

Additional file 2:

Advancing Generative Large Language Models
Toward Discriminative Performance in Protein
Function Prediction

Lv et al.

Detailed descriptions of evaluation metrics

In this study, while our model is a generative large language model (LLM), the tasks under investigation include both classification and text generation. We therefore adopt metrics commonly used for classification tasks, complemented by standard evaluation metrics for LLMs. First, we define general notation:

The test set is denoted as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$, where m represents the number of samples. For single-label classification (subcellular localization prediction), let K denote the number of classes. The true label of sample i is $y_i \in \{1, \dots, K\}$, and the predicted label is $\hat{y}_i \in \{1, \dots, K\}$. For multi-label classification (GO term prediction, EC number prediction, UniProt keyword prediction), let L denote the total number of unique labels. Each sample is represented by binary vectors: $\mathbf{y}_i = (y_{i1}, \dots, y_{iL}) \in \{0,1\}^L$ (true label) and $\hat{\mathbf{y}}_i = (\hat{y}_{i1}, \dots, \hat{y}_{iL}) \in \{0,1\}^L$ (predicted label), where $y_{ij} = 1$ indicates sample i is annotated with label j . For functional description generation, the reference text for sample i is r_i , and the model-generated text is c_i . Any text-based metric is denoted as $s(r_i, c_i)$.

For subcellular localization prediction, we use *Accuracy* to quantify overall prediction performance:

$$\text{Accuracy} = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(\hat{y}_i = y_i)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function (returning 1 if the condition is satisfied, 0 otherwise).

For GO term prediction, EC number prediction, and UniProt keyword prediction, we adopt Precision (P), Recall (R), and F1-score (F1) for per-sample evaluation, then average across the test set to obtain corpus-level results. For a single sample:

Define true positives (TP), false positives (FP), and false negatives (FN):

$$TP = \sum_{j=1}^L \mathbb{I}(y_j = 1 \wedge \hat{y}_j = 1)$$
$$FP = \sum_{j=1}^L \mathbb{I}(y_j = 0 \wedge \hat{y}_j = 1)$$

$$FN = \sum_{j=1}^L \mathbb{I}(y_j = 1 \wedge \hat{y}_j = 0)$$

Then the Precision (P), Recall (R), and F1-score (F1) for this sample are defined as:

$$\begin{aligned} P &= \frac{TP}{TP + FP} \\ R &= \frac{TP}{TP + FN} \\ F1 &= \frac{2PR}{P + R} \end{aligned}$$

For functional description generation, we evaluate generated text quality using 9 widely adopted LLM metrics. For each sample, we compute $s(r_i, c_i)$, and the corpus-level score is the average across all test samples is $S = \frac{1}{m} \sum_i s(r_i, c_i)$.

1) ROUGE series: ROUGE-1 (R1, unigram recall), ROUGE-2 (R2, bigram overlap), ROUGE-L (RL, sentence-level longest common subsequence), and ROUGE-Lsum (RLS, document-level longest common subsequence across concatenated references).

For ROUGE-1 and ROUGE-2, let $\mathcal{G}_n(t)$ denote the multiset of n -grams (including repetitions) in text t , and $\text{cnt}_t(g)$ denote the count of n -gram g in t . The sample-level score is:

$$\text{ROUGE} - n(r, c) = \frac{\sum_{g \in \mathcal{G}_n(r)} \min(\text{cnt}_r(g), \text{cnt}_c(g))}{\sum_{g \in \mathcal{G}_n(r)} \text{cnt}_r(g)}$$

Corpus-level score (average of sample-level scores):

$$Rn_{\text{corpus}} = \frac{1}{N} \sum_{k=1}^N \text{ROUGE} - n(r^{(k)}, c^{(k)})$$

For ROUGE-L, let $\text{LCS}(r, c)$ denote the length of the LCS between r and c , and $|r|$ denote the reference length (tokens). Sample-level score:

$$\text{ROUGE} - L(r, c) = \frac{\text{LCS}(r, c)}{|r|}$$

Corpus-level score:

$$RL_{\text{corpus}} = \frac{1}{N} \sum_{k=1}^N \text{ROUGE} - L(r^{(k)}, c^{(k)})$$

For ROUGE-Lsum, concatenating all references for a sample into $\tilde{r}$, and denote

generated text as $\tilde{c}$. Sample-level score:

$$\text{ROUGE} - \text{Lsum}(\tilde{r}, \tilde{c}) = \frac{\text{LCS}(\tilde{r}, \tilde{c})}{|\tilde{r}|}$$

Corpus-level score:

$$\text{RLS}_{\text{corpus}} = \frac{1}{N} \sum_{k=1}^N \text{ROUGE} - \text{Lsum}(\tilde{r}^{(k)}, \tilde{c}^{(k)})$$

2) BLEU (B) measures n-gram precision with a brevity penalty (BP) to address over-short generated texts. We use BLEU-4 (max n – gram = 4, equal weights $w_n = 1/4$).

Modified n-gram precision:

$$\text{mp}_n(r, c) = \frac{\sum_{g \in \mathcal{G}_n(c)} \min(\text{cnt}_c(g), \text{cnt}_r(g))}{\sum_{g \in \mathcal{G}_n(c)} \text{cnt}_c(g)}$$

BLEU score:

$$\text{BLEU}(r, c) = \text{BP} \cdot \exp \left(\sum_{n=1}^4 w_n \log(\text{mp}_n(r, c)) \right)$$

Brevity penalty:

$$\text{BP} = \begin{cases} 1, & \text{if } |c| > |r|, \\ \exp \left(1 - \frac{|r|}{|c|} \right), & \text{if } |c| \leq |r|. \end{cases}$$

Corpus-level score (concatenate all texts first):

$$\text{B}_{\text{corpus}} = \text{BLEU}(\text{concat}(r^{(k)}), \text{concat}(c^{(k)}))$$

3) BERTScore series quantify semantic similarity using biomedical domain-specific embeddings (BioBERT-large-cased-v1 [48]). We use uniform token weights (IDF weights are ineffective for short domain texts):

Sample-level metrics (Precision: BP, Recall: BR, F1: BF):

$$\text{BP}(r, c) = \frac{\sum_{t=1}^{|c|} w_t \max_{1 \leq s \leq |r|} \cos(u_t, v_s)}{\sum_{t=1}^{|c|} 1}$$

$$\text{BR}(r, c) = \frac{\sum_{s=1}^{|r|} \tilde{w}_s \max_{1 \leq t \leq |c|} \cos(u_t, v_s)}{\sum_{s=1}^{|r|} 1}$$

$$BF(r, c) = \frac{2 \cdot BP(r, c) \cdot BR(r, c)}{BP(r, c) + BR(r, c)}$$

where u_t and v_s are token embeddings of c and r , respectively.

Corpus-level scores (average of sample-level scores):

$$BP_{corpus} = \frac{1}{N} \sum_{k=1}^N BP(r^{(k)}, c^{(k)})$$

$$BR_{corpus} = \frac{1}{N} \sum_{k=1}^N BR(r^{(k)}, c^{(k)})$$

$$BF_{corpus} = \frac{1}{N} \sum_{k=1}^N BF(r^{(k)}, c^{(k)})$$

4) METEOR (M) integrates exact/stem/synonym matching and a fragmentation penalty to balance lexical accuracy and coherence.

Unigram precision/recall:

$$P(r, c) = \frac{m}{|c|}, R(r, c) = \frac{m}{|r|}$$

where m is the number of matched unigrams.

Weighted harmonic mean:

$$F_{mean}(r, c) = \frac{10 P(r, c) R(r, c)}{R(r, c) + 9 P(r, c)}$$

Let ch be the number of contiguous matched chunks. The penalty term is:

$$Pen(r, c) = \gamma \left(\frac{ch}{m} \right)^\beta$$

Sample-level score:

$$METEOR(r, c) = (1 - Pen(r, c)) \cdot F_{mean}(r, c)$$

Corpus-level score:

$$M_{corpus} = \frac{1}{N} \sum_{k=1}^N METEOR(r^{(k)}, c^{(k)})$$

For the subcellular localization task, which is a single-label, 10-class classification problem where each protein is assigned to exactly one location, accuracy is a direct and comprehensive metric. The accuracy measures the exact match between the prediction and the single ground-truth label, effectively capturing the model's overall performance in this closed-set scenario.

In contrast, the tasks of GO term, UniProt keyword, and EC number prediction are inherently multi-label. A single protein can be associated with multiple correct functional annotations, and the label distribution is highly sparse. In such settings, accuracy becomes an inadequate and often misleading metric. Its binary, all-or-nothing nature fails to account for partial correctness and is heavily influenced by threshold choices and extreme class imbalance. Therefore, for these tasks, we adhere to the established standards in the field by employing precision, recall, and F1-score, that are specifically designed to robustly evaluate performance in multi-label, imbalanced classification problems.