

FAST-Eval and FAST-LoRA: Uncertainty Evaluation Framework and Low-Rank Adaptation for Finding Attributes with Supporting Text

Samuel Forster and Inmaculada Arnedillo-Sánchez

School of Computer Science and Statistics

Trinity College Dublin, Dublin, Ireland

{sforster, macu.arnedillo}@tcd.ie

Abstract

Many domains require reasoning over unstructured text to identify attributes of interest that are not stated in a standardised form. Useful judgements on the presence of these attributes should also express calibrated confidence statements and cite supporting text as evidence. We define this task as **Finding Attributes with Supporting Text (FAST)**. We introduce **FAST-Eval**, an uncertainty evaluation framework for confidence calibration, attribute identification and evidential grounding in FAST, together with a dataset of 1,000 human-annotated (*document, attribute*) pairs validated against an independent recruitment expert (Spearman $\rho_s = 0.853$, binary $\kappa = 0.715$). Taking employability skills in CVs as a case study, we use FAST-Eval to test whether **FAST-LoRA**, a Low-Rank Adaptation (LoRA) adapter applied to Qwen3-8B, improves performance in these three dimensions. Trained on 500 pairs, FAST-LoRA outperforms its untuned Qwen3-8B counterpart and GPT-5 on unseen skill identification (F1 0.877 vs 0.736 and 0.766) and calibration, while matching GPT-5 on citation fidelity. Both untuned models assign the highest confidence level to about half their judgements; FAST-LoRA instead distributes confidence across the scale and makes fewer unsupported positive predictions. These results suggest that modest domain adaptation is a more effective strategy than scaling model parameters for FAST.

1 Introduction

In decision-support contexts, identifying attributes of interest is inherently uncertain because source documents rarely express relevant attributes as explicit, standardised statements. A system must therefore determine whether incomplete, indirect or ambiguous evidence supports a target attribute, express its confidence in that judgement and identify the supporting text. Recruitment provides an instructive example: deciding whether a CV demon-

strates a target skill requires more than matching headings or keywords. The same detect-then-verify structure recurs in domains such as legal review, medical assessment and compliance screening. Building on benchmark data and experimental results originally produced as part of an MSc dissertation (Forster, 2026), this paper formalises the task of **Finding Attributes with Supporting Text (FAST)** and introduces **FAST-Eval**, an uncertainty evaluation framework for jointly assessing confidence calibration, attribute identification and evidence attribution.

Large language models are capable generative information extractors (Xu et al., 2024), but three properties obstruct their use where errors are costly. Their confidence estimates can be poorly calibrated (Geng et al., 2024); they generate content unsupported by the source (Ji et al., 2023; Schmidtova et al., 2025); and attaching citations does not guarantee complete or faithful support (Gao et al., 2023; Zhang et al., 2024; Rashkin et al., 2023). AI systems used to analyse job applications or evaluate candidates are listed as high-risk under the EU AI Act, whose human-oversight provisions require deployers to be able to interpret and override outputs (European Parliament and Council, 2024). The risk is concrete: Amazon abandoned an experimental recruiting tool after it learned to penalise CVs containing terms associated with women (Dastin, 2018). This motivates exposing confidence and evidence for inspection at the level of each decision.

We therefore ask whether a model’s stated confidence accurately reflects the reliability of each judgement, allowing uncertain cases to be referred for human review. Parameter-efficient fine-tuning (PEFT), and Low-Rank Adaptation (LoRA) in particular (Hu et al., 2021), offers a low-cost way to teach this behaviour. We test whether it improves calibration without sacrificing identification or grounded evidence.

Contributions. We introduce the task of **Finding Attributes with Supporting Text (FAST)**: given a document and a target attribute, decide whether it is present, how confidently, and with what supporting quotation. For FAST we contribute:

1. **FAST-Eval** (§4), an uncertainty-aware evaluation framework that scores communicated ordinal confidence, attribute identification and evidence attribution together, instantiated on 1,000 human-annotated (*document*, *skill*) pairs with independently validated confidence and evidence-span labels.
2. **FAST-LoRA** (§5.2), a parameter-efficient adapter, built on LoRA (Hu et al., 2021), that trains an open-source model to emit task-facing ordinal confidence and grounded, verbatim citations for a target attribute.
3. An **uncertainty-focused analysis** of the resulting pair-level outputs, testing calibration-target sensitivity and whether the ordinal scale supports selective prediction and deferral.

FAST-Eval is designed as a reusable, domain-independent framework, instantiated here on employability data to benchmark FAST-LoRA. We ask whether adaptation improves (RQ1) confidence calibration, (RQ2) attribute identification and (RQ3) evidence attribution. Under the primary mapping and mapping-independent ordinal metrics, FAST-LoRA is better calibrated than both untuned models; it also improves identification F1 and matches GPT-5 on citation fidelity (§6).

2 Related Work

Skill and attribute extraction. Skill extraction from job-market text is commonly framed as supervised sequence labelling, with SkillSpan (Zhang et al., 2022) providing span-level annotations and BERT baselines. Such systems detect that a skill term appears rather than verify whether its surrounding context actually supports its application. FAST targets verification: simple vocabulary overlap alone is insufficient, and each judgement must include confidence and supporting evidence.

Claim and evidence verification. Closely related tasks pair a label for a targeted proposition with supporting evidence, following the rationalised-prediction paradigm of ERASER (DeYoung et al., 2020). Evidence Inference predicts a categorical treatment-effect relation from a clinical-trial report and structured query, while identifying supporting evidence (Lehman et al.,

2019). FEVER and SciFact instead verify free-form claims using retrieved evidence (Thorne et al., 2018; Wadden et al., 2020). FAST evaluates a pre-specified attribute, returns calibrated ordinal confidence rather than a categorical label, and scores identification, calibration and grounding jointly. An ERASER-style model could be adapted by treating each document–attribute pair as a classification instance and its extracted rationale as the citation. This would provide identification and span-selection baselines, but matching FAST would require a new ordinal output space, confidence supervision and source-fidelity checks. ERASER is therefore a useful design basis rather than a directly applicable off-the-shelf baseline.

Attributed generation. Attributed question answering and citation generation make model outputs verifiable by testing whether generated statements are supported by cited sources (Rashkin et al., 2023; Gao et al., 2023; Zhang et al., 2024). FAST differs in assessing a pre-specified attribute rather than a free-form answer, pairing every judgement with ordinal confidence and a citation, and evaluating identification, calibration and grounding together rather than attribution alone.

Expressed uncertainty and calibration. Linguistic calibration asks whether expressed doubt or confidence tracks correctness (Mielke et al., 2022). Verbalised confidence can be better calibrated than conditional token probabilities in question-answering tasks (Tian et al., 2023), yet need not reflect intrinsic model uncertainty (Yona et al., 2024). Related approaches fine-tune models to output calibrated probabilities (Kapoor et al., 2024), apply post-hoc calibration to entity matching (Kamsteeg et al., 2025), or derive confidence from internal activations for RAG abstention (Huang et al., 2025). FAST instead supervises decision-facing ordinal confidence, anchors it to a targeted attribute judgement, and evaluates it jointly with human confidence labels and source evidence.

Domain adaptation with PEFT. LoRA injects trainable low-rank updates into frozen model weights at a fraction of full fine-tuning’s cost (Hu et al., 2021). Prior work uses it to customise generation style and adapt dialogue models (Liu et al., 2024; Alghisi et al., 2024). We extend the evaluation beyond style or task performance to confidence calibration and evidential grounding in specific domains.

Hallucination and evaluation. Hallucination constitutes fluent output that is unfaithful to or unverifiable from its source (Ji et al., 2023). Human evaluations vary and frequently omit guidelines or inter-annotator agreement (Schmidtova et al., 2025); annotation format can also affect hallucination detection and agreement (Liu et al., 2025). Citation support is itself difficult to assess reliably (Zhang et al., 2024). Following recommendations for reproducible human evaluation (Thomson and Belz, 2024), FAST-Eval pre-specifies its annotation and exclusion criteria, reports inter-annotator agreement, applies deterministic shared scoring, and keeps confidence, identification and evidence separate rather than collapsing them into one score.

3 Task Formulation and Uncertainty Representation

We define the task as **Finding Attributes with Supporting Text (FAST)**. Its generic unit of analysis is a (*document, attribute*) pair, instantiated in this study as a (*CV document, skill*) pair (d, s). Given a CV and a recruiter-specified target skill, the model outputs (i) an ordinal confidence score $\hat{c}_{d,s} \in \{1, \dots, 5\}$ that the document evidences the skill, and (ii) a verbatim citation extract $\hat{e}_{d,s}$ from the document supporting that judgement. For confidence levels 2–5, the extract is the strongest available evidence for the judgement, even when that evidence is weak or inconclusive; level 1 explicitly denotes that no supporting evidence exists and therefore requires an empty extract. The task is targeted: the model assesses an explicit attribute set rather than discovering attributes open-endedly, which bounds output entropy and enables like-for-like cross-model comparison. The ground truth for each pair is $(c_{d,s}^*, e_{d,s}^*)$.

The confidence score states how certain the model is that the attribute is present or absent. FAST-Eval tests whether these reported levels match how often the decisions are correct and how confident human annotators are.

Confidence scale. Confidence uses a five-point ordinal scale with fixed semantic anchors, applied identically to model prompts and human annotation (Table 1). Binary present/absent labels are derived by thresholding at $\hat{c} \geq 4$. A continuous 0–100 scale was trialled and rejected: both models and annotators clustered on round decades, giving the appearance but not the substance of fine resolution. Five anchored levels are genuinely discriminative

at every step.

L	Anchor	Binary
1	Definitely not present	Absent
2	Probably not present	Absent
3	Unclear / mixed	Absent
4	Probably present	Present
5	Definitely present	Present

Table 1: Ordinal confidence scale.

Citation constraint. At levels 2–5, the extract must be a verbatim quote giving the best available textual basis for the score, whether it supports presence, absence or uncertainty. This allows it to be (a) checked against the source for fidelity and (b) compared to the annotated span for relevance. At level 1, where the judgement is that no evidence exists, the extract must be empty. This mirrors a hypothetical real-world deployment: a reviewer should be able to locate the basis of any non-trivial judgement in the original document.

Input	
Skills: ["Project Management", "SQL"]	
Document: "Senior Graphic Designer... Successfully orchestrated complex projects from the idea stage through design, mock-up, and final rendering..."	
Output	
[{"skill_id": "Project Management", "confidence_score": 5, "reference_extract": "Successfully orchestrated complex projects. . ."}, {"skill_id": "SQL", "confidence_score": 1, "reference_extract": ""}]	

Figure 1: Sample input/output.

Figure 1 shows a representative instance: the model returns confidence 5 with a grounded citation for a demonstrated skill, and confidence 1 with an empty citation for a skill with no support in the document.

4 FAST Evaluation Framework

The **FAST Evaluation Framework (FAST-Eval)** is a reproducible framework for evaluating any system that returns the FAST output schema: an ordinal confidence judgement and a supporting citation for each document–attribute pair. It jointly scores three dimensions:

- **Confidence calibration:** whether its ordinal confidence tracks human judgement and the expected correctness of each confidence level;
- **Attribute identification:** whether the system correctly judges the target attribute as present or absent; and

- **Evidence attribution:** whether its citation identifies relevant evidence and is faithfully grounded in the source.

These dimensions are evaluated against independently validated human labels, with deterministic output normalisation, explicit reporting of unparseable outputs, matched seen and held-out partitions, and paired uncertainty and significance analysis. The following subsections specify the benchmark, validation, metrics and procedure so that the same framework can be applied to FAST-LoRA, an untuned model, or another system that emits the required schema. The output schema and scoring procedure are reusable across domains: for example, a contract and a target legal clause could replace a CV and skill without changing what the system outputs or how it is scored. Each application nevertheless requires a domain-specific ground truth comprising new target attributes, documents and expert-validated annotations. The empirical results in this paper therefore establish performance only for English CV skill verification, with merit in investigating other domains in future work.

4.1 Dataset and Ground Truth

Documents are sampled without replacement from Opensporks (Opensporks, 2024), a public corpus of approximately 2,480 English CVs across 24 professional categories, using its plain-text field. Target attributes are sampled from the ESCO taxonomy (European Commission, 2017), drawing from its knowledge and skill concepts as well as transversal skills. The benchmark includes both attributes that tend to be stated explicitly and those typically inferred from described experience, testing lexical and inferential verification behaviour. Skills absent from a document are included at equal frequency to present ones, providing the negatives required to measure precision.

The benchmark comprises 500 training pairs and 500 held-out pairs, stratified across four buckets of one to four skills per document, so that performance can be measured against query complexity rather than in aggregate alone. All 1,000 pairs were manually annotated by the first author, a native English speaker and MSc Computer Science student who designed the task and annotation guidelines; no crowd workers were used. Annotation required approximately 80 hours over six weeks. For each pair, the annotator viewed the plain-text CV and target skill, assigned one of the five anchored con-

fidence levels in Table 1, and selected the minimal verbatim passage that best justified the judgement. Levels 2–5 required a citation, while level 1 required an empty extract. The protocol prioritised explicit evidence, allowed strong indirect evidence when clearly relevant, and resolved remaining ambiguous cases conservatively as absent. Where several passages were relevant, the strongest representative span was retained. Binary labels were derived only afterwards by thresholding the ordinal annotation at level 4.

4.2 Expert Validation of Labels

Inter-annotator agreement is frequently under-reported in NLG human evaluation (Schmidtova et al., 2025). To validate the primary annotations, an independent academic who is a native English speaker and has professional hiring experience re-annotated 100 pairs from the training partition. The expert received the same written definitions, confidence anchors, citation requirements and conservative tie-breaking rule used for the full benchmark, followed by a brief orientation to the response form. No labelled examples were shared, and the first author’s labels remained withheld throughout. The expert then annotated each pair independently using a dedicated questionnaire containing the target skill, CV text and separate fields for confidence and citation. The expert labels served only to validate the reliability of the primary annotations; they did not replace or adjudicate the benchmark labels.

Agreement was strong on all three dimensions (Table 2): ordinal Spearman $\rho_s = 0.853$ and linear-weighted $\kappa_w = 0.664$ (substantial); binary $\kappa = 0.715$ (substantial); and a two-step citation hit rate of 83.6%. All four pre-specified acceptance thresholds ($\rho_s \geq 0.70$, $\kappa_w \geq 0.60$, binary $\kappa \geq 0.60$, citation hit $\geq 70\%$) were met, so no remediation was required.

Metric	Value	Interp.
Spearman ρ_s	0.853	strong ($p < .001$)
Weighted κ (linear)	0.664	substantial
Weighted κ (quadratic)	0.825	near-perfect
Exact / within-one	51% / 92%	—
Binary agreement	86%	—
Binary Cohen’s κ	0.715	substantial
Citation hit rate	83.6%	61/73 pairs
Mean Jaccard	0.608	—

Table 2: Inter-annotator agreement.

Disagreements concentrated at intermediate confidence levels and on transversal or implicit-technical skills, as anticipated: citation agreement rose monotonically with confidence, from 60% at level 2 to 95% at level 5, consistent with ambiguous borderline evidence rather than a systematic difference in annotation strategy. The study serves two purposes. First, it acts as a pre-specified quality gate for the labels, reducing the risk that downstream differences are driven by annotation noise. Second, it provides a human reference point: the 83.6% citation agreement demonstrates the ambiguity inherent in evaluating span selection against a single annotated citation (§6). One directional bias is recorded: our positive rate (38%) is lower than the expert’s (46%), so the primary labels may understate precision.

4.3 Evaluation Metrics

All metrics use a shared, deterministic evaluation codebase and are applied identically to every condition. They operationalise FAST-Eval’s three evaluation dimensions as follows.

RQ1 — Confidence calibration. We map each level to an implied presence probability $p(l) = (l - 0.5)/5$, giving $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. For predictions assigned level l , the level-wise binary accuracy $\text{acc}(l)$ is the proportion whose thresholded binary decision agrees with ground truth: absence at levels 1–3 and presence at levels 4–5. Its target is the confidence assigned to that decision:

$$q(l) = \begin{cases} 1 - p(l), & l \in \{1, 2, 3\}, \\ p(l), & l \in \{4, 5\}. \end{cases}$$

This gives correctness targets $\{0.9, 0.7, 0.5, 0.7, 0.9\}$. We then compute an Ordinal Calibration Error, weighting each level by its prediction count n_l out of N total predictions:

$$\text{OCE} = \sum_{l=1}^5 \frac{n_l}{N} |\text{acc}(l) - q(l)|,$$

an ordinal adaptation of the standard Expected Calibration Error (Geng et al., 2024). We also report Spearman ρ_s between predicted and ground-truth confidence (mapping-independent), exact-match rate, within-one rate and mean absolute error (MAE).

Two supplementary analyses test how useful and robust the communicated confidence is. For selective prediction, we report binary accuracy and

coverage when retaining (i) every level, (ii) all except the maximally uncertain level 3, and (iii) only the most decisive levels 1 and 5. For mapping sensitivity, we recompute OCE with compressed correctness targets $\{0.80, 0.65, 0.50, 0.65, 0.80\}$ and expanded targets $\{0.95, 0.75, 0.50, 0.75, 0.95\}$ around the primary mapping.

RQ2 — Attribute identification. From thresholded binary labels ($\hat{y} = \mathbb{I}[\hat{c} \geq 4]$) we report F1 (primary), precision and recall, plus accuracy. Precision and recall are reported separately so any adaptation-induced shift is visible.

RQ3 — Evidence attribution. *Hit rate:* for eligible pairs (ground-truth positive with a non-empty annotated span), a two-step check is performed: whether the predicted and annotated quotes overlap in position in the document or, failing that, share at least half their words (Jaccard similarity ≥ 0.5). This lenient criterion credits a predicted citation that overlaps the annotated span without requiring an exact match, since a document often contains several acceptable supporting spans. *Hallucination rate:* the complement of the fraction of non-empty predicted citations found verbatim in the source, computed over predictions at every confidence level and polarity. Levels 2–5 require a best-evidence citation; level 1 correctly uses an empty extract and therefore contributes no citation to check. The two metrics are complementary: hit rate measures evidential relevance, hallucination measures raw source fidelity.

4.4 Evaluation Procedure

Each condition is run separately on the seen and unseen partitions, using a single prompt template (containing role framing, skill list, document contents, per-level anchor definitions, and a JSON-only constraint) applied identically for training and for every evaluation condition, so input framing cannot differ between them. Per case, the template is instantiated, inference is run at temperature 0.1, and the output is parsed deterministically into the schema. Each pair is scored against ground truth, and each non-empty citation is checked against the source by a case-insensitive substring match, which underpins the hallucination rate. Syntactically malformed but recoverable outputs are normalised and retained. Only outputs that cannot be mapped to the schema even after deterministic normalisation are classified as unparseable and excluded; these failures are reported separately rather than silently

omitted. Raw outputs, per-pair CSVs and aggregate metrics are archived per run for reproducibility.

5 FAST-LoRA and Experimental Design

5.1 Model Conditions

We use a controlled comparison across three model configurations: (i) **GPT-5**, a frontier closed-source reference; (ii) **Qwen3-8B** (Yang et al., 2025), an open-source base model; and (iii) the same model adapted with **FAST-LoRA**, our adapter (§5.2). Because (ii) and (iii) share the same architecture and tokeniser, their difference isolates the effect of adaptation; the GPT-5 comparison contextualises performance against a much larger frontier model. All conditions receive identical documents, skill lists, prompt template, output schema, parsing and scoring. Dependent variables map onto the three FAST-Eval dimensions: RQ1 confidence calibration, RQ2 attribute identification and RQ3 evidence attribution.

5.2 FAST-LoRA Training

Parameter	Value
Base model	Qwen3-8B
Rank r / scaling α	16 / 32
Dropout	0.05
Target modules	all attn + MLP proj.
Epochs	2
Effective batch size	16 (1×16 accum.)
Learning rate / schedule	2×10^{-4} / cosine
Max sequence length	512
Trainable params	$\approx 42\text{M}$ of 8B

Table 3: LoRA configuration.

FAST-LoRA applies LoRA (Hu et al., 2021) to Qwen3-8B with the Hugging Face PEFT stack (Table 3), targeting all attention and MLP projections ($\sim 42\text{M}$ trainable parameters, $\sim 0.5\%$ of the model) while leaving embeddings and layer norms frozen. It loads over the unmodified base weights, so the conditions differ only through the added low-rank modules.

6 Results

Table 4 reports all metrics on both partitions. Outputs that remain unparseable after deterministic normalisation are excluded from aggregation and reported separately.

6.1 Output Reliability

Seen/unseen evaluable outputs (of 500) were 484/494 (base), 493/490 (FAST-LoRA) and

500/496 (GPT-5); all others remained unmappable even after deterministic normalisation.

6.2 RQ1: Confidence Calibration

OCE shows a clear calibration advantage for FAST-LoRA on unseen data (0.147 vs 0.246 for the base model and 0.210 for GPT-5). The reliability diagrams in Figure 2 visualise the corresponding per-level agreement accuracy and the number of predictions assigned each level; Table 5 gives the underlying values.

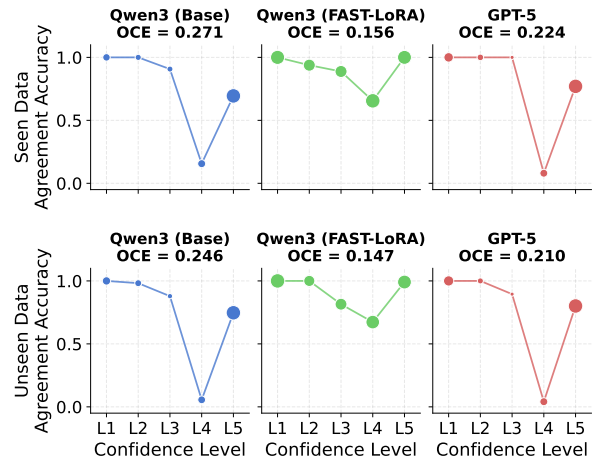


Figure 2: Per-level agreement accuracy. The untuned models fall sharply at level 4; FAST-LoRA approaches the positive-side targets at levels 4 and 5. Circle area is proportional to the number of predictions at each level.

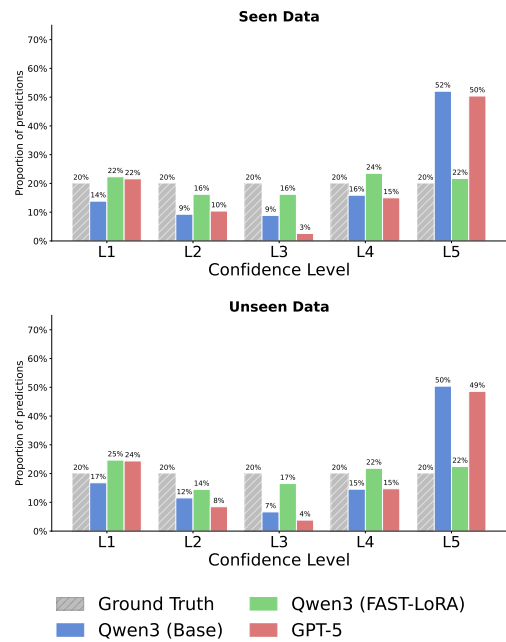


Figure 3: Predicted confidence distributions on seen (top) and unseen (bottom) data.

Metric	Seen (500 pairs)			Unseen (500 pairs)		
	Qwen3-8B (Base)	Qwen3-8B (FAST-LoRA)	GPT-5	Qwen3-8B (Base)	Qwen3-8B (FAST-LoRA)	GPT-5
<i>RQ1 — Confidence calibration</i>						
OCE ↓	0.271	0.156	0.224	0.246	0.147	0.210
Spearman ρ_s	0.755	0.882	0.849	0.813	0.898	0.860
Exact match	0.360	0.633	0.472	0.399	0.620	0.460
Within-one	0.742	0.959	0.816	0.794	0.978	0.845
MAE ↓	0.969	0.410	0.746	0.854	0.402	0.722
<i>RQ2 — Attribute identification</i>						
Precision	0.568	0.821	0.612	0.592	0.834	0.624
Recall	0.979	0.929	1.000	0.974	0.923	0.990
F1	0.719	0.871	0.759	0.736	0.877	0.766
Accuracy	0.698	0.890	0.746	0.725	0.896	0.758
<i>RQ3 — Evidence attribution</i>						
Hit rate	0.435	0.452	0.450	0.441	0.475	0.439
Halluc. ↓	0.253	0.146	0.178	0.232	0.139	0.141

Table 4: Full results on both partitions; best value per row in bold, and ↓ indicates that lower is better. Confidence intervals and significance tests are reported in §6.6.

Both untuned models place about half their predictions at level 5 (base 50.4%, GPT-5 48.6%) against a uniform 20% ground-truth distribution, leaving the intermediate levels underused. FAST-LoRA instead distributes its predictions more evenly across the scale, assigning 14.5–24.7% to each level (Figure 3).

At level 4 (“probably present”) the base model is correct 5.6% of the time and GPT-5 4.1%, whereas FAST-LoRA reaches 67.3%, close to the 70% target. At level 5 FAST-LoRA is correct 99.1% of the time, against 74.7% (base) and 80.1% (GPT-5). A Spearman correlation of $\rho_s = 0.898$ (vs 0.813 and 0.860) and a within-one rate of 0.978 indicate that its confidence ranking also aligns with human judgement.

Level	Base	FAST-LoRA	GPT-5	Target
1	1.000	1.000	1.000	0.90
2	0.982	1.000	1.000	0.70
3	0.879	0.815	0.895	0.50
4	0.056	0.673	0.041	0.70
5	0.747	0.991	0.801	0.90

Table 5: Per-level agreement accuracy on unseen data.

OCE’s exact ordering is mapping-sensitive. Under compressed correctness targets, GPT-5 has lower OCE than FAST-LoRA (0.184 vs 0.200; Qwen3 base 0.211); FAST-LoRA leads under the primary and expanded targets (0.127 vs 0.264 and 0.225). We therefore interpret OCE alongside the mapping-independent rank correlation and ordinal-agreement metrics, on which FAST-LoRA leads

consistently.

The ordinal scores also support selective prediction (Table 6). Deferring only level 3 raises FAST-LoRA’s accuracy from 89.6% to 91.2% at 83.5% coverage; it does not help the untuned models because many of their errors occur at level 4. Retaining only levels 1 and 5 yields 99.6% accuracy for FAST-LoRA at 47.1% coverage, compared with 81.0% for the base model and 86.7% for GPT-5 at their respective coverage levels.

Model	Accuracy / coverage (%)		
	All	Exclude L3	L1 and L5
Base	72.5 / 100	71.4 / 93.3	81.0 / 67.2
FAST-LoRA	89.6 / 100	91.2 / 83.5	99.6 / 47.1
GPT-5	75.8 / 100	75.3 / 96.2	86.7 / 73.0

Table 6: Selective prediction on unseen data. Coverage is the proportion retained after deferring the omitted confidence levels.

6.3 RQ2: Attribute Identification

On unseen data FAST-LoRA reaches F1 0.877, versus 0.736 (base, +14.1pp) and 0.766 (GPT-5, +11.1pp). The gain is primarily precision: 0.834 versus 0.592, while recall falls only from 0.974 to 0.923. Both untuned models show a *default-to-presence* pattern, combining near-perfect recall with precision near 0.6. FAST-LoRA also leads every one-to-four-skill query bucket on both partitions; its largest unseen margin over the base model is +20.8pp in the single-skill bucket.

6.4 RQ3: Evidence Attribution

Hit rates are similar across conditions (FAST-LoRA 0.475 vs 0.441 and 0.439 unseen). This metric is constrained by annotation ambiguity: the expert matched the original citation on 83.6% of eligible pairs under the same procedure, and documents may contain multiple valid spans. FAST-LoRA reduces hallucination from 23.2% (base) to 13.9%, statistically matching GPT-5’s 14.1% verbatim fidelity.

6.5 Error Analysis

Four patterns recur:

- **Adjacent-vocabulary false positives** dominate untuned errors: a bare “project management” heading receives level 5 from the base model but level 2 from FAST-LoRA.
- **Indirect-evidence false negatives** are FAST-LoRA’s main residual error (15 unseen vs 5 for base), mostly on implicitly evidenced transversal skills where FAST-LoRA is harsher than the ground truth.
- **High-confidence errors** affect $\sim 25\%$ of the base model’s 249 level-5 predictions and $\sim 20\%$ of GPT-5’s, versus $< 1\%$ of FAST-LoRA’s.
- **Confident negatives** at levels 1–2 are essentially always correct, locating the untuned models’ errors in over-prediction of presence.

FAST-LoRA thus trades a small recall reduction (-0.051) for far fewer false positives.

6.6 Statistical Significance

Sampling uncertainty in the reported estimates is assessed by bootstrap resampling of the evaluated pairs (5,000 replicates), while differences in paired binary correctness are tested with McNemar’s test. On unseen data, FAST-LoRA’s F1 of 0.877 has a 95% bootstrap CI of $[0.841, 0.908]$, separated from the base model’s 0.736 $[0.692, 0.777]$ and GPT-5’s 0.766 $[0.723, 0.805]$. Separately, McNemar’s test confirms that FAST-LoRA makes significantly more correct binary decisions than either comparator ($p < 10^{-11}$); this test concerns paired correctness rather than F1 itself. Paired bootstrap differences in OCE on commonly evaluated unseen pairs favour FAST-LoRA over both the base model (-0.096 , 95% CI $[-0.115, -0.067]$) and GPT-5 (-0.063 , $[-0.082, -0.034]$) under the primary mapping. The hallucination reduction over the base model is significant (paired difference 95% CI $[-0.117, -0.023]$), whereas FAST-LoRA and

GPT-5 are statistically indistinguishable on hallucination (difference CI $[-0.022, +0.070]$) and on hit rate (overlapping CIs), consistent with the span-selection ambiguity discussed above.

7 Discussion

One pattern, three measurements. The three improvements are consistent with one coherent behavioural change. Untuned models often predict presence based on vocabulary proximity alone, whereas FAST-LoRA requires substantive evidence of skill *application* before committing to a positive prediction. Better-filtered positive predictions drive a rise in precision, while a more accurate, even use of the confidence distribution improves calibration. Closer use of the source text lowers hallucination. Because this pattern holds on held-out data, FAST-LoRA appears to have learned a general threshold on required evidence strength rather than just memorising the training pairs.

Beyond distribution matching. The ground-truth confidence distribution is uniform by construction, and FAST-LoRA is trained on it, so its aggregate distribution alone is not evidence of a useful representation of uncertainty in general. However, two results argue against the behaviour being explained by simple frequency matching. First, FAST-LoRA has the strongest case-level rank correlation with human confidence ($\rho_s = 0.898$), which cannot be guaranteed by simply matching the marginal distribution. Second, it has better per-level correctness and 99.6% accuracy at the decisive levels 1 and 5. Further, a model can place 20% of its predictions at, for instance, level 4, while being either 4% or 67% correct for that level; this conditional accuracy is what separates FAST-LoRA from the untuned models.

Specialisation over scale. FAST-LoRA leads GPT-5 under the primary OCE mapping, every mapping-independent ordinal metric and identification F1, with compressed OCE targets the lone exception. The substantive finding is the confidence gap, not merely that a fine-tuned model beats a zero-shot one. GPT-5 assigns its top confidence level to nearly half of its predictions, yet is correct only 4.1% of the time at *probably present*. The shared zero-shot prompt defines the scale but does not convey the task’s evidential threshold, whereas fine-tuning learns that threshold from hundreds of labelled examples without adding inference-time

context. While including an equivalent number of behavioural examples at inference time would be impractical, whether few-shot prompting could achieve similar performance merits investigation. Adaptation is therefore a sufficient and practical route to more reliable performance in FAST, though not necessarily the only one.

Future directions. Four directions follow. First, compare few-shot threshold demonstrations directly with PEFT under identical prompts, data budgets and FAST-Eval metrics, varying the number of examples to establish when weight adaptation is worthwhile. Second, apply FAST-Eval in other domains such as legal and medical review, with new expert-validated labels to test transfer beyond CVs. Third, separate span selection from source fidelity through character-offset supervision, multiple acceptable reference spans and distinct metrics. Finally, test calibration under shifted attribute prevalence and evaluate selective prediction in a live review study, asking whether deferral helps people make more accurate decisions and recognise when model advice should be questioned.

8 Conclusion

This paper formalises the task of Finding Attributes with Supporting Text (FAST) and introduces FAST-Eval, which jointly evaluates communicated uncertainty, attribute identification and evidence grounding. On 500 held-out pairs, FAST-LoRA enables an 8B model to outperform both its untuned counterpart and the much larger GPT-5 on the primary calibration and identification measures, while matching GPT-5 on citation fidelity. Behaviourally, FAST-LoRA is less willing to infer presence from adjacent vocabulary alone and reserves high confidence only for strongly supported decisions, demonstrating the effectiveness of adaptation in the domain of employability skills. FAST-Eval is provided as a reusable uncertainty evaluation framework for FAST: its output schema and evaluation metrics apply across domains, provided suitable target documents and attributes are available.

Limitations

Data and generalisation. The evaluation covers 500 held-out pairs drawn from the same English CV corpus and annotation process as the training data. It therefore demonstrates within-corpus generalisation, not robustness across domains, document

genres or populations. Applying FAST to legal, clinical or multilingual material would require new domain-specific attributes, annotations and expert validation. The deliberately balanced benchmark also differs from deployment, where attribute prevalence and the number of queries per document may vary; because these shifts can affect precision, calibration and deferral, the reported values should not be read as deployment estimates.

Models and adaptation. The experiments use one open base model, Qwen3-8B, and one LoRA configuration (rank 16, $\alpha=32$), selected without a systematic hyperparameter search. Additional model families, LoRA ranks, training-set sizes and full fine-tuning are needed to establish how broadly the gains transfer. GPT-5 is evaluated zero-shot under the shared task instructions, so few-shot prompting, prompt optimisation or retrieved demonstrations may narrow the gap without changing its weights. The comparison between the Qwen3-8B conditions isolates adaptation, whereas comparison with GPT-5 also changes architecture, training data, tokenisation and serving infrastructure. No existing baseline produces directly equivalent outputs across all three FAST dimensions.

Post-hoc calibration. We do not compare against logit-based recalibration such as temperature scaling. This could be tested for the open-weight conditions, but not consistently for GPT-5 because the required logits were unavailable through the API. Moreover, it would recalibrate reported confidence without addressing the evidential threshold or citation grounding; future work should compare both logit-based and label-level recalibration methods.

What the confidence scores represent. The five levels communicate how strongly a document supports an attribute and are supervised against human ordinal labels. They are not direct measurements of predictive entropy and do not separate epistemic uncertainty from ambiguity in the document. OCE also requires numerical targets for the verbal levels: FAST-LoRA leads under the primary and expanded mappings, while GPT-5 has lower OCE when targets are compressed towards 0.50. OCE should therefore be interpreted alongside the mapping-independent Spearman correlation and ordinal-agreement metrics. The selective-prediction analysis is retrospective: it estimates accuracy and coverage at different confidence thresholds, but no model abstained during inference and

no human referral process was tested.

Annotation and evidence. The original annotator used a more conservative threshold than the independent expert (38% vs 46% positive in the re-annotated sample), so measured precision and recall partly reflect whose threshold is treated as ground truth. Beyond the 100 independently re-annotated pairs, the benchmark relies on one annotator. Each pair also contains only one reference citation, although several passages may validly support an attribute; hit rate can therefore reject reasonable alternatives. Conversely, hallucination rate establishes that a quotation occurs verbatim in the source, not that it fully entails the predicted attribute. These metrics capture complementary but incomplete aspects of evidence quality.

Output reliability and reproducibility. Outputs that remained unmappable after deterministic normalisation were excluded and reported separately. Because failure rates differ slightly by model, aggregate metrics compare evaluable outputs rather than an identical set of all 500 pairs; paired analyses use commonly evaluable cases. Finally, GPT-5 is a point-in-time closed-model comparison. Archiving its outputs permits exact re-scoring but cannot make the underlying inference process independently reproducible.

Ethical Considerations

The work targets a high-risk domain (EU AI Act Annex III) and is positioned as assistive, evidence-first decision support with a human in the loop, not autonomous screening. Meaningful oversight requires reviewers to question and override outputs: confidence should therefore be surfaced only from calibration-verified models (untuned level-5 predictions are wrong ~20–25% of the time, versus < 1% for FAST-LoRA), and citations should be treated as verification pointers rather than verdicts because model-selected spans match the annotated reference substantially less often than the independent expert’s. Predictions at levels 2–5 are paired with a source-checkable citation to make assessments contestable; level 1 explicitly communicates that no supporting evidence was found. CVs are treated as sensitive personal data: documents are used under the source dataset’s research terms, identified by anonymous numeric IDs, and processed locally except for the GPT-5 baseline API calls. Automated hiring tools can encode and

amplify bias; the evidence-citation design improves detectability but does not remove the risk, and any real deployment should include demographic auditing and a documented bias assessment. We follow the ACL Code of Ethics.

References

- Simone Alghisi, Massimo Rizzoli, Gabriel Roccabruna, Seyed Mahed Mousavi, and Giuseppe Riccardi. 2024. [Should we fine-tune or RAG? evaluating different techniques to adapt LLMs for dialogue](#). In *Proceedings of the 17th International Natural Language Generation Conference*, pages 180–197. Association for Computational Linguistics.
- Jeffrey Dastin. 2018. [Amazon scraps secret AI recruiting tool that showed bias against women](#). Reuters.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. [ERASER: A benchmark to evaluate rationalized NLP models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458. Association for Computational Linguistics.
- European Commission. 2017. [ESCO: European skills, competences, qualifications and occupations](#).
- European Parliament and Council. 2024. [Regulation \(EU\) 2024/1689 of the European Parliament and of the Council \(Artificial Intelligence Act\)](#). Official Journal of the European Union.
- Samuel Forster. 2026. Parameter-efficient domain adaptation for attribute extraction and reasoning in LLMs. Master’s dissertation, Trinity College Dublin.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. [Enabling large language models to generate text with citations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488. Association for Computational Linguistics.
- Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppel, Preslav Nakov, and Iryna Gurevych. 2024. [A survey of confidence estimation and calibration in large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [LoRA: Low-rank adaptation of large language models](#). *arXiv preprint arXiv:2106.09685*.

- Zhiqi Huang, Vivek Datla, Chenyang Zhu, Alf Samuel, Daben Liu, Anoop Kumar, and Ritesh Soni. 2025. [Confidence-based response abstinence: Improving LLM trustworthiness via activation-based uncertainty estimation](#). In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainNLP 2025)*, pages 184–193, Suzhou, China. Association for Computational Linguistics.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Computing Surveys*, 55(12):1–38.
- Iris Kamsteeg, Juan Cardenas-Cartagena, Floris van Beers, Tsegaye Misikir Tashu, and Matias Valdenegro-Toro. 2025. [Confidence calibration in large language model-based entity matching](#). In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainNLP 2025)*, pages 120–137, Suzhou, China. Association for Computational Linguistics.
- Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. 2024. [Calibration-tuning: Teaching large language models to know what they don’t know](#). In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertainNLP 2024)*, pages 1–14, St Julians, Malta. Association for Computational Linguistics.
- Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. [Inferring which medical treatments work from reports of clinical trials](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717. Association for Computational Linguistics.
- Xinyue Liu, Harshita Diddee, and Daphne Ippolito. 2024. [Customizing large language model generation style using parameter-efficient finetuning](#). In *Proceedings of the 17th International Natural Language Generation Conference*, pages 412–426. Association for Computational Linguistics.
- Xu Liu, Guanyi Chen, Kees van Deemter, and Tingting He. 2025. [Annotating hallucinations in question-answering using rewriting](#). In *Proceedings of the 18th International Natural Language Generation Conference*, pages 823–832. Association for Computational Linguistics.
- Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. [Reducing conversational agents’ overconfidence through linguistic calibration](#). *Transactions of the Association for Computational Linguistics*, 10:857–872.
- Opensporks. 2024. [Resumes](#). Hugging Face Hub dataset opensporks/resumes. English-language CV corpus; documents taken from the Resume_str plain-text field.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. [Measuring attribution in natural language generation models](#). *Computational Linguistics*, 49(4):777–840.
- Patricia Schmidtova, Eduardo Calò, Simone Balloccu, Dimitra Gkatzia, Rudali Huidrom, Mateusz Lango, Fahime Same, Vilém Zouhar, Saad Mahamood, and Ondrej Dusek. 2025. [Do my eyes deceive me? a survey of human evaluations of hallucinations in NLG](#). In *Proceedings of the 18th International Natural Language Generation Conference*, pages 60–79. Association for Computational Linguistics.
- Craig Thomson and Anya Belz. 2024. [\(mostly\) automatic experiment execution for human evaluations of NLP systems](#). In *Proceedings of the 17th International Natural Language Generation Conference*, pages 272–279. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. [Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. [Fact or fiction: Verifying scientific claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7534–7550. Association for Computational Linguistics.
- Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. [Large language models for generative information extraction: A survey](#). *Frontiers of Computer Science*, 18(6):186357.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.

Gal Yona, Roei Aharoni, and Mor Geva. 2024. [Can large language models faithfully express their intrinsic uncertainty in words?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764, Miami, Florida, USA. Association for Computational Linguistics.

Mike Zhang, Kristian Nørgaard Jensen, Sif Sonniks, and Barbara Plank. 2022. [SkillSpan: Hard and soft skill extraction from English job postings](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4962–4984. Association for Computational Linguistics.

Weijia Zhang, Mohammad Aliannejadi, Yifei Yuan, Jiahuan Pei, Jia-hong Huang, and Evangelos Kanoulas. 2024. [Towards fine-grained citation evaluation in generated text: A comparative analysis of faithfulness metrics](#). In *Proceedings of the 17th International Natural Language Generation Conference*, pages 427–439. Association for Computational Linguistics.

A Data and Code Availability

The benchmark is derived from the publicly available Opensporks CV corpus ([Opensporks, 2024](#)) and ESCO taxonomy ([European Commission, 2017](#)). Source code and benchmark annotations will be released upon publication at <https://github.com/FAST-Eval/FAST-LoRA>.