SpanUQ: Span-Level Uncertainty Quantification for Large Language Model Generation

Yimeng Zhang[1*] Yingying Zhuang[1] Ziyi Wang[2] Yuxuan Lu[2] Pei Chen[1]
Aman Gupta[1] Zhe Su[1] Ming Tan[1] Zhilin Zhang[1] Qun Liu[1]
Manikandarajan Ramanathan[1] Rajashekar Maragoud[1] Edward Vul[1] Jing Huang[1] Dakuo Wang[2]
[1]Amazon [2]Northeastern University [*]Corresponding author

NEURAL INFORMATION PROCESSING SYSTEMS
Atlanta · 2026

Paper | Code | Dataset | Project

01 · MOTIVATION

Why span-level?

Query: *"Who is Marie Curie?"*
Response: Marie Curie was a Polish physicist who won the Nobel Prize in 1901.

(a) **Token-Level Uncertainty:** hard to interpret

Marie | Curie | was | a | Polish | physicist | who | won | the | Nobel | Prize | in | 19 | 01 | ·

(b) **Sequence-Level Uncertainty:** cannot localize *which part* is unreliable

Marie Curie was a Polish physicist who won the Nobel Prize in 1901. $u = 0.52$

(c) **Span-Level Uncertainty (Ours):** interpretable and localizable

Marie Curie was a Polish physicist (u=0.08) who won the Nobel Prize (u=0.12) in 1901 (u=0.81) ·

$u \in [0, 1]$: uncertainty score 0 1

Tokens are **noisy**; one sequence score (**u = 0.52**) **blends** right and wrong claims. Spans **pinpoint** the error: the Physics Nobel was 1903, not **1901**.

02 · NEW TASK

SLUE

Span-Level Uncertainty Estimation. From a **single forward pass**, jointly **detect** spans $s_k = (b_k, e_k)$, each conveying one assessable unit of meaning, and **score** each with uncertainty $u_k \in [0, 1]$ (spans ≈ claims / atomic facts, and they may overlap).

Why it matters: span scores underpin self-refinement and act as a decision signal downstream, from verifiers to collaborating agents.

0.939 span AUROC

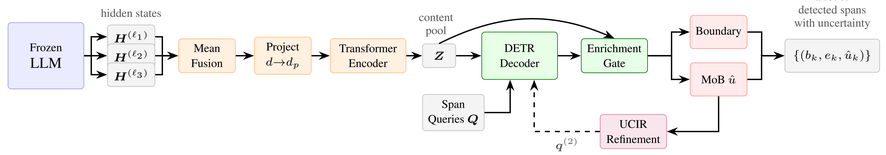
0.910 span detection F1

10–20× faster than sampling

25M probe params (<0.2%)

03 · METHOD

Distill 20 samples into one forward pass



A lightweight probe on a **frozen LLM**: only ~25M params are trained; the probe adds <50 ms.

Multi-Layer Fusion
Average hidden states from peak-zone layers (22/24/26 of Qwen3-14B), where the uncertainty signal is strongest.

DETR Span Decoder
32 learnable span queries + Hungarian matching; regress (b, e) with L1+GIoU. Handles **overlapping** spans.

Mixture of Beta ($K = 3$)
Fits the bimodal labels (certain ≈0, hallucinated ≈1) plus a middle mode. Nearly self-calibrated: **ECE 0.020**.

UCIR Refinement
Feed round-1 $\hat{u}$ back through the shared decoder for a "second look": **+3.0%** ρ_{span}, <15% overhead.

04 · BENCHMARK

SpanUQ-Bench: the first span-level UQ benchmark

1 Sample 20 responses at $T = 1$ → **2 Decompose** atomic claims → spans → **3 Verify** each claim vs. Wikipedia → **4 Soft label** $u_k^* = 1 - \frac{1}{S}\sum_s \mathbb{1}[\text{supp.}]$

20K prompts · **293K** spans · **5** domains (NQ, TriviaQA, ELI5, Bio, FELM) · **5** LLMs

Label distribution

62% $u^* < 0.1$ **20%** mid **18%** $u^* > 0.9$

05 · RESULTS

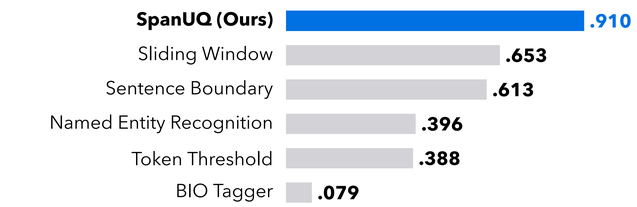
Best on all 5 LLMs

AUROC↑	14B	8B	4B	30B	M-7B
Token Entropy	.603	.580	.575	.592	.569
P(True)	.526	.546	.529	.541	.505
Verbalized	.593	.574	.595	.598	.498
FActScore	.625	.762	.714	.701	.661
SelfCheck-NLI	.808	.801	.782	.813	.816
MLP Probe†	.881	.884	.873	.893	.863
SpanUQ	**.939**	**.930**	**.944**	**.936**	**.908**

Qwen3-14B/8B/4B/30B-A3B and Mistral-7B. †Oracle span boundaries. SpanUQ also has the lowest MAE (0.110–0.129).

06 · DETECTION

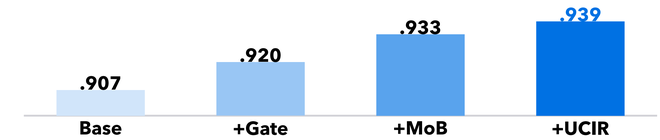
Learned spans beat heuristics



Span detection F1: **+39.4%** over the best heuristic. BIO tagging cannot represent overlapping claims.

07 · ABLATION

Every component helps



AUROC on Qwen3-14B. ECE: 0.058 → 0.052 → **0.020** → **0.013**.

Takeaways

Spans are the natural unit for LLM uncertainty: interpretable & localizable.

One pass, 25M params beats 20-sample methods on 5 LLMs, 10–20× faster.

Next: regenerate only uncertain spans, route trust between collaborating agents, and use span scores as dense RL rewards.