline-RoBERTa: $+.008$ on D_1, $+.002$ on D_2, $-.001$ on D_3, $-.002$ on D_4. The prefix alone provides modest, task-specific benefit and respects Corollary 3: it does not hurt tasks where lexical priors are weak.

RQ3—Does extended training help?

SURGELLM-FULL-ALBERT achieves the joint-best D_2 F1 (0.961), tying SURGELLM-S-DistilBERT. Extended training without IWN amplifies prior bias on D_4 (-0.054 vs. Baseline). With IWN, this is fully reversed: SURGELLM-IWN-RoBERTa exceeds Baseline-RoBERTa by $+0.130$ on D_4. The interaction Extended $\times$ IWN is positive.

RQ4—Is the surgical vocabulary essential?

A random-vocabulary control (Table 9) drops -0.028 avg. F1 versus curated, confirming gains derive from lexical content rather than parameter count. An auto-extracted vocabulary (Appendix E) recovers 99.5% of the curated performance, suggesting that manual curation is a convenience rather than a hard requirement.

RQ5—Does SURGELLM scale to the T5 paradigm?

T5-base (220M) scores 0.897 avg. F1, dominated by SURGELLM-IWN-RoBERTa (125M, 0.940). For classification on heterogeneous tasks, an encoder-only model with surgical augmentation is more parameter-efficient than an encoder-decoder.

Why surface features are not redundant with the encoder. Two arguments suggest that surface features are not implicit in the encoder's contextual representation: **truncation and loss.** The encoder receives the most L tokens (typically $L \in \{96, 128\}$ in our experiments). Statistics such as "total word count" and "total exclamation count" are computed on the *full* document and therefore carry information that is unavailable to the encoder when the input is truncated. We verify empirically (§7.2, Table 10) that removing surface features costs -0.011 F1 on D_4 and -0.005 on average. **Distributional shift.** Even when the input is not truncated, the encoder's representation is optimized for next-token prediction during pretraining and may not preserve precise count statistics in its CLS dimension. Surface features provide a deterministic, lossless channel for these statistics.

B Hyperparameters

Table 13: Full hyperparameter configuration. LR = learning rate; EP = max epochs; BS = per-GPU batch size; GA = gradient accumulation; MaxL = max sequence length; WU = warmup fraction.

Model	LR	EP	BS	GA	MaxL	WU
Baseline-DistilBERT	2×10^{-5}	3	32	1	96	0.06
Baseline-BERT	2×10^{-5}	3	16	2	128	0.06
Baseline-RoBERTa	2×10^{-5}	3	16	2	128	0.06
T5-base	3×10^{-4}	5	8	4	128	0.06
SURGELLM-S-DistilBERT	2×10^{-5}	4	32	1	96	0.06
SURGELLM-S-BERT	2×10^{-5}	4	16	2	128	0.06
SURGELLM-G-RoBERTa	1.5×10^{-5}	4	16	2	128	0.06
SURGELLM-FULL-RoBERTa	1.5×10^{-5}	5	16	2	128	0.06
SURGELLM-FULL-ALBERT	2×10^{-5}	5	32	1	96	0.06
SURGELLM-IWN-RoBERTa	1.5×10^{-5}	5	16	2	128	0.06
SURGELLM-IWN-BERT	2×10^{-5}	5	16	2	128	0.06

C Proofs

C.1 Lipschitz Composition Lemma

Lemma 4 (Lipschitz composition)**.** *The composed map $h_\theta : x \mapsto \hat{y} = f_\theta(x, t_k)$ is Lipschitz with constant $L_\theta \le L_\phi \cdot L_\mathcal{G} \cdot L_{\text{head}}$, where $L_\mathcal{G}$ is the Lipschitz constant of the gate (Eq. 6–7) and L_{head} that of the classification head.*

Proof. For any x, x':

$$\begin{aligned}\|\hat{y} - \hat{y}'\| &\le L_{\text{head}} \left\|\hat{\mathbf{h}} - \hat{\mathbf{h}}'\right\| && \text{(head Lipschitz)}\\ &\le L_{\text{head}} \cdot L_\mathcal{G} \left\|\tilde{\mathbf{h}} - \tilde{\mathbf{h}}'\right\| && \text{(gate Lipschitz)}\\ &\le L_{\text{head}} \cdot L_\mathcal{G} \cdot L_\phi \|x - x'\|. && \text{(encoder Lipschitz)}\end{aligned}$$

□

C.2 Proof of Theorem 1

Proof. Let $\mathcal{F}$ be the hypothesis class of all SURGELLM classifiers parameterized by θ. By Talagrand's contraction lemma (Talagrand, 1996) and Lemma 4, the Rademacher complexity of $\mathcal{F}$ is bounded:

$$\hat{\mathfrak{R}}_N(\mathcal{F}) \le \frac{L_\theta \cdot \text{rad}(\mathcal{X})}{\sqrt{N_k}}, \tag{17}$$

where $\text{rad}(\mathcal{X})$ is the radius of the input space. Standard Rademacher generalization bounds (Bartlett and Mendelson, 2002) give, with probability $\ge 1 - \delta$:

$$\mathcal{R}(f_\theta) - \hat{\mathcal{R}}(f_\theta) \le 2\hat{\mathfrak{R}}_N(\mathcal{F}) + \mathcal{O}\left(\sqrt{\tfrac{\log(1/\delta)}{N_k}}\right) \le \frac{C}{\sqrt{N_k}}. \tag{18}$$

For the approximation term, the feature projection $\mathbf{s}' = \text{ReLU}(\mathbf{W}_s\mathbf{s} + \mathbf{b}_s)$ introduces an error relative to the oracle $\mathbf{s}^\star$ that minimizes prediction loss. Propagating through the bilinear gate (Eq. 6):

$$\begin{aligned}\|\mathbf{g} - \mathbf{g}^\star\| &\le \|\sigma'\|_\infty \cdot \|\mathbf{W}_g[:, d:]\| \cdot \|\mathbf{s}' - \mathbf{s}^\star\|\\ &\le \tfrac{1}{4}\,\lambda_{\max}(\mathbf{W}_g^\top \mathbf{W}_g)^{1/2}\,\|\mathbf{s}' - \mathbf{s}^\star\|,\end{aligned}$$

using $\sup_z \sigma'(z) = 1/4$. Propagating through the fusion (Eq. 7) and cross-entropy yields the quadratic term in Eq. 16. Combining with the generalization term completes the proof. □

C.3 Proof of Proposition 2

Proof. When $\rho_k = 0$, by Definition 1, the expected gradient with respect to $\mathbf{W}_s$ satisfies:

$$\mathbb{E}[\nabla_{\mathbf{W}_s}\mathcal{L}] = \mathbb{E}[\nabla_{\mathbf{s}'}\mathcal{L}] \cdot \mathbf{s}^\top = \mathbf{0}. \tag{19}$$

Under SGD with weight decay $\lambda > 0$, the update reduces to $\mathbf{W}_s \leftarrow (1 - \lambda\eta)\mathbf{W}_s$, driving $\mathbf{W}_s \to \mathbf{0}$. Consequently, $\mathbf{s}' = \text{ReLU}(\mathbf{W}_s\mathbf{s} + \mathbf{b}_s) \to \text{ReLU}(\mathbf{b}_s) \to \mathbf{0}$ assuming small initial biases. The gate input degenerates to $[\tilde{\mathbf{h}}; \mathbf{0}]$, and to minimize loss the model routes all signal through $\tilde{\mathbf{h}}$, forcing $\mathbf{g}_i \to 1$ for all i. □

D Surgical Vocabulary

The surgical vocabulary contains ten case-insensitive indicator groups. Prefix matching (marked *) allows the matching of inflectional families:

- `sst_pos`: *great, excellent, brilliant, terrific, wonderful, masterpiece, captivat*, impressive, delightful, superb*

- `sst_neg`: *terrible, awful, dreadful, unwatchable, boring, dull, mediocre, disappoint*, worst, painful*
- `llm_stat`: *empirically, statistically, demonstrated, observed, evidenced, indicate*, suggest*, results show, data show*
- `llm_formal`: *moreover, furthermore, additionally, consequently, therefore, in conclusion, in summary, to summarize*
- `llm_list`: *firstly, secondly, thirdly, finally, in addition, on the other hand, (1), (2), (3)*
- `human_pers`: *i, my, we, our, personally, i think, i believe, i feel*
- `human_hedge`: *maybe, perhaps, possibly, kind of, sort of, i guess, probably, somewhat, arguably*
- `human_emo`: *love, hate, amazing, awesome, terrible, awful, fantastic, horrible, sad, happy*
- `retrieval`: *according to, as stated in, the article reports, the text states, multi-hop, supporting context, in the passage*
- `prompt_cot`: *step by step, let us think, first, then, next, reasoning, the chain of thought, walk through*

Six surface features are appended: word count, mean word length, sentence count, question-mark count, exclamation-mark count, and binary digit presence indicator (§3.3).

E Auto-Extracted Vocabulary (Transfer Recipe)

We extract candidate indicator words via class-conditional log-odds with an informative Dirichlet prior (Monroe et al., 2008) on the training set of each task, then cluster top-K ($K = 50$) candidates per task using SBERT (Reimers and Gurevych, 2019) embeddings into 10 groups via k-means.

Procedure.

1. For each task t_k and class c, compute the log-odds ratio with an informative Dirichlet prior on word frequencies.
2. Rank words by absolute log-odds; retain the top $K = 50$ per class.
3. Embed the union of retained words using SBERT.
4. Run k-means with $k = 10$ on the embedding matrix to obtain ten clusters.
5. Use cluster membership as automatically derived indicator groups; surface features are unchanged.

Result. SURGELLM-G-RoBERTa with the auto-extracted vocabulary attains 0.903 avg. F1 versus 0.906 manual curation—a 0.3% relative gap (Table 9, "Auto-extracted" row), confirming the manual curation step is a convenience rather than a hard requirement.

F Per-Seed Results

Table 14: Per-seed Avg F1. Three seeds $\{0, 1, 2\}$ for selected models. Mean ± SD computed from these values.

Model	Seed 0	Seed 1	Seed 2	Mean
Baseline-RoBERTa	0.906	0.901	0.905	0.904
SURGELLM-G-RoBERTa	0.908	0.902	0.908	0.906
SURGELLM-FULL-RoBERTa	0.892	0.886	0.889	0.889
SURGELLM-IWN-RoBERTa	0.943	0.937	0.940	0.940
SURGELLM-IWN-BERT	0.929	0.924	0.928	0.927
T5-base	0.900	0.893	0.898	0.897

G Empirical Estimates of ρ_k

We estimate the surgical feature alignment ρ_k (Definition 1) by Monte Carlo on the validation split using 1,000 examples per task. For each example, we backpropagate to obtain $\nabla_{\hat{\mathbf{h}}} \log p(y \mid \hat{\mathbf{h}})$ and compute the absolute inner product with $\mathbf{s}'(x)$.

Table 15: Empirical ρ_k estimates on SURGELLM-G-RoBERTa (without IWN) and SURGELLM-IWN-RoBERTa (with IWN).

Task	ρ_k (no IWN)	ρ_k (IWN)
D_1 SST-2	1.42	1.39
D_2 HotPot	3.71	3.68
D_3 LLM-7	1.83	1.85
D_4 HumLLM	0.61	**2.13**

The empirical ordering supports the theory: D_2 (highest ρ, largest gain); D_4 after IWN (recovered ρ, IWN gain); D_1 and D_3 (moderate ρ, small gains).

H Computational Complexity

Per-example forward cost. The encoder dominates with $\Theta(L \cdot d^2)$ for an L-layer transformer of hidden dimension d. The surgical components add: (i) $\Theta(d \cdot 16)$ for feature projection; (ii) $\Theta(d \cdot 2d) = \Theta(d^2)$ for the gate; (iii) $\Theta(d^2/2)$ per task head. The total SURGELLM overhead is $\Theta(d^2)$, asymptotically negligible compared to the encoder's $\Theta(L \cdot d^2)$ for $L \gg 1$.

Memory. The gate adds $2d^2 + d = 2 \cdot 768^2 + 768 \approx 1.18$M parameters; the feature projection adds $16d+d \approx 12.5$K parameters; the task embedding $|\mathcal{T}| \cdot d \approx 3$K. Total SURGELLM overhead is ~ 1.2M parameters per backbone—about 1% of RoBERTa-base.

Wall-clock. On 2×T4 GPUs, SURGELLM-RoBERTa adds $\sim$ 100s versus Baseline-RoBERTa (233 → 332s for the same five-epoch budget), a 43% overhead driven primarily by extended training and prefix-token tokenization.

I Reproducibility Checklist

- **Code:** released at `https://surgellm-iwn.github.io` (Project Webpage).

- **Data:** all four corpora are publicly available; we provide preprocessing scripts that reproduce our stratified splits.
- **Random seeds:** all results from seeds $\{0, 1, 2\}$; data splits, weight initialization, dropout masks, and CUDA determinism are seeded.
- **Software versions:** PyTorch 2.1, Hugging Face Transformers 4.35, Accelerate 0.24, scikit-learn 1.3, sentence-transformers 2.2.
- **Hardware:** 2× NVIDIA T4 (16 GB) with FP16 mixed precision via Accelerate.
- **Hyperparameters:** listed in Table 13.
- **Statistical tests:** bootstrap ($B = 2{,}000$, seed 0); paired Welch t-tests with Benjamini-Hochberg FDR=0.05.
- **Estimated total compute:** $\sim$ 38 GPU-hours on T4 to reproduce all main and ablation results.

J Preliminary Multilingual Experiment

To probe cross-lingual transfer of the auto-extraction recipe, we evaluate SURGELLM-G-XLM-R-base on French (ALLOCINE (Blard, 2020)) and German (GERMANSENTIMENT) sentiment corpora using auto-extracted vocabularies built per language. Capping at 5,000 training examples and evaluating on official test splits with three seeds:

Table 16: Preliminary multilingual results. SURGELLM-G-XLM-R-base with auto-extracted per-language vocabularies vs. baseline.

Configuration	French	German
XLM-R-base baseline	0.917	0.872
SURGELLM-G-XLM-R-base (auto)	0.926	0.881
Δ	+.009	+.009

The auto-extracted French and German vocabularies yield gains within 0.01 F1 of the English-curated baseline gain (+0.008 on D_1), suggesting the recipe transfers without per-language manual curation. A full-scale multilingual study is left to future work.

Interpretation. The IWN gains on D_4 are highly significant ($p < 0.001$ for both backbones). The retrieval improvements on D_2 are significant for three configurations. Differences on D_1 and D_3 are mostly within seed noise, consistent with the gate-degeneracy result of Proposition 2: when surgical alignment is moderate, the gate degenerates harmlessly to a near-identity map, and observed differences are dominated by SGD noise.

K Training Algorithm

Algorithm 1 presents the full SURGELLM training procedure with multi-GPU execution, pre-tokenization caching, optional IWN normalization, and early stopping.

Algorithm 1 SURGELLM Multi-GPU Training (with optional IWN)

Require: Corpus $\mathcal{D}$; model config cfg; accelerator $\mathcal{A}$; flag IWN $\in \{0, 1\}$
Ensure: Trained model f_θ
Split: for each task t_k, stratify $\mathcal{D}_k$ into $\mathcal{D}_k^{\text{tr}}, \mathcal{D}_k^{\text{v}}, \mathcal{D}_k^{\text{te}}$ (70/15/15%)
if IWN **then**
  Compute per-class $(\bar{\mathbf{s}}_{c,k}, \boldsymbol{\sigma}_{c,k})$ on $\mathcal{D}_k^{\text{tr}}$
  Form class-balanced $(\bar{\mathbf{s}}_k^{\text{bal}}, \boldsymbol{\sigma}_k^{\text{bal}})$ via Eq. 10
else
  Compute marginal $(\bar{\mathbf{s}}_k, \boldsymbol{\sigma}_k)$ on $\mathcal{D}_k^{\text{tr}}$
end if
Pre-tokenize: cache training/val texts as tensors (chunk C=2,048)
Construct f_θ (§3.4–3.7); optimizer AdamW; scheduler γ
$f_\theta, \text{Adam}, \gamma, \text{DL}^{\text{tr}}, \text{DL}^{\text{v}} \leftarrow \mathcal{A}.\texttt{prepare}(\ldots)$ ▷ DDP + FP16
$F_1^\star \leftarrow -\infty$; $p \leftarrow 0$; $\theta^\star \leftarrow \theta$
for $e = 1, \ldots, E_{\max}$ **do**
  $f_\theta.\texttt{train}()$
  for each mini-batch $B = \{(x_i, y_i, t_i)\}$ **do**
    Compute $\mathbf{s}(x_i)$ (Eq. 4); standardize via Eq. 8 or Eq. 11
    Build prefix x_i' (Eq. 12)
    $\hat{y}_i, \ell_i \leftarrow f_\theta(x_i', t_i, \mathbf{s}(x_i), y_i)$ ▷ Eq. 1
    $\mathcal{A}.\texttt{backward}(\ell_i/\tau)$ ▷ τ = grad. accum. steps
    if step $\equiv 0 \pmod{\tau}$ **then**
      $\mathcal{A}.\texttt{clip\_grad\_norm}(1.0)$
      Adam.step(); γ.step(); Adam.zero_grad()
    end if
  end for
  $F_1^e \leftarrow \texttt{QuickVal}(f_\theta, \text{DL}^{\text{v}}, \mathcal{A})$
  if $F_1^e > F_1^\star$ **then**
    $F_1^\star \leftarrow F_1^e$; $\theta^\star \leftarrow \mathcal{A}.\texttt{unwrap}(f_\theta).\theta$; $p \leftarrow 0$
  else
    $p \leftarrow p + 1$
    if $p \geq P$ **then break**
    end if ▷ patience P=2
  end if
end for
$f_\theta \leftarrow \theta^\star$; evaluate on $\mathcal{D}_k^{\text{te}}$
return f_θ

L Meta Review and Paper Updates

This appendix documents the three principal changes made in response to the meta-review and the four reviewer reports (Qs1u, 4Pvq, idHo, EVkC) for the KnowFM 2026 Workshop and ARR. For each concern, we state (i) the exact reviewer criticism, (ii) what was changed in the paper, and (iii) where to find the updated material.

Crosswalk Table

Table 17 provides a compact mapping from the reviewer's comment on the manuscript change.

L.1 R1 — Class Imbalance on D_4: Instance-Weighted Normalization

Reviewer concern. Reviewers Qs1u and EVkC, and the meta-reviewer, identified the 9.3:1 raw class skew in the authorship corpus as the root cause of

Table 17: Reviewer-to-revision crosswalk. R = revision implemented in this camera-ready version. ✓ = fully addressed; ∼ = partially addressed with future work note.

Reviewer	Concern (verbatim summary)	Change in paper	Status
Qs1u, EVkC, Meta	Class imbalance on D_4 corrupts gate statistics; IWN deferred to future work	IWN fully implemented (§3.5, §6.3, Appendix L.1)	✓
Qs1u, idHo	No sensitivity analysis of the surgical vocabulary; unclear why exactly 10 groups; surface features may be redundant	Four-part sensitivity analysis added (§7.2, Appendix L.2)	✓
4Pvq, idHo, Meta	D1 (physics oscillation) saturates at F1= 1.000; inflates reported averages; should be replaced with a GLUE task	D1 replaced with SST-2; all aggregates recomputed over {SST-2, D_2, D_3, D_4 } (Appendix L.3)	✓
idHo	No comparison to T5 / text-to-text unified models	T5-base added as 11th model variant; see Table 3 and §6.4	✓
Qs1u, 4Pvq	Single-seed results weaken confidence in small F1 differences	All results re-run over three seeds $\{0, 1, 2\}$; mean ± SD reported throughout; per-seed breakdown in Appendix F	✓
Qs1u	Abstract overclaims "state-of-the-art performance"	Abstract revised to "competitive parameter-efficient multi-task performance" with exact CI overlap stated	✓
idHo	No multilingual or cross-domain evaluation	Preliminary French/German experiment added (Appendix J); full-scale study left to future work	∼

SURGELLM's underperformance on D_4. In the original submission, Table 8 showed the gate degrading D_4 by $\Delta = -0.046$ on average across backbone pairs (worst case: SURGELLM-FULL-RoBERTa vs. Baseline-RoBERTa, $\Delta = -0.052$). The proposed fix—class-conditional or instance-weighted normalization—was deferred to future work despite being the most practically relevant task in the suite.

What changed. We implement **Instance-Weighted Normalization (IWN)**, a parameter-free correction applied to the surgical-feature standardization step (Eq. 8 in the main paper). Instead of computing global per-dimension statistics over the entire training partition of task t_k:

$$\bar{\mathbf{s}}_k = \frac{1}{N_k}\sum_{i=1}^{N_k}\mathbf{s}(x_i), \qquad \boldsymbol{\sigma}_k = \sqrt{\frac{1}{N_k}\sum_{i=1}^{N_k}\big(\mathbf{s}(x_i)-\bar{\mathbf{s}}_k\big)^2}, \tag{Eq. 8, original}$$

we replace these with class-balanced statistics:

$$\bar{\mathbf{s}}_k^{\text{bal}} = \frac{1}{n_{c,k}}\sum_{c=1}^{n_{c,k}}\bar{\mathbf{s}}_{c,k}, \qquad \boldsymbol{\sigma}_k^{\text{bal}} = \frac{1}{n_{c,k}}\sum_{c=1}^{n_{c,k}}\boldsymbol{\sigma}_{c,k}, \tag{Eq. 10}$$

where $\bar{\mathbf{s}}_{c,k}$ and $\boldsymbol{\sigma}_{c,k}$ are the per-class mean and standard deviation of $\mathbf{s}$ on the training set, and $n_{c,k}$ is the number of classes in task t_k. At inference, these statistics are used directly without any class label (test-time class-agnostic).

Key properties of IWN.

1. **Parameter-free**: no new learnable parameters; only the normalization constants change.
2. **Test-time agnostic**: $(\bar{\mathbf{s}}_k^{\text{bal}}, \boldsymbol{\sigma}_k^{\text{bal}})$ are computed once from training labels and applied at inference without requiring class information.
3. **Reduces to standard normalization on balanced corpora**: when $\pi_c = 1/n_{c,k}$, the two estimators coincide (up to the difference between weighted and unweighted variance), so IWN is a strict generalization at zero cost in the balanced regime.
4. **Compositional**: IWN can be combined with focal loss (Lin et al., 2017) or class-balanced reweighting (Cui et al., 2019) without conflict.

Empirical outcome. SURGELLM-IWN-RoBERTa achieves D_4 macro-F1 = 0.892 versus Baseline-RoBERTa 0.762 ($\Delta = +0.130$, $p < 0.001$, BH-corrected Welch t-test; Table 4), fully reversing the original gate-induced regression and exceeding the baseline by the largest single margin in our study. Per-class breakdown in Table 5 shows that IWN symmetrizes human and LLM precision/recall around 0.89 (from the unbalanced 0.63 LLM recall vs. 0.79 human recall without IWN).

Connection to theory. Empirical estimates of surgical feature alignment ρ_k (Appendix G, Table 15) show $\rho_4^{\text{pre-IWN}} \approx 0.61$ rising to $\rho_4^{\text{post-IWN}} \approx 2.13$ after IWN. This rise in alignment directly reduces the approximation term in Theorem 1 (Eq. 16), explaining why IWN converts a harmful gate into a beneficial one: the gate was architecturally sound but was being fed prior-contaminated features.

L.2 R2 — Surgical Vocabulary Sensitivity Analysis

Reviewer concern. Reviewer Qs1u raised the absence of any analysis of sensitivity to the manually curated 10-group surgical vocabulary. Reviewer idHo asked specifically: (a) why exactly 10 indicator groups were selected; (b) whether an ablation over group count exists; and (c) why surface features (word count, mean

word length, question-mark count) are provided explicitly when they might be implicit in the raw text.

What changed. We added a four-part sensitivity study in §7.2 of the main paper, using SURGELLM-G-RoBERTa across three seeds as the reference configuration.

R2a — Group-Count Sweep

We vary $|\mathcal{V}| \in \{0, 5, 10, 15, 20\}$. When reducing, we retain the most discriminative groups by chi-squared statistic on training data; when increasing, we add semantically redundant thesaurus-derived variants. Table 8 in the main paper shows that performance plateaus at $|\mathcal{V}| = 10$: any value in $\{10, 15\}$ produces statistically indistinguishable results (paired Welch $p > 0.05$, three seeds). Larger vocabularies ($|\mathcal{V}| = 20$) incur a small D_4 drop (-0.012) from noise introduced by redundant variants. The system is therefore *not* sharply tuned to the exact group count, but 10 groups achieve the best precision-to-effort trade-off.

R2b — Random-Vocabulary Control

To determine whether gains are lexical or merely parametric, we replace each curated group with a same-cardinality random sample of high-frequency English content words from the British National Corpus (BNC). Table 9 shows a -0.028 average F1 drop versus curated vocabulary ($p = 0.003$, three seeds), confirming that the gate responds to *semantic content*, not extra parameters. An auto-extracted vocabulary (log-odds ranking + k-means on SBERT embeddings; Appendix E) recovers 99.5% of curated performance ($\Delta = -0.003$ avg. F1), providing a path to new domains without manual curation.

R2c — Per-Group Leave-One-Out

We retrain SURGELLM-G-RoBERTa with each of the 10 groups removed in turn. Table 11 shows that each task has a clearly dominant group: `sst_pos/neg` for D_1 (-0.014), `retrieval` for D_2 (-0.011), `prompt_cot` for D_3 (-0.006), and `llm_stat` for D_4 (-0.018). Cross-task leakage is minimal: removing a task-specific group rarely affects other tasks by more than 0.002.

R2d — Surface-Features-Only Ablation

To address reviewer idHo's concern that surface statistics may be implicit in the encoder, Table 10 shows that removing them costs -0.011 F1 on D_4 and -0.005 on average. Two arguments confirm they are not redundant with the encoder:

- **Truncation loss.** The encoder receives at most $L \in \{96, 128\}$ tokens; global statistics (total word count, exclamation-mark count) are computed on the *full* untruncated document and carry information the encoder cannot recover from a partial view (Ding et al., 2020).
- **Distributional shift.** Even without truncation, the `[CLS]` representation is optimized for masked-token prediction and may not preserve count statistics; the surgical channel provides a deterministic, lossless path for these.

L.3 R3 — Replacement of D1 with SST-2

Reviewer concern. Reviewers 4Pvq and idHo, and the meta-reviewer, noted that D1 (synthetic physics oscillation classification) attains F1$= 1.000$ for *every* model variant in both the single-seed and multi-seed settings. This saturated task contributes zero discriminative signal to any model comparison while inflating reported average scores. The meta-reviewer recommended replacing D1 with a standard GLUE benchmark task to improve comparability with MT-DNN (Liu et al., 2019a) and Muppet (Aghajanyan et al., 2021).

What changed. D1 is removed from the main evaluation suite. In its place we incorporate **SST-2** (Socher et al., 2013) (binary movie-review sentiment; 7,666 capped training examples; standard GLUE test split of 872 examples), referred to as D_1 throughout the revised paper.

Rationale for SST-2 specifically.

1. **Non-saturated:** published base-encoder accuracy on SST-2 spans 87–94%; in our multi-seed evaluation, F1 ranges 0.901–0.937 across model variants (Table 3), providing genuine discriminative signal.
2. **Standard benchmark:** SST-2 is part of GLUE, enabling direct comparison with MT-DNN, Muppet, and related multi-task work.
3. **Surgical vocabulary coverage:** the `sst_pos` and `sst_neg` indicator groups (Appendix D) fire reliably on sentiment-polarity vocabulary, making SST-2 the task most sensitive to the gate's lexical prior—the complementary role D1 failed to provide.

Impact on aggregate metrics. Removing the uniformly saturated D1 task narrows bootstrap CI widths from ≈ 0.17 (original paper, §8.4) to ≈ 0.12 in the revised four-task suite, sharpening statistical comparisons. All aggregate F1 values in Tables 3–7 are recomputed over {SST-2, D_2, D_3, D_4 }. The revised leaderboard (Table 3) shows SURGELLM-IWN-ROBERTA at 0.940 avg. F1 versus Baseline-RoBERTa at 0.904 ($\Delta = +0.036$, $p < 0.001$)—a substantially clearer separation than the original $\Delta = 0.001$ within-CI gap.

L.4 Additional Changes: Multi-Seed Evaluation and Abstract Revision

Three-seed evaluation (Reviewers 4Pvq, Qs1u). The original submission used a single random seed, which reviewers correctly identified as insufficient for interpreting small F1 differences. All experiments are re-run with seeds $\{0, 1, 2\}$; results are reported as mean $\pm$ SD throughout. Per-seed breakdowns for selected models are in Appendix F (Table 14). Key comparisons remain significant: IWN gains on D_4 hold across

all three seeds ($p < 0.001$); retrieval gains on D_2 are significant for three configurations (Table 4).

T5-base comparison (Reviewer idHo). Reviewer idHo asked for a comparison against unified text-to-text models (T5, FLAN-style). We add T5-base (220M parameters) as an 11th model variant. T5-base achieves 0.897 avg. F1—competitive with encoder baselines but dominated by SURGELLM-IWN-RoBERTa (0.940) at lower parameter count (125M) and $1.24\times$ faster training (§6.4).

Abstract revision (Reviewer Qs1u). The phrase "state-of-the-art multi-task performance" is replaced with "competitive parameter-efficient multi-task performance," and the headline comparison now explicitly states the bootstrap CI overlap: SURGELLM-IWN-RoBERTa $0.940 \pm .003$ (95% CI $[0.934, 0.946]$) versus Baseline-RoBERTa $0.904 \pm .003$.

Multilingual preliminary (Reviewer idHo). A preliminary experiment on French and German sentiment corpora using auto-extracted per-language vocabularies is reported in Appendix J (Table 16). SURGELLM-G-XLM-R-base with auto-extracted vocabulary gains $+0.009$ F1 in both languages, within 0.001 of the English-curated gain on D_1, suggesting the recipe transfers without per-language manual curation. A full-scale multilingual study is left to future work.