

Peer Review

Review of: "SMOOTHIE: Label Free Language Model Routing"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper asks a practical question that most routing work sidesteps: can you pick, per input, which of several LLMs to trust when you have no labels at all? SMOOTHIE answers with a weak-supervision construction. Each LLM output is embedded together with its input (SentenceBERT all-mpnet-base-v2 by default), and the difference between each output embedding and an unobserved "true" embedding is modelled as a zero-mean Gaussian with diagonal covariance whose variance is inversely proportional to a quality score $\theta_i(x)$ (Eq. 1). Because the errors are assumed independent, the expected squared distance between two LLM outputs splits into the sum of their individual error terms (Proposition 1, Eq. 2), and a triplet of LLMs gives a closed-form estimate of each one (Eq. 3 and Eq. 4). SMOOTHIE-GLOBAL averages the pairwise distances over the whole test set; SMOOTHIE-LOCAL averages them only over the n_0 nearest neighbors of the test input, and the router then picks the argmax of θ .

Empirically, the authors report that the global weights correlate with true model quality (average Spearman 0.72 over the NLG tasks, 0.94 on MixInstruct, 0.46 on AlpacaEval, §5.1), that the top-weighted model is the true best on 9/14 task-ensemble pairs (abstract), that SMOOTHIE-LOCAL beats every baseline on both mixed-task distributions in Table 2 including the labeled PAIRRM and LABELED-KNN, and that the same machinery works for choosing among prompt templates for a fixed model (Table 3). What I like most is how little the method needs: no gradient steps, a few seconds per thousand samples (§5.4, "2.14 seconds per 1000 samples" for SMOOTHIE-LOCAL on the 7B ensemble), and a derivation short enough to check by hand. I did check it, and Appendix D is correct as far as it goes.

2. Whether the conclusion holds on the data

The central empirical claim, that label-free local routing beats both random selection and the best single model, is supported by Table 2. SMOOTHIE-LOCAL is the top row in all four columns: 58.7 vs 53.9 (PAIRRM) on 3B DISTR-ACC, 20.2 vs 19.0 (PAIRRM) on 3B DISTR-ROUGE2, 75.0 vs 73.2 (BEST-MODEL) on 7B DISTR-ACC, 26.9 vs 26.4 (BEST-MODEL) on 7B DISTR-ROUGE2. The 3B gains are large. That BEST-MODEL is itself an oracle choice made with labels makes the result more convincing, not less. The ensemble-sampling ablation (§5.4, "Different ensembles") is also a good design: 350 settings instead of two hand-picked pools.

Where I think the text runs ahead of the numbers is in several places where the prose and the tables disagree. I recomputed every comparison from the tables as rendered on Qeios.

Table 1 versus the §5.1 prose. The text says SMOOTHIE-GLOBAL beats RANDOM "on 6/7 tasks for the 3B ensemble." In Table 1, 3B block, SMOOTHIE-GLOBAL is higher on CNN (14.3 vs 12.9), Def. Ext. (61.5 vs 52.4), E2E (31.8 vs 27.3), SQuAD (60.7 vs 59.6), and WebNLG (30.7 vs 23.4), ties on XSum (4.5 vs 4.5), and is lower on TriviaQA (32.1 vs 32.7). That is 5 wins, 1 tie, 1 loss. The 7B claim (7/7) and the BEST-ON-VAL counts (4/7 and 5/7) do check out, as do "up to 7pts" (WebNLG 3B, $30.7 - 23.4 = 7.3$) and "up to 12pts" (Def. Ext. 7B, $70.9 - 58.5 = 12.4$).

Table 9 versus the "Choice of embeddings" paragraph. This one is inverted. The text says SentenceBERT beats BGE on 3B DISTR-ACC and 7B DISTR-ROUGE2, and loses on 3B DISTR-ROUGE2 and 7B DISTR-ACC. Table 9 shows the opposite in every cell: 3B DISTR-ACC, BGE 59.3 vs SBERT 58.7; 3B DISTR-ROUGE2, SBERT 20.2 vs BGE 19.7; 7B DISTR-ACC, SBERT 75.0 vs BGE 74.6; 7B DISTR-ROUGE2, BGE 27.1 vs SBERT 26.9. The substantive conclusion (both embeddings work, neither dominates) survives. The sentence does not. The same paragraph also opens with "affects SMOOTHIE-GLOBAL performance" while Table 9 reports only SMOOTHIE-LOCAL.

Headline numbers in §1 that I could not reconcile. §1 says SMOOTHIE-LOCAL improves on SMOOTHIE-GLOBAL "by up to 7 points accuracy"; Table 2, 7B DISTR-ACC gives $75.0 - 66.5 = 8.5$, so the bound is understated. §1 says SMOOTHIE-GLOBAL beats random selection "by up to 15 points win-rate on AlpacaEval", while §5.1 says it does so "by an average of 15pt win-rate, and up to 27pts." Those are different statements. §1 also says prompt selection gains are "up to 18 points"; the largest SMOOTHIE-GLOBAL minus RANDOM gap I find in Table 3 is Llama-2, TriviaQA, $68.7 - 45.9 = 22.8$, and none of the

gaps to BEST-ON-VAL are anywhere near 18. I may be missing the intended comparator, but the reader should not have to guess.

Small mismatches between text and appendix tables. §5.1 gives RANDOM on MixInstruct as a ChatGPT-rank of 5.95; Table 10, row RANDOM, gives 5.96. §5.1 gives RANDOM on GSM8K as 28.3%; Table 11, row RANDOM, gives 28.4. In §5.4, 211 of 350 is 60.29%, which rounds to 60.3, not the printed 60.2 (and both 60.2 and 83 lack a percent sign).

How strong is "substantially"? §5.2 says SMOOTHIE-LOCAL "substantially betters BEST-MODEL." On 3B that is fair (58.7 vs 52.3; 20.2 vs 18.1). On 7B the margins are 1.8 points (75.0 vs 73.2) and 0.5 points (26.9 vs 26.4), and the local-over-global gap on 7B DISTR-ROUGE2 is 0.8 (26.9 vs 26.1). No variance is reported for any SMOOTHIE row, and RANDOM (Table 2) and BEST-ON-VAL (Table 1) are reported as means over 10 draws without spread. With test sets of roughly a thousand items per task, a paired bootstrap over test items would tell the reader whether 0.5 ROUGE-2 points is signal. I would soften the 7B wording or add the intervals.

The missing baseline is the one the paper itself names. Appendix C, "Consistency-based selection", explains that Minimum Bayes Risk selection (pick the generation with the highest average similarity to the others) is close to SMOOTHIE, and that SMOOTHIE with $n_0 = 1$ reduces to it if the $-\delta_{jk}$ term in Eq. 3 is dropped and the average runs over more than the two terms δ_{ij} and δ_{ik} . That is exactly why an MBR or medoid baseline belongs in Tables 1 to 3. Without it, I cannot tell how much of the gain comes from the Gaussian model and the triplet correction, and how much from plain consensus. The observation in §5.4 that SMOOTHIE-LOCAL is best at $n_0 = 1$ and "steadily degrades as n_0 increases" makes the question sharper: at $n_0 = 1$ each $\hat{\delta}$ is a single squared distance on one neighbour, so the "expectation" in Eq. 3 is estimated from one draw. The same applies to GSM8K (Table 11), where majority vote over final answers is the natural label-free comparator.

Modelling assumptions. The diagonal covariance means independent errors across LLMs; the Limitations paragraph in §6 acknowledges this, which I appreciate. Models from shared pretraining data will make correlated errors, and a consensus-based estimator will then reward the majority rather than the correct minority. One experiment that would help: an ensemble in which two near-duplicate models (for instance Gemma-2B and Gemma-7B, both already listed in Table 6) are added, to see whether their θ is inflated.

Train-variant results worth a sentence. In Table 7, SMOOTHIE-GLOBAL-Train on 3B E2E drops to 24.7 from 31.8, below RANDOM (27.3). §5.4 counts this as one of the "2/14" worse pairs, which is honest, but a drop below random deserves an explanation. In Table 8, SMOOTHIE-GLOBAL-train is identical to SMOOTHIE-GLOBAL in all four cells (51.3, 18.1, 66.5, 26.1), even though Table 7 shows the two differ on E2E. It may well be correct; I would say why.

3. Reproducibility

This is better than most routing papers. Code is linked (Appendix A, github.com/HazyResearch/SMOOTHIE), datasets are listed with Hugging Face paths (Table 5), every open model except Falcon-1B (used in §5.3) is listed with a Hugging Face path (Table 6), the ten AlpacaEval systems are named (Appendix E.1), the embedding checkpoints are named (`all-mpnet-base-v2`, `bge-small-en-v1.5`), and the method is deterministic given embeddings. What I could not find in the Qeios text:

- Decoding parameters for any generation: temperature, top-p, maximum new tokens, stop sequences, and whether sampling or greedy decoding was used. For the GSM8K chain-of-thought runs, this matters a lot.
- Model revisions. Table 6 gives repository names without commit hashes, and several of those repositories have been updated since release.
- The prompt templates. §5.1 says a single template was written per task, and §5.3 says 3 to 5 per task were written; none are reproduced in the Qeios version, which makes Table 3 impossible to replicate from the paper alone.
- How the 1000-sample subsets (§5.1) and the 50-sample validation draws for BEST-ON-VAL and LABELED-KNN were drawn, and with which seeds.
- Which split each "test" set comes from, and for SQuAD, why the `hazyresearch/based-squad` variant was used instead of the original.
- How the MixInstruct ChatGPT-rank was computed (which ChatGPT snapshot, which prompt). A rank judged by an unversioned API model cannot be recomputed later.
- Compute and hardware for producing the generations. The routing cost is tiny; the generation cost is not.

The rendering of Algorithm 1 on Qeios shows only its title ("Algorithm 1. Estimate Scores"); the body is not legible in this version, nor are Figures 1 to 8, so I could not verify the Figure 2 correlations, the Figure

3 rank histograms, or the Figure 4 claim that a 410M model beats a 6.9B one on E2E.

4. What specifically to fix

1. §5.4, "Choice of embeddings". Swap the two clauses so they match Table 9 (SBERT wins on 3B DISTR-ROUGE2 and 7B DISTR-ACC; BGE wins on 3B DISTR-ACC and 7B DISTR-ROUGE2), and change "SMOOTHIE-GLOBAL" to "SMOOTHIE-LOCAL" in the first sentence.
2. §5.1: "6/7 tasks for the 3B ensemble" should read 5/7 with one tie (XSum) and one loss (TriviaQA), per Table 1.
3. §1, the three "up to" claims. Align them with the tables: local-over-global is up to 8.5 (Table 2, 7B DISTR-ACC), AlpacaEval is "average 15, up to 27" per §5.1, and the prompt-selection "18 points" needs its comparator named, because Table 3 gives 22.8 against RANDOM on Llama-2 TriviaQA.
4. Reconcile 5.95/5.96 (§5.1 vs Table 10) and 28.3/28.4 (§5.1 vs Table 11); print 60.3% for 211/350.
5. §5.1 cites "Figure 10" for the best-model counts. The paper has Figures 1 to 8; I think Figure 8 is meant.
6. Add an MBR/medoid baseline to Tables 1, 2, and 3, and majority vote over final answers to Table 11. Appendix C already derives the relationship to MBR; the experiment would show what the $-\delta_{jk}$ correction buys.
7. §3.2, after Eq. 1: Z is written as $1/Z$ in front of the exponential, so it is the partition function, not the "log partition function."
8. §3.2, covariance: $\Sigma_{jj} = 1/(2\theta_{|j|/m})(x)$. The stacked vector has m blocks of length d , so the block index of coordinate j is $\lfloor j/d \rfloor$. With $\lfloor j/m \rfloor$ the indexing is wrong whenever $d \neq m$.
9. §4.1, "Sample-conditional estimation": $\delta_{ij}(x)$ is written as $E[\|\lambda_1(x) - \lambda_j(x)\|]$ without the square, whereas Eq. 3 and the estimator use the squared norm. The same paragraph defines neighbours "in f_0 's embedding space", but the embedding model is g_0 everywhere else.
10. §4.1: say what happens when $\delta_{ij} + \delta_{ik} - \delta_{jk} \leq 0$ for some triplet. With noisy one-neighbour estimates ($n_0 = 1$) that denominator can be zero or negative, and θ^{jk}_i is then undefined or negative before averaging. How often did it happen?
11. §5.4, "Improving efficiency": " $n \times m$ generations for a D_{train} of size n " should say D_{test} . The text also refers to "Algorithm 4.1" in two places where Algorithm 1 is meant.
12. Report a paired bootstrap interval (or standard deviation over the 10 draws) for every row of Tables 1 to 3 and 7 to 9, and temper "substantially" in §5.2 for the 7B columns if the intervals overlap.

13. Explain the Table 7, 3B E2E drop to 24.7 for SMOOTHIE-GLOBAL-Train, and why SMOOTHIE-GLOBAL-train equals SMOOTHIE-GLOBAL to the decimal in all four cells of Table 8.
14. In Appendix E add decoding parameters, model commit hashes, the prompt templates used in §5.1 and §5.3, subsample seeds, the ChatGPT snapshot used for MixInstruct ranking, and generation compute.
15. A correlated-errors stress test (two sibling models from Table 6 in the same ensemble) would turn the first limitation in §6 from a caveat into a measured property.

5. Recommendation

Minor revisions. The method is simple, well-motivated, correctly derived, and the key result in Table 2 holds on the numbers shown: a label-free router that beats labelled baselines on the 3B ensemble by a clear margin and beats them by smaller margins on 7B. Nothing I found undermines that. What needs fixing is mostly bookkeeping: an inverted paragraph against Table 9, a miscount against Table 1, several headline bounds in §1 that do not match the tables, two notational errors in §3.2 and §4.1, and a reproducibility appendix without decoding settings or prompts. The one substantive addition I would ask for is the MBR baseline, because the authors themselves show it is the closest competitor and the reader deserves to see the gap.

Declarations

Potential competing interests: No potential competing interests to declare.