ICML 2026 INSTRUCTION-FOLLOWING EMBEDDINGS

InstEmb: Instruction-Following Embeddings through Glimpses of the Future

Tianhao Gao*, Jun Fang*, Xiaohui Zhang, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Qixia Jiang

JD.com, Beijing, China · *Equal contribution

Core idea one prefilling pass, two semantics

Look-ahead tokens give embeddings a "glimpse" of output semantics — without decoding.

- **Problem:** last-token pooling misses output signals.
- **Method:** distill teacher outputs into look-ahead tokens.
- **Inference:** one prefilling pass.

Main results state-of-the-art instruction following

28.5
FollowIR score

+15.6
FollowIR p-MRR

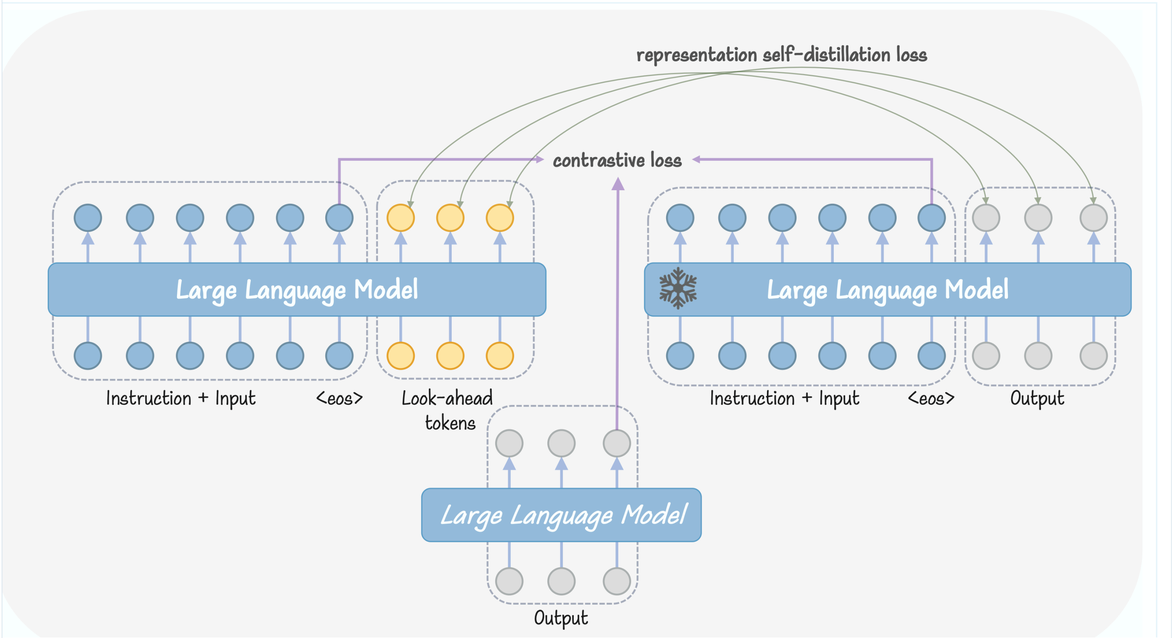
67.08
Instruction mean

63.39
Generic mean

Benchmark	Strong baseline	InstEmb	Gain
FollowIR score	Promptriever 26.1	**28.5**	+2.4
FollowIR p-MRR	FollowIR-7B +12.2	**+15.6**	+3.4
InfoSearch p-MRR	Llama-3 6.6	**7.7**	+1.1
Instruction mean	Inbedder 59.9	**67.08**	+7.18

No FollowIR-specific supervised training.

Method framework



Architecture. Student uses input + look-ahead tokens; frozen teacher uses input + target output.

Training objectives

OUTPUT-AWARE
Distill future states.

INPUT-INTRINSIC
Contrast views.

JOINT
Add both losses.

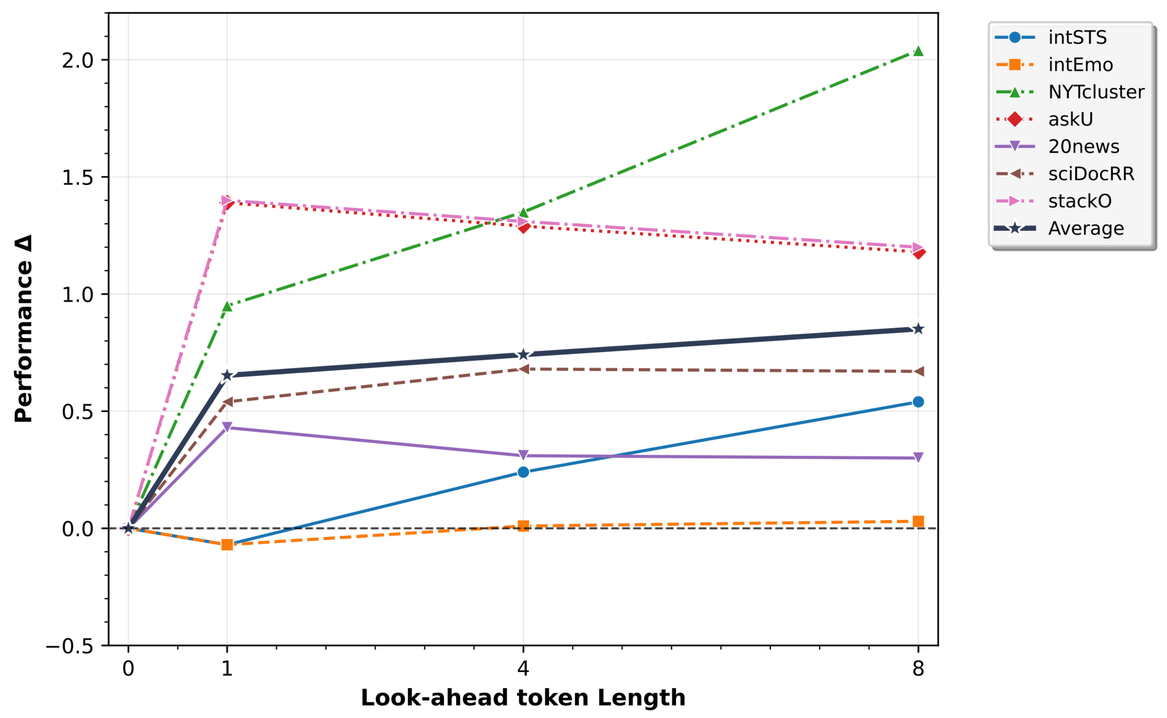
$$L_{Distill} = \frac{1}{L} \sum_{j=|x|+1}^{|x|+L} \| h^S_j - h^T_j \|_2^2$$

$$L_{CL} = -\frac{1}{4N} \sum_{i=1}^{N} \sum_{z \in V_i} \log \frac{\sum_{z' \in V_i \setminus \{z\}} \exp(\mathrm{sim}(z,z')/\tau)}{Z_i(z)}$$

$$V_i = \{s^1, s^2, t, a\}$$

$$L_{InstEmb} = L_{Distill} + L_{CL}$$

Look-ahead token length



Ablation. One look-ahead token already helps; longer look-ahead benefits richer output semantics.

Example why instruction matters

Same words, different intent:

Q1: "Is this tent durable for outdoor use?"

Q2: "Is this tent compact for outdoor use?"

- Baseline over-matches shared words.
- InstEmb separates intent.

Ablation highlights

DAAP
best pooling: 67.08 inst.

MSE
best instruction objective

KL
best generic transfer

Full
all contrastive views help

DAAP pooling

DAAP averages the two anchors optimized by InstEmb.

$$e = \tfrac{1}{2} \left(h^S_{|x|} + \mathrm{mean}(h^S_{\text{look-ahead}}) \right)$$

- No pooling search.
- Input + output anchors.

Instruction robustness

0.26
Correct → incorrect gap

0.26
Implicit → incorrect gap

Higher gap = stronger instruction sensitivity.

Takeaways

- **Future semantics** without decoding.
- **DAAP** aligns training and pooling.
- **SOTA** instruction-following embeddings.