ICML
International Conference
On Machine Learning

Token Cleaning: Fine-Grained Data Selection for LLM Supervised Fine-Tuning

Jinlong Pang, Na Di, Zhaowei Zhu, Jiaheng Wei,
Hao Cheng, Chen Qian, Yang Liu

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

Motivation

New Censensus on SFT: data quality matters far more than quantity.

- Superficial Alignment Hypothesis: *LIMA [NeurIPS '23]*
- Empirical Observations: *ALPAGASUS [ICLR '24] , LESS [ICML'24], DS2 [ICLR'25], etc.*

Limitation: Even in high-quality samples, patterns or phrases that are not task-related can be redundant, uninformative, or even harmful

Sample-level —> Fine-grained token-level

Token Scoring Mechanism

Main components:

- Influence-guided Scoring Function[1]

$$\text{Infl}\,(x_{i,j}|\boldsymbol{x}_{i,:j};\theta,\theta') := \ell(x_{i,j}|\boldsymbol{x}_{i,:j};\theta') - \ell(x_{i,j}|\boldsymbol{x}_{i,:j};\theta).$$

$$\text{Score}(x_{i,j}|\boldsymbol{x}_{i,:j};\theta,\theta') = -\text{Infl}\,(x_{i,j}|\boldsymbol{x}_{i,:j};\theta,\theta'),$$

θ (θ_0) : Base model

θ' (θ_t) : Better (Reference) model

$\boldsymbol{x}$: token vector (sample)

$\boldsymbol{x}_{i,:j}$: previous $j-1$ tokens

$x_{i,j}$: target j-th token

Note: A higher score indicates a higher token quality.

- Threshold (empirical value, 60%)

$$\hat{y}_{i,j} = \begin{cases} 1 & \text{if Score}(x_{i,j}|\boldsymbol{x}_{i,:j};\theta,\theta') \text{ ranks top } k\%, \forall i,j; \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

[1] P. W. Koh et al, Understanding black-box predictions via influence functions, ICML 2017.

Token-Cleaning Pipeline

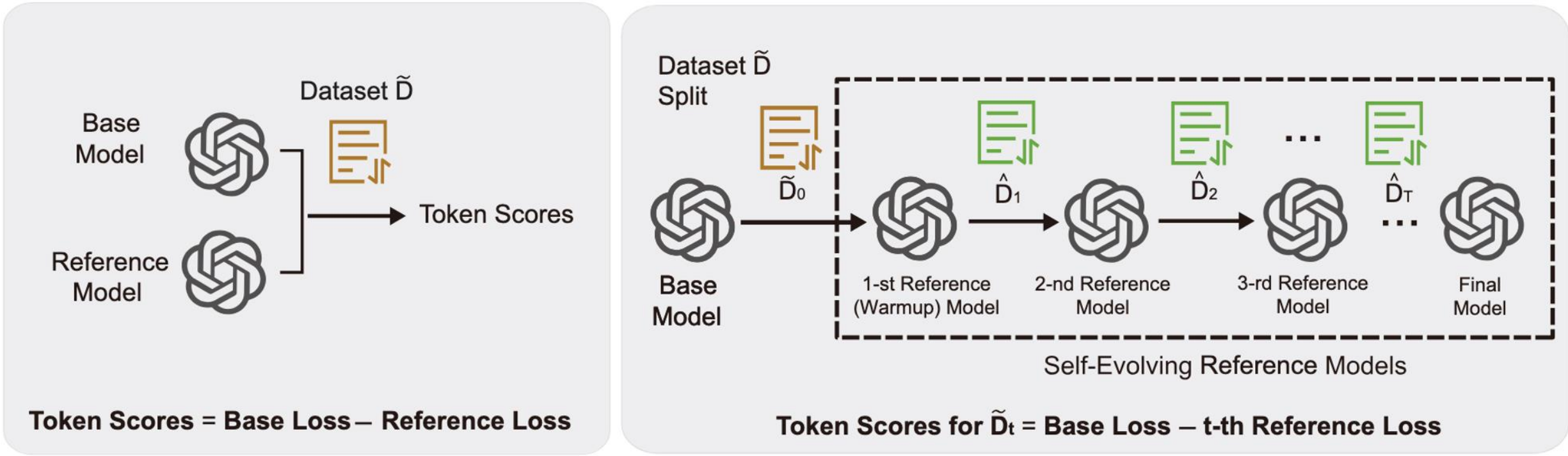

Figure: Overview of token cleaning pipeline

Theoretical Analyses: A Noisy Label Perspective

When and why SFT with cleaned tokens outperforms the full tokens?

Theorem 3.1 (Error of learning with full tokens) *With probability at least* $1-\delta$*, the generalization error of learning with full tokens is upper-bounded by*

$$\mathcal{L}_{\mathcal{D}}(\hat{\theta}_{\widetilde{D}}) \leq \underbrace{\eta(\widetilde{D})}_{\text{Data quality}} + \underbrace{\sqrt{\frac{2\log(4/\delta)}{M}}}_{\text{Data quantity}}, \tag{6}$$

where $M := \sum_{i=1}^{N} L_i$ *denotes the number of tokens.*

- Depends on two factors:
  - Data quality: $\eta(\widetilde{D}) := \mathbb{P}(\widetilde{Y} \neq Y)$ -- the noise rate of full tokens.
  - Data quantity: the number of token M

Theoretical Analyses: A Noisy Label Perspective

Which condition hold if token cleaning works?

Corollary 3.1.1 *With probability as least* $1-2\delta$*, token cleaning performs better than full-tokens in terms of the error upper bound when*

$$\eta(\widetilde{D}) - \hat{\eta} \geq \sqrt{2\log(4/\delta)} \cdot \sqrt{\frac{1}{M}} \cdot \left(\sqrt{\frac{1}{\hat{r}}} - 1\right), \tag{7}$$

where $\hat{\eta} := (\widehat{Y} \neq Y)$ *denotes the noise rates of cleaned labels and* $\hat{r} := (\widehat{Y} = 1)$ *denotes the ratio of positive tokens after token cleaning.*

Conclusion: token cleaning is preferred when the positive impact of reducing noise rate outweighs the negative impact of reducing the number of feasible tokens.

Main Experiments

- **Experimental Setup**
  - Base model: LLaMA-3.2-3B, LLaMA-3.1-8B, Mistral-7B-v0.3
  - Data Pool: DS^2 50k samples[1]
  - Selected token proportion: 60%

Model	TruthfulQA	TydiQA	LoqiQA	MMLU	HellaSwag	ARC-C	BoolQ	AVG
Base model: LLaMA-3.2-3B								
BASE	39.39	21.10	22.17	56.29	55.24	42.20	72.95	44.19
DS^2 (10K)	43.35	41.20	24.96	56.93	55.64	44.62	74.80	48.79
FULL TOKENS (50K)	43.32	49.60	24.34	56.87	55.57	44.44	74.98	49.87
UNIFORM RANDOM (50K×0.6)	43.79	47.00	23.41	56.96	55.37	44.44	75.05	49.43
RHO	45.57	<u>53.60</u>	<u>26.05</u>	**57.10**	55.16	<u>45.39</u>	<u>77.36</u>	51.46
FIXED-MODEL CLEANING	<u>48.96</u>	52.60	25.89	<u>57.09</u>	**56.43**	<u>45.39</u>	**77.52**	<u>51.98</u>
SELF-EVOLVING CLEANING	**51.07**	**56.38**	**28.22**	56.18	<u>55.81</u>	**45.99**	77.33	**53.00**

[1] J. Pang et al, Improving Data Efficiency via Curating LLM-Driven Rating Systems, ICLR 2025.

Summary

- We systematically analyze when and why SFT with the cleaned tokens outperforms the full tokens.
- Proposed token cleaning pipeline effectively remove uninformative tokens while preserving task-relevant information.
- Token cleaning pipeline further boost the performance of sample-level work.

https://github.com/UCSC-REAL/TokenCleaning

Code