

463 A Limitations

464 In this work, several limitations exist that should be acknowledged for a balanced understanding of
465 the results and methodology. First, while the Hierarchical Demonstration Order Optimization (HIDO)
466 framework effectively reduces the search space for many-shot in-context learning (ICL), its reliance
467 on clustering introduces an additional layer of complexity that may not always generalize well to
468 all datasets or language models. The clustering process itself, especially with a limited number of
469 clusters, may not capture intricate interdependencies between demonstrations. Furthermore, although
470 the dynamic update mechanism improves the accuracy of the score function, it also increases the
471 overall computational cost, particularly when applied to very large datasets or when running a high
472 number of optimization iterations.

473 Additionally, the current framework assumes that performance improvements arise primarily from the
474 optimized demonstration order, but factors such as the inherent instability of large language models
475 (LLMs) across varying contexts might also contribute to observed fluctuations. Finally, the probing
476 set generation step introduces potential noise, and while the system attempts to mitigate this through
477 iterative updates, inaccuracies in probing may still affect the final demonstration order selection.

478 B Broader Impact

479 This work on Hierarchical Demonstration Order Optimization (HIDO) for many-shot in-context
480 learning has significant potential societal implications. Positively, by improving the performance
481 and reliability of large language models across diverse domains, HIDO could enhance AI applica-
482 tions in education, healthcare, and scientific research, making these systems more accessible and
483 effective for users without extensive prompt engineering expertise. However, potential negative
484 impacts include reinforcing existing biases in training data through optimized demonstration orders,
485 increasing compute requirements for determining optimal orderings (raising environmental and
486 resource accessibility concerns), and potentially widening capability gaps between organizations with
487 resources to implement such optimization techniques and those without. As many-shot in-context
488 learning becomes more widely deployed, careful consideration should be given to monitoring how
489 optimization techniques like HIDO affect fairness, bias, and resource distribution in AI systems.

490 C Licenses for existing assets

491 Datasets:

- 492 • AGNews [46]: asked the permission for using dataset per the data creator’s requirements.
- 493 • CB [7]: released under MIT License.
- 494 • CR [11]: publicly available for research use.
- 495 • DBPedia [46]: available under Creative Commons Attribution-ShareAlike License.
- 496 • MPQA [39]: Cite Wiebe et al. (2005) [38], available for research purposes.
- 497 • MR [33]: publicly available for research use.
- 498 • RTE [6]: publicly available for research use.
- 499 • SST-5 [35]: released under Stanford CoreNLP License.
- 500 • TREC [37]: available for research purposes.

501 Models:

- 502 • GPT-3.5-Turbo and GPT-4o-Mini [29, 30]: used under the OpenAI API Terms of Use.
- 503 • SciPhi-Mistral-7B-32k [13]: released under Apache 2.0 License.
- 504 • Zephyr-7b-beta [14]: Cite Huggingface [14], released under MIT License.
- 505 • LLaMa-3-8B-Instruct-Gradient-1048k [12]: released under Llama 3 Community License.

506 We have added this license information in Section 5.1 and ensured proper attribution throughout the
507 paper. All assets are used in accordance with their respective licenses and terms of use.

508 D Complete Development of ICD-OVI Metric

509 Enlightened by \mathcal{V} -usable information, our ICD-OVI, measures the usable information that an LLM
 510 can capture from ordered demonstrations $\Pi(\mathcal{D})$. First, we define the predictive family corresponding
 511 to the ordered demonstrations $\Pi(\mathcal{D})$ as

$$\mathcal{V}_{\Pi} := \{P_{\text{LLM}}(\cdot|\Pi(\mathcal{D}) \oplus q) | q \in \mathcal{Q}_P\} \cup \{P_{\text{LLM}}(\cdot|q) | q \in \mathcal{Q}_P\}, \quad (6)$$

512 where \mathcal{Q}_P represents the set of all possible queries in the sample space of input demonstrations' data
 513 distribution P , and $\{P_{\text{LLM}}(\cdot|q) | q \in \mathcal{Q}_P\}$ is added to satisfy the optimal ignorance requirement for a
 514 predictive family [42]. Then, ICD-OVI, the information that the model can capture from $\Pi(\mathcal{D})$, can
 515 be defined as the expected information the model with predictive family \mathcal{V}_{Π} can capture from query
 516 random variable (r.v.) Q for predicting label r.v. A , i.e.,

$$\begin{aligned} \text{ICD-OVI} &= H_{\mathcal{V}_{\Pi}}(A) - H_{\mathcal{V}_{\Pi}}(A|Q), \\ &= \inf_{f \in \mathcal{V}_{\Pi}} \mathbb{E}_{q, a \sim \mathcal{D}} [-\log f[\emptyset](a)] - \inf_{f \in \mathcal{V}_{\Pi}} \mathbb{E}_{q, a \sim \mathcal{D}} [-\log f[q](a)], \\ &= \mathbb{E}_{(q, a) \sim P} [\log_2 P_{\text{LLM}}(a|\Pi(\mathcal{A}) \oplus \emptyset) - \log_2 P_{\text{LLM}}(a|\Pi(\mathcal{D}) \oplus q)], \end{aligned} \quad (7)$$

517 where $\Pi(\mathcal{A}) := \bigoplus_{i=1}^n \mathcal{T}(\emptyset, a_{\pi(i)})$. The third equation follows the definition of in-context \mathcal{V} -
 518 information from Eq. 1 of [23]. Practically, denoting $P_{\text{LLM}}^i(\hat{a}) := P_{\text{LLM}}(\hat{a}|\hat{q}_i)$, we may approximate
 519 the Eq. 7 with the probing samples \hat{D} generated by LLM with

$$\frac{1}{|\hat{D}|} \sum_i (-\log_2 P_{\text{LLM}}^{\Pi, i}(\hat{a}) + \log_2 P_{\text{LLM}}^i(\hat{a})). \quad (8)$$

520 However, Eq. 8 involves the LLM-generated labels \hat{a} s for the probing samples, which can be factually
 521 incorrect. Utilizing those incorrect labels may lead to bias in the computation of ICD-OVI. Fortunately,
 522 the theory of V -usable information [9, 23] provide a effective tool called point-wise \mathcal{V} -informationn
 523 threshold (*PVI threshold*) which assists deciding if one generated sample label is reliable. Here, PVI
 524 is defined as

$$\text{PVI}_{(\hat{q}, \hat{a})}^{\Pi(\mathcal{D})} = -\log_2 P_{\text{LLM}}(\hat{a}|\Pi(\mathcal{D}) \oplus \hat{q}) + \log_2 P_{\text{LLM}}(\hat{a}|\Pi(\mathcal{A}) \oplus \hat{q}). \quad (9)$$

525 By Eq. 9, the ICD-OVI is the mean of PVIs for all probing samples \hat{D} . Built upon PVI, the PVI
 526 threshold is a scalar characterizing the likelihood of the correctness of the sample label. Specifically,
 527 when the PVI of a probing sample (\hat{q}, \hat{a}) is smaller than a constant τ , the label \hat{a} is possibly incorrect;
 528 otherwise, the label \hat{a} is highly likely to be correct for query \hat{q} . Actually, the fact of the existence of a
 529 PVI threshold is extensively validated by (author?) [9] and (author?) [23] in multiple LLMs and
 530 datasets of various semantic scenarios.

531 With the aid of the PVI threshold, we can address the potential bias caused by incorrect LLM-
 532 generated labels. Specifically, for a probing sample (\hat{q}, \hat{a}) , we first calculate its PVI; if it is higher
 533 than a predefined \mathcal{V} -information threshold τ , then we adopt the PVI of the sample (\hat{q}, \hat{a}) into the
 534 ICD-OVI calculation of ordered demonstrations $\Pi(\mathcal{D})$. Otherwise, we relax the PVI to its expectation
 535 for labels set $\{a | a \in \mathcal{A}\}$, i.e.,

$$\text{EPVI}_{(\hat{q}, \hat{a})}^{\Pi(\mathcal{D})} = \sum_{a \in \mathcal{A}} [-P_{\text{LLM}}^{\Pi, \hat{q}}(a) \log_2 P_{\text{LLM}}^{\Pi, \hat{q}}(a) + P_{\text{LLM}}^{\hat{q}}(a) \log_2 P_{\text{LLM}}^{\hat{q}}(a)]. \quad (10)$$

536 Conclusively, by denoting point-wise ICD-OVI (PICD-OVI) as

$$\text{PICD-OVI}_{(\hat{q}, \hat{a})}^{\Pi(\mathcal{D})} = \mathbb{I}(\text{PVI}_{(\hat{q}, \hat{a})} \geq \tau) \text{PVI}_{(\hat{q}, \hat{a})} + \mathbb{I}(\text{PVI}_{(\hat{q}, \hat{a})} < \tau) \text{EPVI}_{(\hat{q}, \hat{a})}, \quad (11)$$

537 our ICD-OVI can be approximated as

$$\text{ICD-OVI}(\Pi(\mathcal{D})) \approx \frac{1}{|\hat{D}|} \sum_{(\hat{q}, \hat{a})} \text{PICD-OVI}_{(\hat{q}, \hat{a})}. \quad (12)$$

538 Thus, our proposed ICD-OVI can effectively estimate the V -usable information despite noisy labels.

E Theorems and Proofs

Lemma 1. Let $f(x_1, \dots, x_n) = \sum_{i=1}^n x_i \log x_i$ be defined for $x_i > 0$, with the constraint $\sum_{i=1}^n x_i = c$, where $0 < c < \frac{1}{e}$. Then:

1. f reaches its minimum when all x_i are equal, i.e., $x_i = \frac{c}{n}$ for all i .
2. f reaches its maximum when one x_i equals c and the rest are zero.

Proof. We will use the method of Lagrange multipliers.

Let $g(x_1, \dots, x_n) = \sum_{i=1}^n x_i - c = 0$ be our constraint. The Lagrangian is:

$$L(x_1, \dots, x_n, \lambda) = \sum_{i=1}^n x_i \log x_i - \lambda \left(\sum_{i=1}^n x_i - c \right)$$

We set the partial derivatives to zero:

$$\frac{\partial L}{\partial x_i} = \log x_i + 1 - \lambda = 0 \quad \text{for } i = 1, \dots, n$$

$$\frac{\partial L}{\partial \lambda} = \sum_{i=1}^n x_i - c = 0$$

From $\frac{\partial L}{\partial x_i} = 0$, we get:

$$x_i = e^{\lambda-1}$$

This shows that all x_i are equal at the critical points.

Minimum Point: When all x_i are equal, let $x_i = \frac{c}{n}$ for all i . The function value is:

$$f\left(\frac{c}{n}, \dots, \frac{c}{n}\right) = c \log \frac{c}{n}$$

Maximum Point: Consider $x_1 = c$ and $x_i = 0$ for $i > 1$. The function value is:

$$f(c, 0, \dots, 0) = c \log c$$

To show that $f(\frac{c}{n}, \dots, \frac{c}{n}) < f(c, 0, \dots, 0)$, we need to prove:

$$c \log \frac{c}{n} < c \log c$$

This is equivalent to $\frac{c}{n} < c$, which is true for $n > 1$ and $c > 0$. Therefore, we have shown that the minimum occurs when all $x_i = \frac{c}{n}$, and the maximum occurs when one $x_i = c$ and the rest are zero. \square

Theorem 1 We assume that given a LLM, a probing sample (\hat{q}, \hat{a}) and an ordered demonstration text $\Pi(\mathcal{D})$,

- When $PVI_{(\hat{q}, \hat{a})}^{\Pi(\mathcal{D})} \geq \tau$, then $\hat{a} = a^*$, where the a^* is the ground-truth label corresponding to the generated query \hat{q} .
- The LLM predict the label \hat{a} with the highest probability when query by \hat{q} with $\Pi(\mathcal{D})$ as its context, i.e., $P(\hat{a} | \Pi(\mathcal{D}) \oplus \hat{q}) = \arg \max_{a \in \mathcal{A}} P(a | \Pi(\mathcal{D}) \oplus \hat{q})$.
- Assume that for any two ordered demonstration texts $\Pi_1(\mathcal{D})$ and $\Pi_2(\mathcal{D})$, the $P_{LLM}(a | \Pi_1(\mathcal{A}) \oplus \emptyset) = P_{LLM}(a | \Pi_2(\mathcal{A}) \oplus \emptyset)$ for all $a \in \mathcal{A}$.

Without loss of generalizability, for any two ordered demonstrations $\Pi_1(\mathcal{D})$ and $\Pi_2(\mathcal{D})$, there is a $\epsilon (\frac{1}{e} \leq \epsilon \leq 1)$ such that $P(\hat{a} | \Pi_i(\mathcal{D}) \oplus \hat{q}) > 1 - \epsilon$. We additionally assume that when $PVI_{(\hat{q}, \hat{a})}^{\Pi(\mathcal{D})} < \tau$:

566 • The a^* is the second most probable label given by the LLM when prompted by
 567 query \hat{q} with any ordered demonstration context $\Pi(\mathcal{D})$, i.e., $P(a^*|\Pi(\mathcal{D}) \oplus \hat{q}) =$
 568 $\arg \max_{a \in \mathcal{A} \setminus \{\hat{a}\}} P(a|\Pi(\mathcal{D}) \oplus \hat{q})$; we write $P(a^*|\Pi_i(\mathcal{D}) \oplus \hat{q}) = \lambda_i \epsilon$, where $0 \leq \lambda_i \leq 1$,
 569 $i \in \{1, 2\}$.

570 • By symmetry, we only consider the case $\lambda_1 < \lambda_2$. In this case, we assume that $\frac{1}{2} - \delta <$
 571 $\lambda_1 < \frac{1}{2} + \delta$ (δ is a constant) such that

$$(\lambda_1 \epsilon) \log \lambda_1 \epsilon + (1 - \lambda_1 \epsilon) \log (1 - \lambda_1 \epsilon) < \epsilon \log \epsilon - (2 - \lambda_1) \epsilon. \quad (13)$$

572 Meanwhile, we require $\lambda_2 - \lambda_1 > (1 - \frac{1}{\log(n-2)})(1 - \lambda_1)$.

573 With the assumptions above, if

$$PICD-OVI_{(\hat{q}, \hat{a})}^{\Pi_1(\mathcal{D})} > PICD-OVI_{(\hat{q}, \hat{a})}^{\Pi_2(\mathcal{D})}, \quad (14)$$

574 then we have

$$PVI_{(\hat{q}, a^*)}^{\Pi_1(\mathcal{D})} > PVI_{(\hat{q}, a^*)}^{\Pi_2(\mathcal{D})}. \quad (15)$$

575 Therefore, if $\Pi_1(\mathcal{D})$ is more performant demonstration order than $\Pi_2(\mathcal{D})$, i.e., Eq. 15 establish for
 576 any probing sample (\hat{q}, \hat{a}) , then

$$ICD-OVI(\Pi_1(\mathcal{D})) > ICD-OVI(\Pi_2(\mathcal{D})). \quad (16)$$

577 *Proof.* First, in the case that $PVI_{(\hat{q}, \hat{a})}^{\Pi(\mathcal{D})} \geq \tau$, by Assumption 1, we have $\hat{a} = a^*$. Therefore, we have

$$PICD-OVI_{\hat{q}, \hat{a}}^{\Pi(\mathcal{D})} = P(\hat{a}|\Pi(\mathcal{D}) \oplus \hat{q}) - P(\hat{a}|\Pi(\mathcal{A}) \oplus \emptyset) = PVI_{\hat{q}, \hat{a}}^{\Pi(\mathcal{D})} = PVI_{\hat{q}, a^*}^{\Pi(\mathcal{D})}. \quad (17)$$

578 Eq. 17 enforces the establishment of Eq. 15.

579 Next, in the case where $PVI_{(\hat{q}, \hat{a})}^{\Pi(\mathcal{D})} < \tau$, with Assumption 3, it suffices to prove that $|\lambda_1 \epsilon \log \lambda_1 \epsilon| \geq$
 580 $|\lambda_2 \epsilon \log \lambda_2 \epsilon|$ gives rise to

$$|\lambda_1 \epsilon \log \lambda_1 \epsilon + \sum_{\Sigma_i x_i = (1 - \lambda_1) \epsilon} x_i \log x_i + x_{\hat{a}, 1}| \geq |\lambda_2 \epsilon \log \lambda_2 \epsilon + \sum_{\Sigma_i x_i = (1 - \lambda_1) \epsilon} x_i \log x_i + x_{\hat{a}, 2}|. \quad (18)$$

581 Now, by utilizing the Assumption 5, we claim that Eq. 18 establish, thus the theorem is proved.

582 To prove Eq. 18, we start from the known inequality

$$\lambda_2 - \lambda_1 > (1 - \frac{1}{\log(n-2)})(1 - \lambda_1). \quad (19)$$

583 For simplicity, we represent $\lambda_2 - \lambda_1$ as Δ in the following texts. We rewrite the Eq. 19 as

$$\begin{aligned} \Delta &> \frac{1 + 1/\epsilon \log e^{-\epsilon(1 - \lambda_1) - \epsilon + \log 2/2}}{\log(n-2)} + (1 - \lambda_1), \\ &= \frac{1}{\epsilon} \left[\frac{\epsilon(\log \epsilon - \log 2) + (\epsilon \log 2 - \epsilon(2 - \lambda_1))}{\log(n-2)} \right] - \frac{\log \epsilon}{\log(n-2)} + (1 - \lambda_1) + \frac{1}{\log(n-2)}. \end{aligned} \quad (20)$$

584 By Assumption 5, we substitute terms appears in Eq. 20 with left hand side (LHS) of Eq. 13 and
 585 $\log[1 - \lambda_1 \epsilon] > \log[(1 - \lambda_1) \epsilon]$, further relax the bound as

$$\begin{aligned} \Delta &> \lambda_1 \frac{\log \lambda_1 \epsilon}{\log(n-2)} - \frac{\log \epsilon}{\log(n-2)} + (1 - \lambda_1) + \frac{1 - \lambda_1}{\log(n-2)} \log[(1 - \lambda_1) \epsilon] + \frac{1}{\log(n-2)} \\ &= -\frac{1}{\epsilon \log(n-2)} \{ -\lambda_1 \epsilon \log \lambda_1 \epsilon + \epsilon \log \epsilon - [(1 - \lambda_1) \epsilon] \log(n-2) - (1 - \lambda_1) \epsilon \log[(1 - \lambda_1) \epsilon] - \epsilon \}. \end{aligned} \quad (21)$$

586 By multiplying $\epsilon \log(n-2)$ to both sides of the inequality, we have

$$-\lambda_1 \epsilon \log(\lambda_1 \epsilon) + \epsilon \log \epsilon - [\log(n-2)](1 - \lambda_1 - \Delta) \epsilon - (1 - \lambda_1) \epsilon \log(1 - \lambda_1) \epsilon - \epsilon > 0. \quad (22)$$

587 Eq. 22 is equivalent to

$$-\lambda_1 \epsilon \log \lambda_1 \epsilon + \log \epsilon (\lambda_1 + \Delta) \epsilon + (n-2) \frac{(1 - \lambda_1 - \Delta) \epsilon}{n-2} \log \frac{\epsilon}{n-2} - (1 - \lambda_1) \epsilon \log(1 - \lambda_1) \epsilon - \epsilon > 0. \quad (23)$$

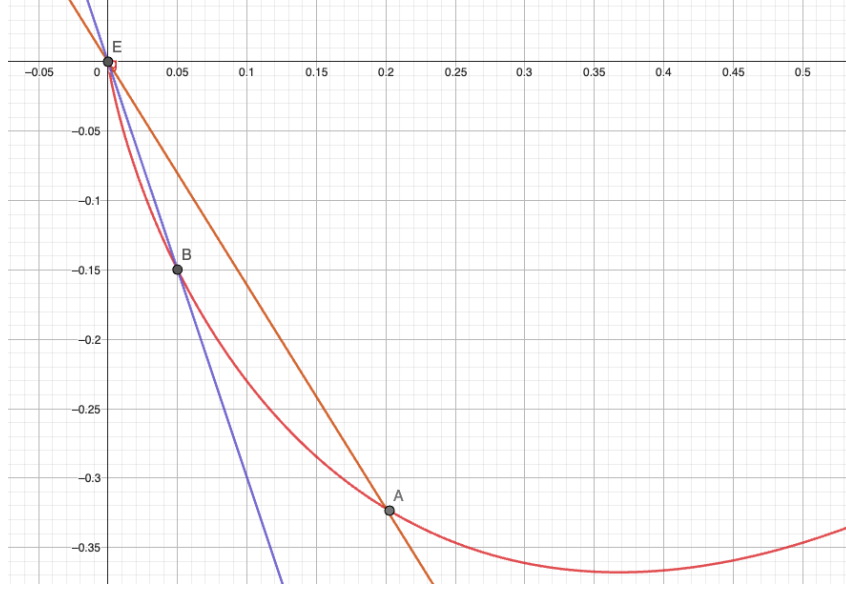


Figure 4: Illustration of the observation of Eq. 24 and Eq. 25. The red, orange, and blue curves are $x \log x$, $\log \epsilon x$ and $\log \frac{\epsilon}{n-2} x$ (where $n = 6$ and $\epsilon = 0.2$), respectively. It is clear that $x \log x \leq \log \epsilon x$ between point E and A ; $x \log x \leq \log \frac{1}{n-2} \epsilon x$ between point E and B .

Now, we observe that since $\lambda_1 + \Delta = \lambda_2 < 1$, thus $(\lambda_1 + \Delta)\epsilon < \epsilon$. Therefore

$$(\lambda_1 + \Delta)\epsilon \log (\lambda_1 + \Delta)\epsilon \leq -\log \epsilon (\lambda_1 + \Delta)\epsilon. \quad (24)$$

Here, the $\log \epsilon$ is the slope of the linear function composed by $(0, 0)$ and $(\epsilon, \epsilon \log \epsilon)$. Analogously, we have

$$\frac{1 - \lambda_1 - \Delta}{\epsilon} \log \frac{1 - \lambda_1 - \Delta}{n-2} \epsilon \leq \log \frac{\epsilon}{n-2} \frac{(1 - \lambda_1 - \Delta)\epsilon}{n-2}. \quad (25)$$

By substituting the terms of RHS of Equ 24 and Equ 25 appeared in Equ 23 with the LHS of Equ 24 and Equ 25, we further relax our inequality as

$$-\lambda_1 \epsilon \log \lambda_1 \epsilon + (\lambda_1 + \Delta)\epsilon \log (\lambda_1 + \Delta)\epsilon + (1 - \lambda_1 - \Delta)\epsilon \log \left(\frac{1 - \lambda_1 - \Delta}{n-2} \epsilon \right) - (1 - \lambda_1)\epsilon \log (1 - \lambda_1)\epsilon + (1 - \epsilon) \log (1 - \epsilon) > 0. \quad (26)$$

We now rearrange the Eq. 26 and substitute $\lambda_1 + \Delta$ with λ_2 , we have

$$-\lambda_1 \epsilon \log \lambda_1 \epsilon - (1 - \lambda_1)\epsilon \log (1 - \lambda_1)\epsilon > -(\lambda_2 \epsilon) \log (\lambda_2 \epsilon) - (1 - \lambda_2)\epsilon \log \left(\frac{1 - \lambda_1 - \Delta}{n-2} \epsilon \right) - (1 - \epsilon) \log (1 - \epsilon). \quad (27)$$

We observe that, by Lemma 1, we have that

$$\begin{aligned} \min_{(x_1, \dots, x_{n-2})} \sum_{\Sigma x_i = (1 - \lambda_2)\epsilon} x_i \log x_i &= (1 - \lambda_2)\epsilon \log \left(\frac{1 - \lambda_2}{n-2} \epsilon \right), \\ \max_{(x_1, \dots, x_{n-2})} \sum_{\Sigma x_i = (1 - \lambda_1)\epsilon} x_i \log x_i &= (1 - \lambda_1)\epsilon \log (1 - \lambda_1)\epsilon. \end{aligned} \quad (28)$$

In other words,

$$\begin{aligned} \max_{(x_1, \dots, x_{n-2})} |\sum_{\Sigma x_i = (1 - \lambda_2)\epsilon} x_i \log x_i| &= -(1 - \lambda_2)\epsilon \log \left(\frac{1 - \lambda_2}{n-2} \epsilon \right), \\ \min_{(x_1, \dots, x_{n-2})} |\sum_{\Sigma x_i = (1 - \lambda_1)\epsilon} x_i \log x_i| &= -(1 - \lambda_1)\epsilon \log (1 - \lambda_1)\epsilon. \end{aligned} \quad (29)$$

Besides, it is direct to show that

$$(1 - \epsilon) \log (1 - \epsilon) \leq x_{\hat{a}, i} \log x_{\hat{a}, i} \leq 0, \quad (30)$$

597 i.e.,

$$-(1 - \epsilon) \log(1 - \epsilon) \geq |x_{\hat{a},i} \log x_{\hat{a},i}| \geq 0, \quad (31)$$

598 Hence, we rewrite the Eq. 27 to

$$\begin{aligned} & |\lambda_1 \epsilon \log \lambda_1 \epsilon| + \min_{(x_1, \dots, x_{n-2})} |\sum_{\Sigma x_i = (1-\lambda_1)\epsilon} x_i \log x_i| + \min |x_{\hat{a},1} \log x_{\hat{a},1}| > \\ & |(\lambda_2 \epsilon) \log(\lambda_2 \epsilon)| + \max_{(x_1, \dots, x_{n-2})} |\sum_{\Sigma x_i = (1-\lambda_2)\epsilon} x_i \log x_i| + \max |x_{\hat{a},2} \log x_{\hat{a},2}|. \end{aligned} \quad (32)$$

599 Therefore, we are able to write that

$$|\lambda_1 \epsilon \log \lambda_1 \epsilon + \sum_{\Sigma x_i = (1-\lambda_1)\epsilon} x_i \log x_i + x_{\hat{a},1}| \geq |\lambda_2 \epsilon \log \lambda_2 \epsilon + \sum_{\Sigma x_i = (1-\lambda_1)\epsilon} x_i \log x_i + x_{\hat{a},2}|, \quad (33)$$

600 which is exactly Eq. 18. \square

601 **Theorem 2.** Randomly flipping K entries from a sequence of length N will always keep the rank
602 correlation within a range characterized by the lower bound $1 - 6 \sum_{i=1}^K (a_i - a_{K+1-i})^2 / N(N^2 - 1)$
603 and upper bound 1. Here a_i is the original position index of the i -th perturbed element. The lower
604 bound is achieved with a probability of $1/K!$ when the perturbed sequence is the reverse of the
605 original sequence. The upper bound is achieved with a probability of $1/K!$ when the perturbed
606 sequence is identical to the original sequence.

607 To prove the above theorem, we first present the lemma:

608 **Lemma 2.** Given a list of N integers $\{a_1, a_2, \dots, a_N\}$ with $a_i < a_{i+1}, i = 1, 2, \dots, N-1$ and its
609 random perturbation $\{a_1^*, a_2^*, \dots, a_N^*\}$, the maximum value of $\sum_{i=1}^N (a_i - a_i^*)^2$ is achieved by reversing
610 the list, i.e., $a_i^* = a_{N+1-i}$.

611 *Proof.* To prove that the maximum value of the sum:

$$S = \sum_{i=1}^N (a_i - a_i^*)^2$$

612 is achieved by reversing the list $\{a_i^*\}_{i=1}^N$, we need to show that this arrangement maximizes the
613 squared differences between the original list $\{a_i\}_{i=1}^N$ and the perturbed list $\{a_i^*\}_{i=1}^N$, where a_i^* is the
614 perturbed element in the i -th position.

615 We know that

$$a_1 < a_2 < \dots < a_N.$$

616 Considering the sum $S = \sum_{i=1}^N (a_i - a_i^*)^2$, each term in this sum is of the form $(a_i - a_i^*)^2$, which
617 measures how far apart a_i and a_i^* are. Thus, to maximize the sum, we need to maximize each
618 individual squared difference $(a_i - a_i^*)^2$.

619 The largest possible difference between any two elements of the list $\{a_i\}_{i=1}^N$ occurs when the largest
620 element a_N is paired with the smallest element a_1 , the second largest element a_{N-1} is paired with
621 the second smallest element a_2 , and so on. In other words, the maximum possible difference occurs
622 when $a_i^* = a_{N+1-i}$ for all i . This arrangement is precisely the reverse of the original list.

623 To prove that reversing the list maximizes the sum, we propose to prove that when swapping any two
624 elements in the perturbed list, the sum will always decrease. Suppose we swap two elements a_p^* and
625 a_q^* (with $p < q$, without loss of generality) in the reversed list. Before the swap, the contributions to
626 the sum from the two positions are:

$$(a_p - a_p^*)^2 + (a_q - a_q^*)^2.$$

627 After swapping a_p^* and a_q^* , the new contributions become:

$$(a_p - a_q^*)^2 + (a_q - a_p^*)^2.$$

628 The change in the sum, ΔS , is the difference between these two expressions:

$$\Delta S = ((a_p - a_q^*)^2 + (a_q - a_p^*)^2) - ((a_p - a_p^*)^2 + (a_q - a_q^*)^2).$$

629 We expand these terms as follows:

630 - Before the swap:

$$(a_p - a_p^*)^2 + (a_q - a_q^*)^2 = (a_p - a_{N+1-p})^2 + (a_q - a_{N+1-q})^2$$

631 - After the swap:

$$(a_p - a_q^*)^2 + (a_q - a_p^*)^2 = (a_p - a_{N+1-q})^2 + (a_q - a_{N+1-p})^2$$

632 Because $a_p < a_q$ and the list is ordered, swapping two elements in the reversed list *decreases* the
633 squared differences, leading to a decrease in the sum S . Thus, reversing the list maximizes the
634 absolute differences $|a_i - a_i^*|$ for all i , and any deviation from the reversed order will result in a
635 smaller sum. \square

636 With this lemma, now we prove Theorem 2.

637 *Proof.* Given two ranking sequences $\{s_i\}_{i=1}^N$ and $\{s_i^*\}_{i=1}^N$, the Spearman’s rank correlation coeffi-
638 cient is represented as follows:

$$\rho = 1 - \frac{6 \sum_{i=1}^N (s_i - s_i^*)^2}{N(N^2 - 1)}. \quad (34)$$

639 In our case, one ranking sequence is obtained by perturbing K elements in another ranking sequence.
640 Denote the selected elements as $\{a_i\}_{i=1}^K$, and the elements after perturbation as $\{a_i^*\}_{i=1}^K$

641 according to Lemma 2, we know the maximum value of $\sum_{i=1}^K (a_i - a_i^*)^2$ is achieved when $a_i^* = a_{K+1-i}$.

642 For other elements that are not perturbed satisfy that their d_i equals 0. Therefore, the Spearman’s
643 rank correlation coefficient reaches the minimum value:

$$\rho_{\min} = 1 - \frac{6 \sum_{i=1}^K (a_i - a_{K+1-i})^2}{N(N^2 - 1)}. \quad (35)$$

644 Similarly, the maximum value is $\rho_{\max} = 1$ when the perturbed sequence is exactly the same as
645 the original sequence. Since each perturbation has an equal probability, and there are $K!$ different
646 perturbations, we know the probabilities are both $1/K!$. \square

647 F Supplementary Experiments

648 F.1 Implementation Details

649 Our conduct experiments using a system equipped with four NVIDIA A100 80GB PCIe GPUs.
650 The system ran NVIDIA driver version 550.54.14 and CUDA 12.4. We implement the project with
651 Python, mainly relying on the PyTorch [34] and Transformers [40] packages for the implementation.

652 F.2 Ablation Experiment Results

653 F.3 Accuracy difference between few-shot (10 shots) and many-shot (150) ICL

654 As mentioned earlier, we want to confirm that ICL-DOI still exists in many shot ICL. Thus, we
655 randomly select orders with 10 or 150 demonstrations and measure the model accuracy. The following
656 figures present the distribution of model performance under few-shot and many-shot settings on
657 various datasets.

658 F.4 Quantitative Analysis of Generated Probing Sets

659 In the method development, we assume that the demonstration order optimized for answer prediction
660 can also be used for sample generation. Since each additional iteration of HIDO optimizes the order

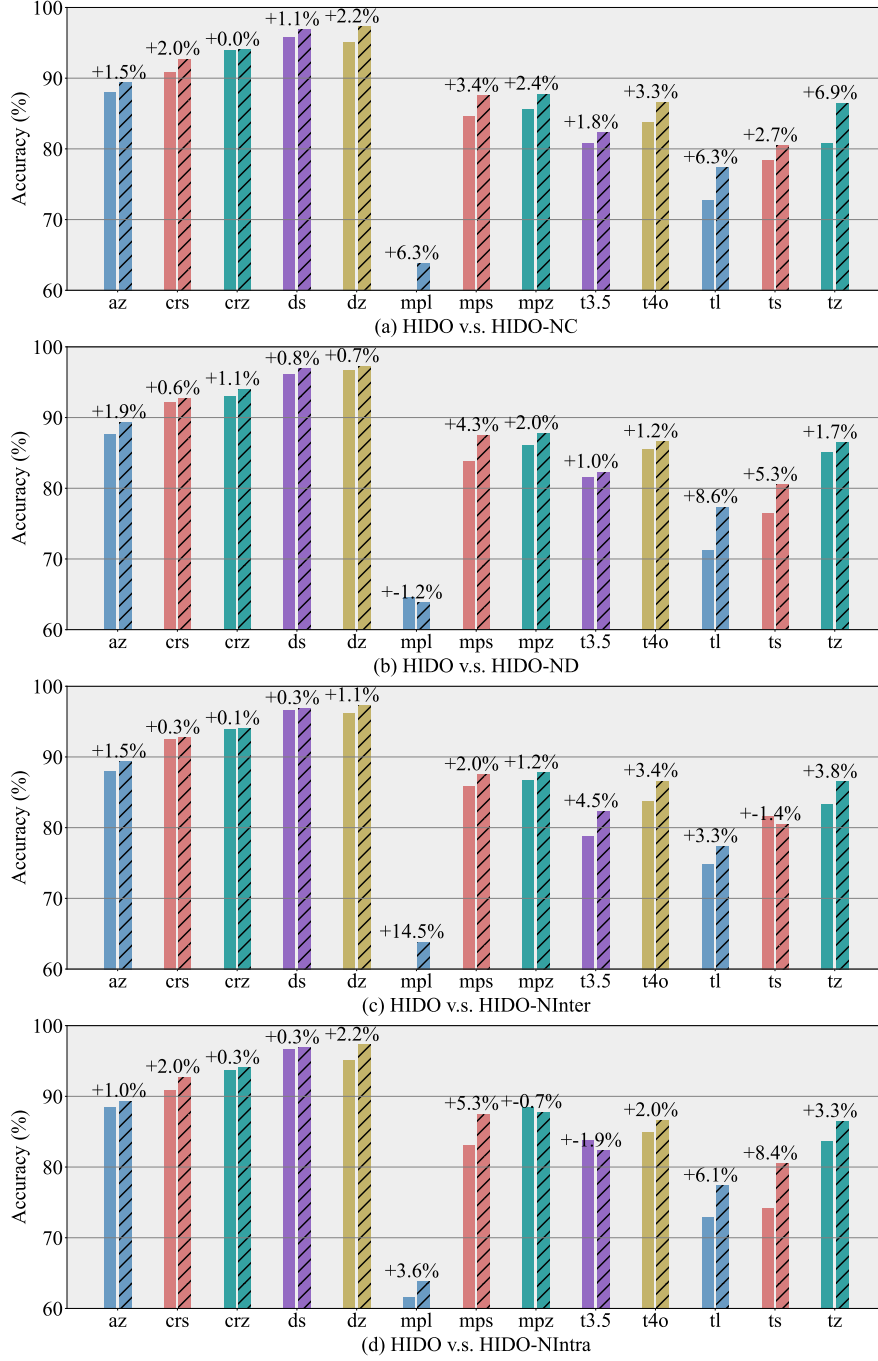


Figure 5: The performance of our proposed HIDO and its variants tested with different LLMs on various datasets. The first one or two characters indicate the dataset (i.e. 't' represents TREC and 'mp' represents 'MPQA'). The remaining characters represent the model (i.e. 'z' represents Zephyr and '3.5' represents GPT-3.5T).

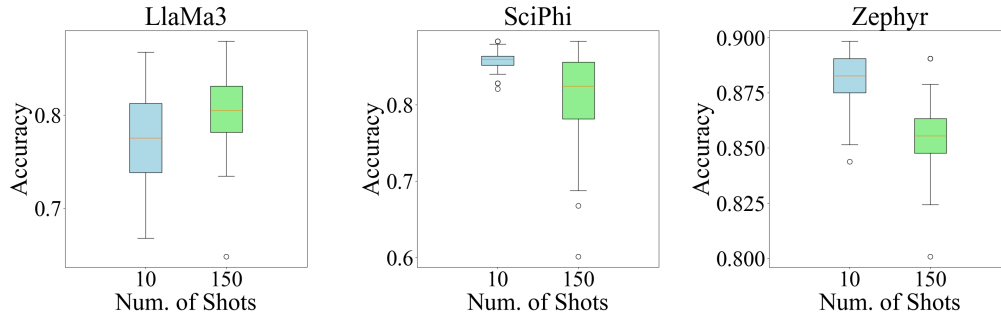


Figure 6: AGNews. Many shot ICL generally improves the best model accuracy (i.e. increases maximum accuracy), which causes the range to be larger.

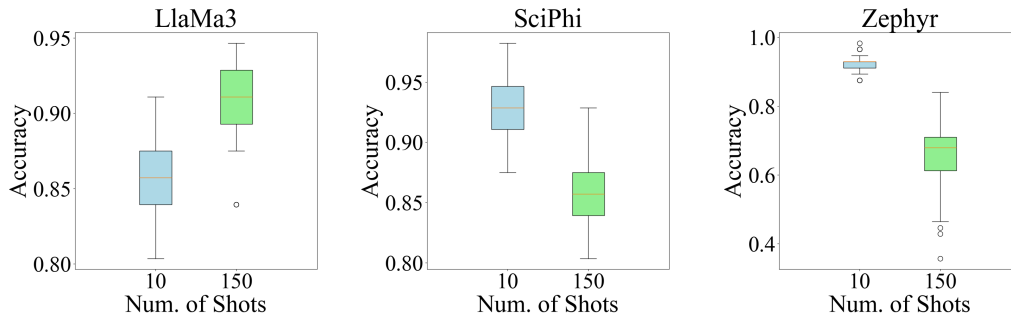


Figure 7: CB. Here, the figure shows that many shot learning causes model performance to degrade. This could be a result of CB having less test samples (56 samples compared to 256 samples for other datasets). Regardless, there is large variance in the results, indicating demonstration order instability.

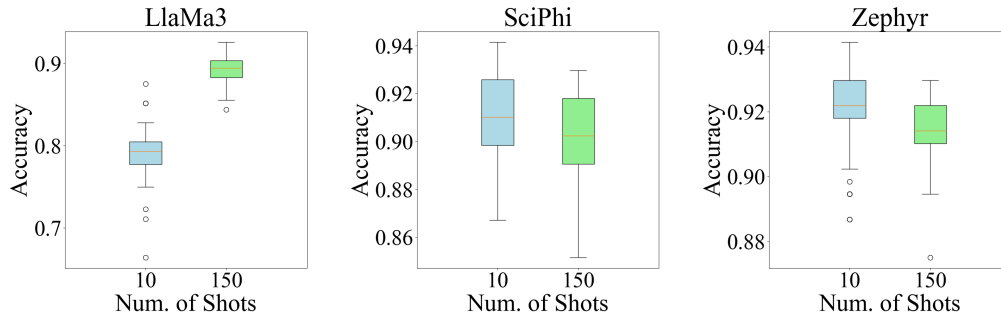


Figure 8: CR. SciPhi and Zephyr exhibit a wider variance in accuracy. In Zephyr, there is an extremely low outlier, emphasizing the importance of order on model performance.

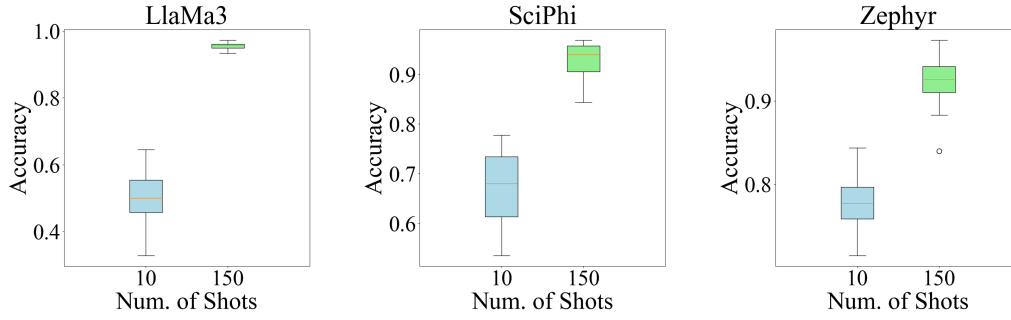


Figure 9: DBPedia. Many shot learning improved model performance for all models; however, for LLaMa3, the variance becomes smaller but stays the same or increases for the other models. Taking a look at DBPedia, the samples in general give more context in comparison to the others, which suggests that LLaMa3 is better at retaining and exploiting the information given from the demonstrations when completing the task of interest.

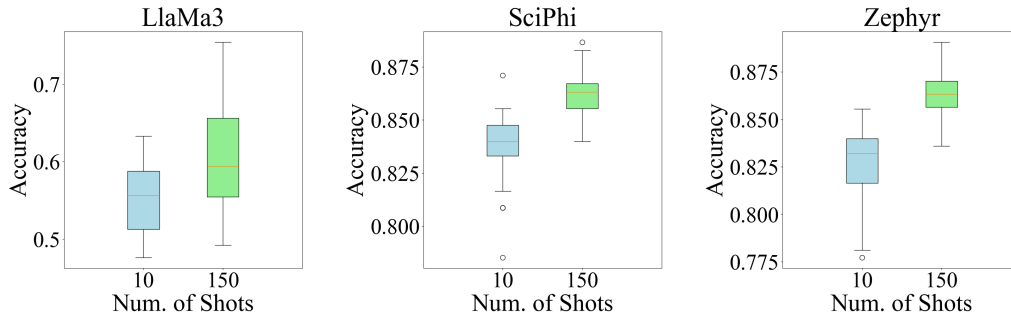


Figure 10: MPQA. Again, many shot ICL improved model accuracy, but also caused the variance to increase in general. LLaMa3 especially exhibits the problem of ICL-DOI with over 25% difference between the best and worst accuracy.

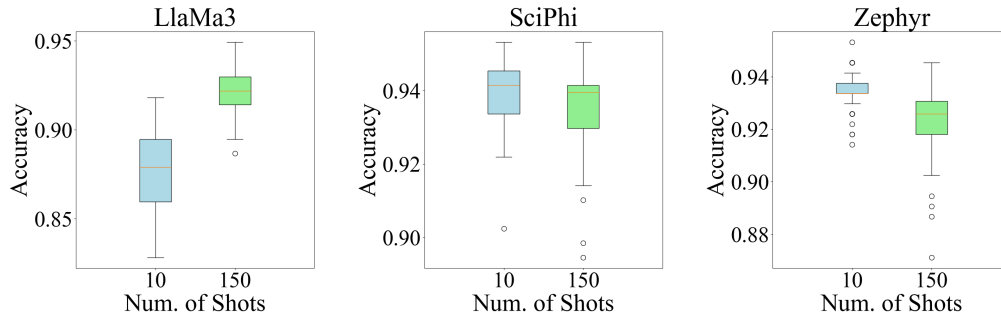


Figure 11: MR. Model performance only improved for LLaMa3, but the other two models illustrate a wider variance. For SciPhi and Zephyr, the model performance under the few-shot and many shot settings is comparable, but in many-shot, the worst accuracy is much lower than that of few-shot performance.

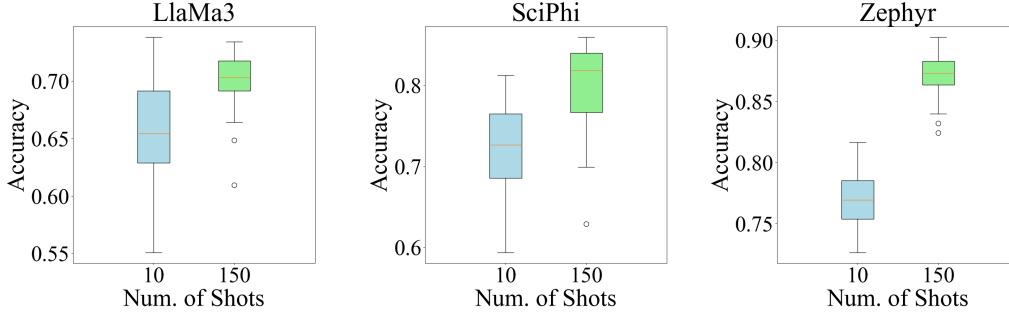


Figure 12: TREC. Increasing the number of demonstrations increased average accuracy for all models, and the variance did not improve much, other than for LLaMa3. LLaMa3 has 8 billion paramters, compared to only 7 billion for the other two models, which means that it has more capability to learn and retain information. This can potentially be the reason for its superior performance against the other two LLMs.

661 such that it can achieve a higher accuracy, the probing set from the inter-cluster optimization round is
 662 generated from the current optimized order. Thus, we can compare the probing set to the original
 663 demonstrations, which should be of high quality. Ideally, as the number of iterations increases (i.e.
 664 the order becomes more optimized), the distance between the two should decrease (i.e. the quality
 665 of the probing set increases). The following figures measure the average L_2 norm between the
 666 demonstration embeddings and the probing set embeddings generated by various LLMs on different
 667 datasets. In general, the experiments support the assumption, presenting a negative trend between
 668 iterations and distance.

669 F.5 Example samples from each dataset

670 Below, we provide some samples in each dataset, which can be compared to the probing sets presented
 671 in F.6.

672 F.6 Qualitative Analysis of Generated Probing Sets

673 In addition to Appendix F.4, we display the generated probing sets, along with example samples from
 674 each dataset, to qualitatively analyze the generated text. Because the embeddings may not completely
 675 capture the semantics and syntax of the text, we want to use human evaluation to determine if the
 676 quality of the probing set improves.

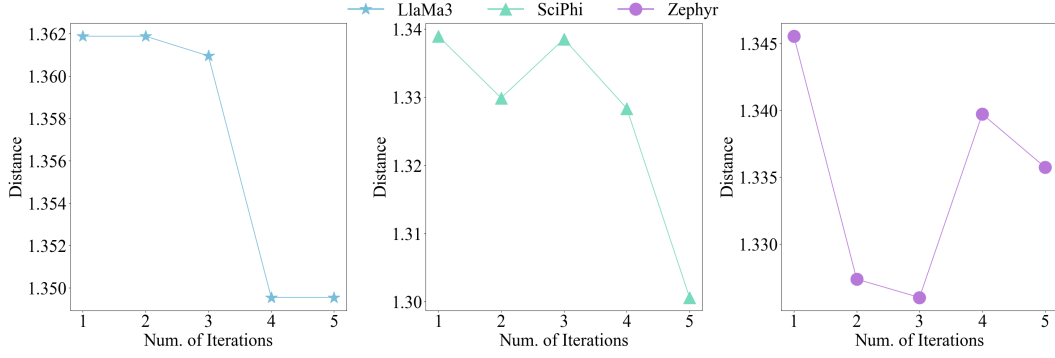


Figure 13: AGNews. Embedding distance for both LLaMa3 and Zephyr consistently decrease as the number of iterations increase; however, Zephyr reaches its optimal at three iterations, and additional iterations will cause the resulting order to deviate, as indicated by the spike at the fourth iteration. SciPhi has a peak at three iterations but decreases after that point.

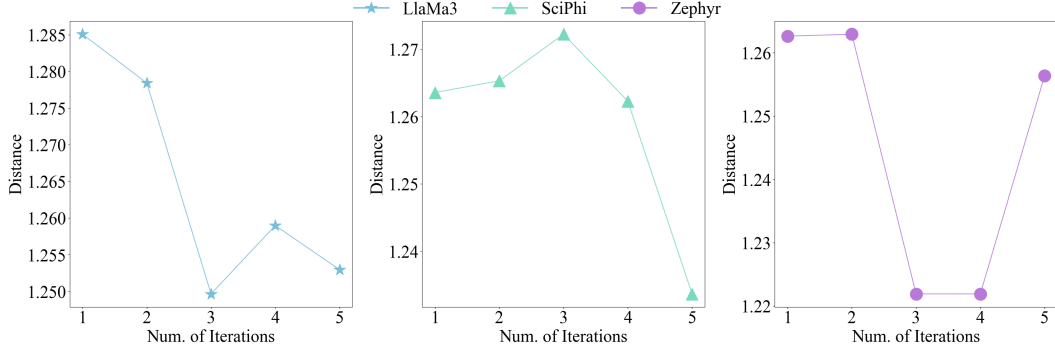


Figure 14: CR. Similar to the previous figure, the probing sets generated by LLaMa3 and Zephyr consistently drop, and SciPhi displays a peak and then a major drop in embedding distance. The figures suggest that after some iterations (i.e. as the order becomes more optimal), the LLM can generate samples close to the original text.

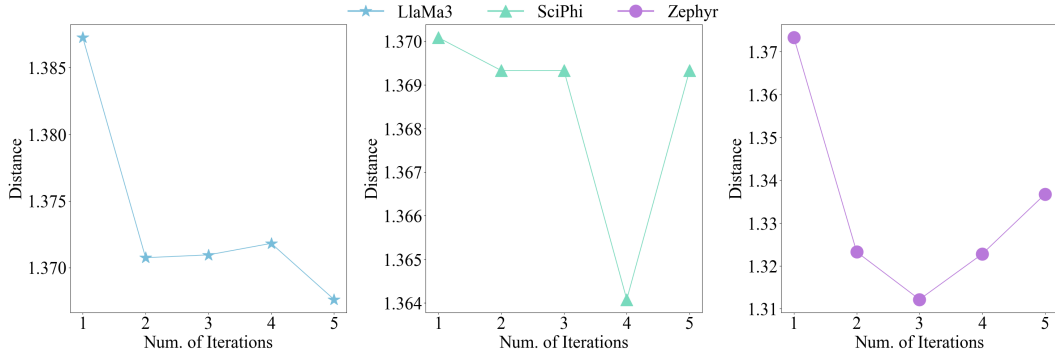


Figure 15: DBpedia. All models demonstrate a negative trend between distance and iteration. The figure for SciPhi displays a plateau between the second and third iteration, which could imply that the probing set (i.e. the actual text) or the semantics did not change much.

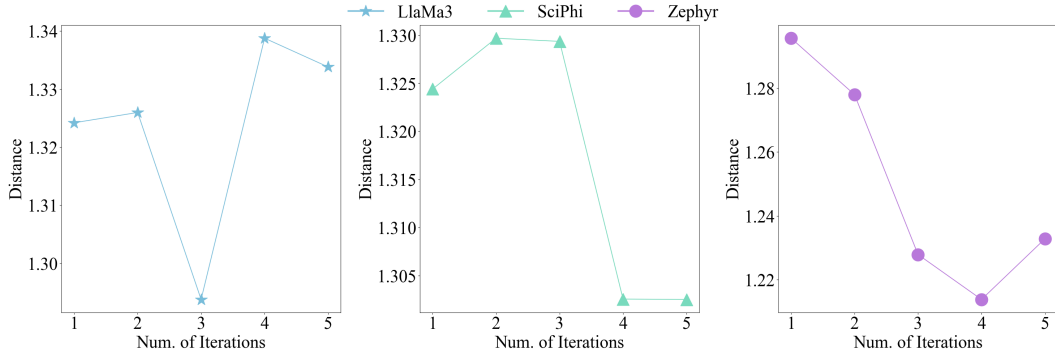


Figure 16: MPQA. The figures in general demonstrate a negative trend. For SciPhi, the distance increases first then drops after the second iteration. However, the difference is relatively small, about 0.05 difference, indicating that the generated samples are similar to the demonstrations.

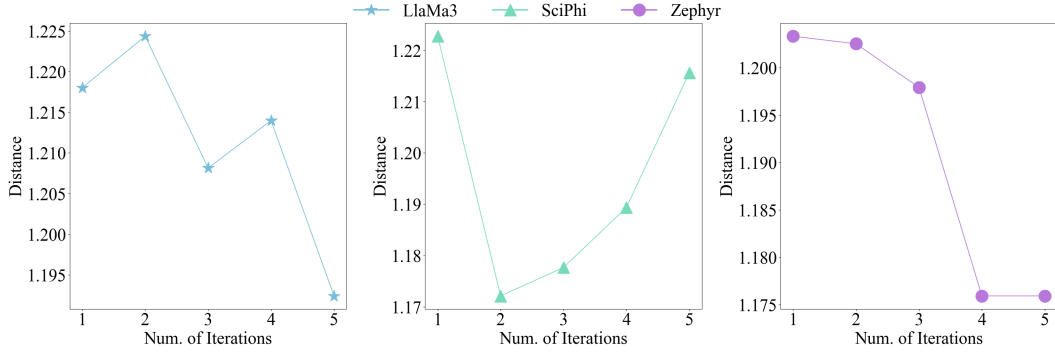


Figure 17: MR. For LLaMa3, the distance peaks at iteration two and iteration four, but generally decreases. This could be due to HIDO trying to find the best order in the neighborhood space but selecting one that does not perform well; however, it is able to find the best order in the end.

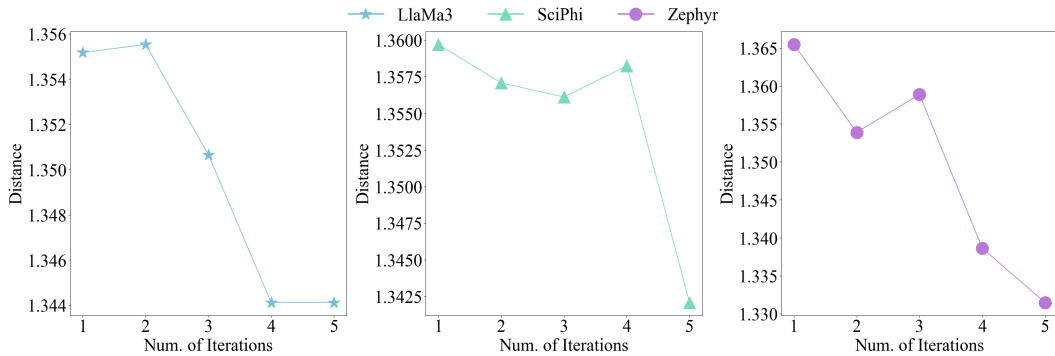


Figure 18: TREC. Like before, the general trend is negative in all the figures. However, the plots for SciPhi and Zephyr both have a peak but drops in the next iteration, which indicates that the model diverges from the optimal and corrects itself.