

Text Capability Loss in Vision-Language Adaptation: An Attention-Sink Diagnosis

Minsik Choi^{1,4*}, Geewook Kim^{2,3*}, Young Geun Kim^{1†}

¹Korea University · ²NAVER Cloud AI · ³KAIST AI · ⁴FuriosaAI



Paper

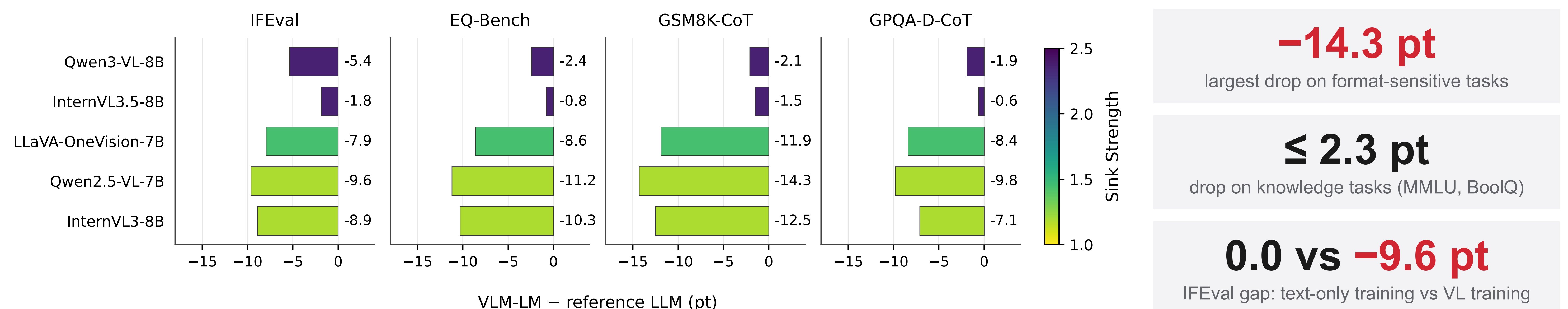


Code

Turning an LLM into a VLM degrades instruction following.

Sink Strength, one scalar on the base LLM, predicts the loss before any VL training.

1 VL adaptation erodes format-sensitive text capability



INSTRUCTION · IFEval

**Answer in ALL CAPS.
Use no commas.**

paraphrased from the IFEval prompt

BASE LLM · Qwen2.5-7B-Instruct

“... THE PRACTICE OF USING NATURAL
FEATURES TO CREATE A MORE SUSTAINABLE
ENVIRONMENT”

no commas

PASS

AFTER VL TRAINING · Qwen2.5-VL-7B

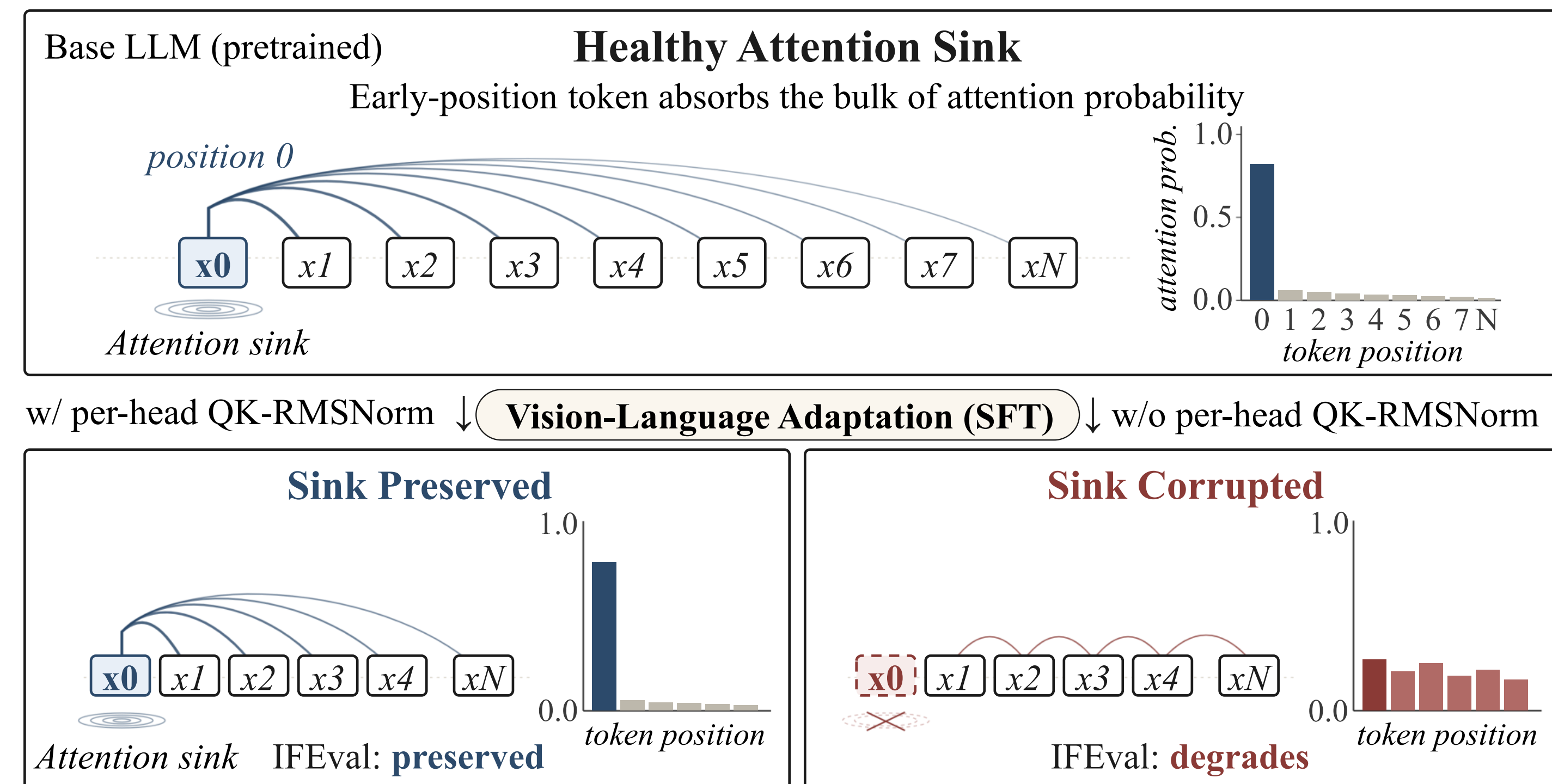
“... IT PROMOTES BIODIVERSITY,
SUSTAINABLE HABITATS, AND ECOLOGICAL
SERVICES ...”

uses commas

FAIL

2 Why: VL fine-tuning corrupts the attention sink

Attention Sink Corruption in Vision-Language Adaptation



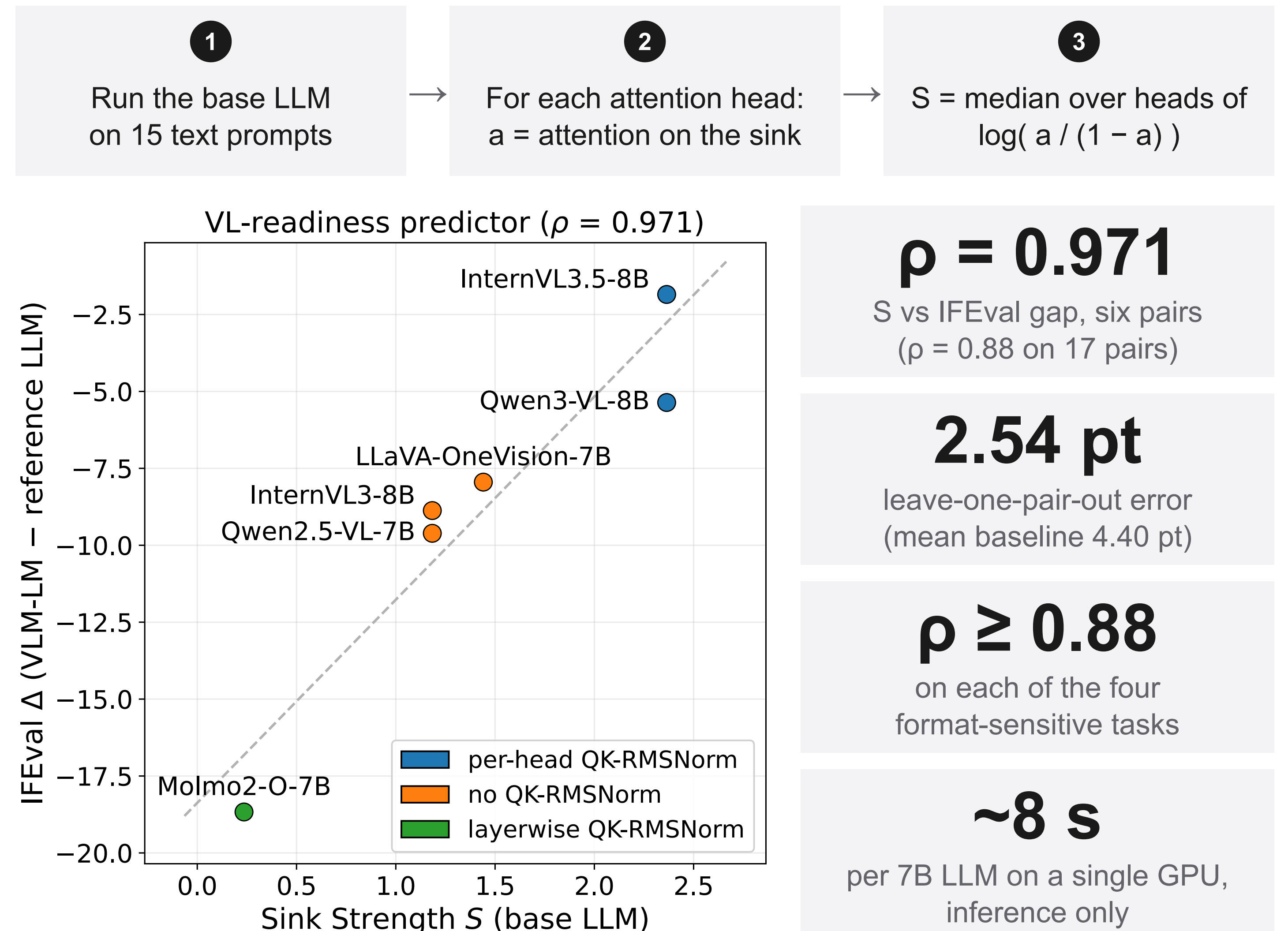
Sink. An early token that absorbs most attention.

Damage. VL updates shift attention scores, amplified at the sink.

Protection. Per-head QK-RMSNorm removes it (Theorem 1).

3 Sink Strength S: screen before VL training

HOW S IS MEASURED



4 Granularity matters: per-head vs layerwise

BASE-LLM DESIGN	SINK STRENGTH S	IFEVAL GAP
Per-head QK-RMSNorm Qwen3-VL, InternVL3.5	2.36	-1.8 to -5.4
No QK-RMSNorm Qwen2.5-VL, InternVL3, LLaVA-OV	1.18–1.44	-7.9 to -9.6
Layerwise QK-RMSNorm Molmo2-O (control)	0.24	-18.7

A “has QK-norm?” check passes Molmo2-O. S flags it as the highest risk.

5 Finding: simple fixes do not undo the damage

Adding QK-RMSNorm later
does not protect

11 vs 59.9

IFEval: injected model
(first checkpoint) vs base LLM

Merging back toward the LLM
does not recover it

0 of 7

merging settings
recover IFEval

**Takeaway: screen the base LLM with
Sink Strength before VL training.**