SkillSight

Detecting Explicit vs. Implied Skills in Text (0.5 / 1)

Presented by: Yonatan Elman, Michael Kovalchuk, Roni Fadlon

Project Review

Motivation
Extract CS/Hi-Tech skills from resumes with Explicit/Implicit labels (from a predefined Global Skill Vector).

1. Vector → Sparse list: only relevant skills are stored, labeled 1.0 / 0.5.
2. We added skill aliases (for explicit mentions).
3. We defined a fixed global vector of 110 CS-related skills.
4. We switched from 0/1/2 to 1.0/0.5, where 0 means the skill is not present / not predicted.

Novelty
The model extracts implicit skills from context, not only explicit keywords.

Literature Review

Paper (Year)	Task	Methods	Data	Results	Relation to our project
Implicit Skills Extraction Using Document Embedding (2020)	Skill ext. + CV↔JD match; add implicit	Doc2Vec similar JDs + transfer	1.1M JDs	F1=0.83; MRR=0.88	implicit via similar JDs; ours per-text evidence
A Survey on Skill Identification From Online Job Ads (2021)	Survey of skill ID from job ads	Taxonomy (bases/methods /granularity)	108 papers	challenges + trends	background only (no explicit/implicit evidence)
SKILLSPAN Hard and Soft Skill Extraction (2022)	Span-level skill extraction + dataset	BERT + domain-adaptive; single vs multi	391 JPs; 14.5K sentences	domain-adapted better; single>multi	explicit spans; ours evidence (incl. implicit)

Dataset

- Synthetic JSONL dataset: job_description (4–6 sentences) + skills (3–6 skills) from a predefined GLOBAL_SKILL_VECTOR.
- Labels per skill: 1.0 = Explicit, 0.5 = Implicit.

- Generated with LLM (gpt-4o mini) using a structured prompt
- aliases for Explicit skill.
- Post-generation: automatic validation & repair to prevent leaks/missing explicit mentions.
- 473 synthetic samples
- 110 skills in the Global Skill Vector

EDA

EDA Summary

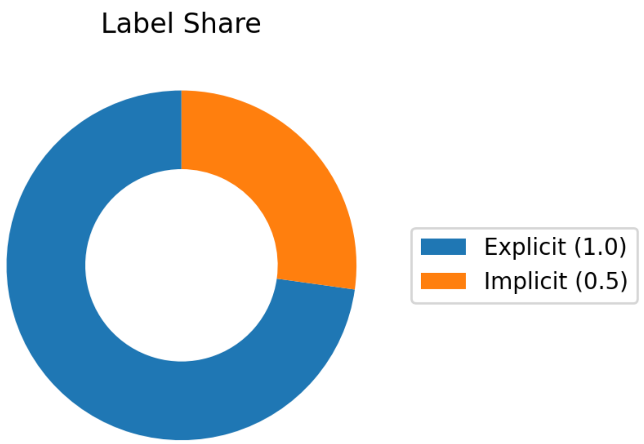
Samples: 473

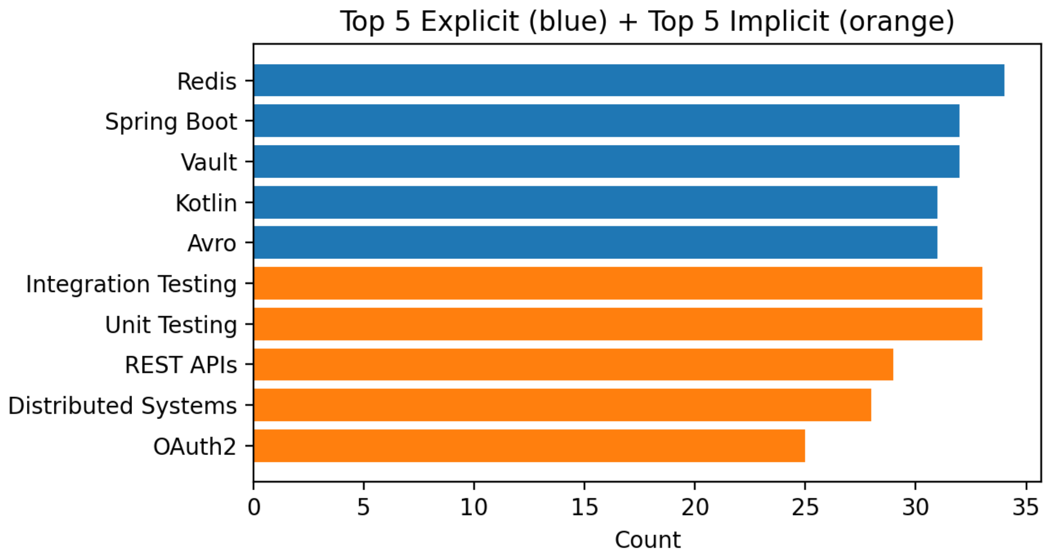
Global skills: 110

Total labeled skill occurrences: 2318

Explicit vs Implicit: 72.8% / 27.2%

Samples with 0 implicit: 20.9%





Quality (Types) + Length

Missing keys: 0 (0.0%)
Bad skills type: 0 (0.0%)
Duplicates: 0 (0.0%)
Non-ASCII: 1 (0.2%)
Double spaces: 0 (0.0%)
Control chars: 0 (0.0%)
Unknown skills: 1
Bad label values: 0

Sentence count: mean 5.53, median 6.0, min/max 4/8
Words: mean 92.4, median 91.0, min/max 60/131
Chars: mean 701.6, median 696.0, min/max 474/1009

Imbalance: zero-occ 1, <10 4, <20 44, Top10 cover 14.15%

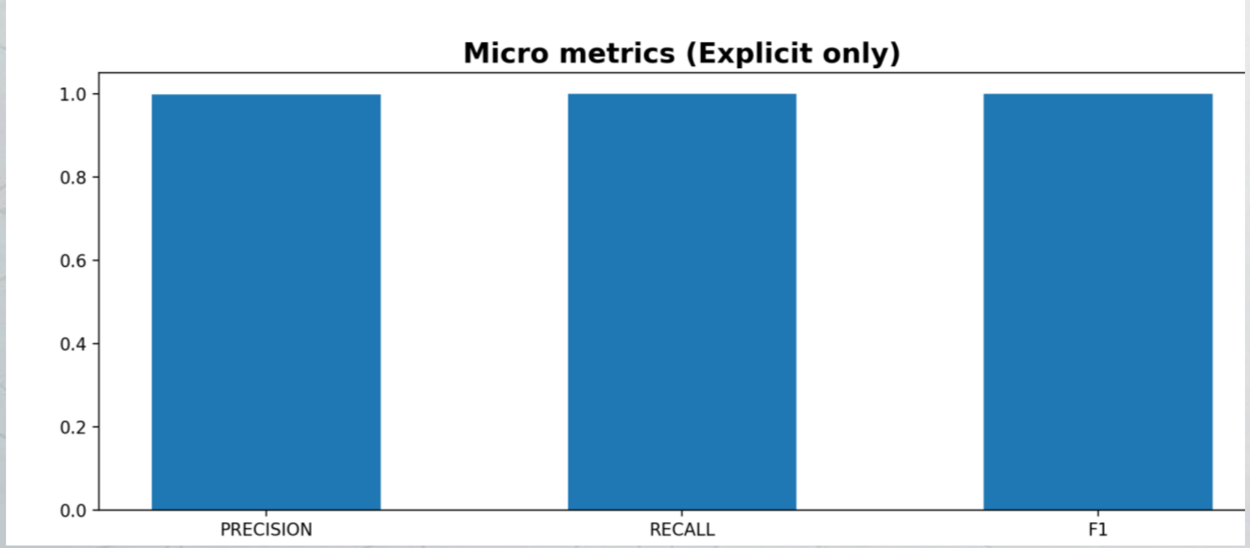
Co-occurrence: Top pair (Integration Testing, Unit Testing) appears 11/473.

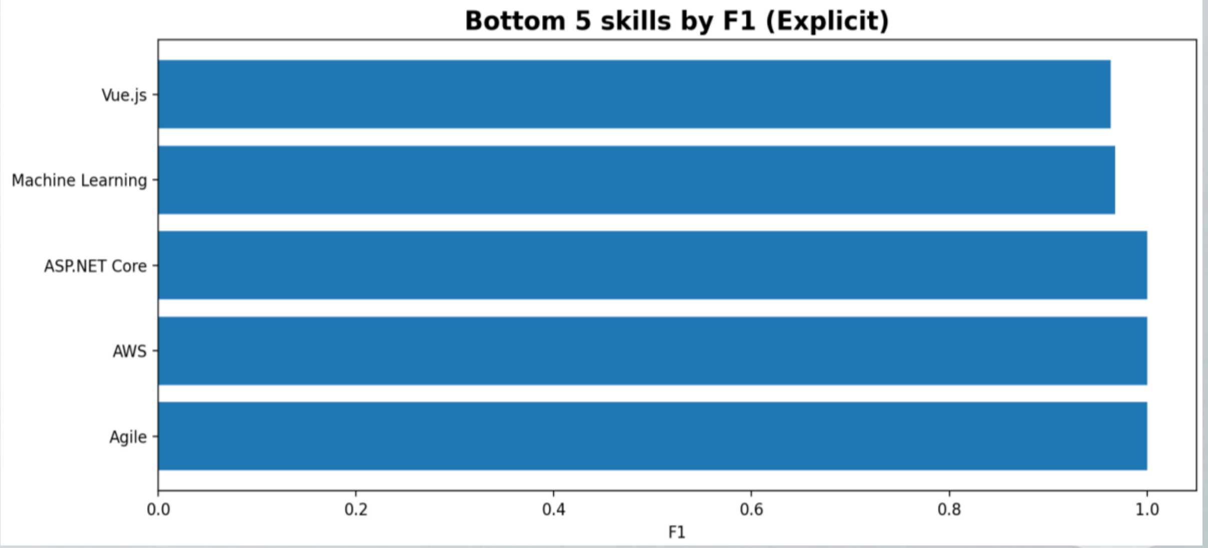
Baseline 1- keyword matching

- Type: keyword matching (aliases + word-boundary regex)
- Output: Explicit only (1.0 if match else 0.0)
- N=473 | Micro F1=0.999 | TP/FP/FN=1688/2/0
- Errors: FP=2, FN=0
- High explicit scores are expected because labels rely on alias presence.

FP #1 — Machine Learning (gt=0.5 → pred=1.0)
Trigger: “machine learning initiatives”
Reason: Explicit-only baseline

FP #2 — Vue.js (gt=0.5 → pred=1.0)
Trigger: “Vue.a”
Reason: spurious match





Baseline 2- zero shot

- Type: Zero-shot classification (LLM + prompt) over GLOBAL_SKILL_VECTOR
- Output: predicts 0.5 (Implicit) / 1.0 (Explicit)
- N=473

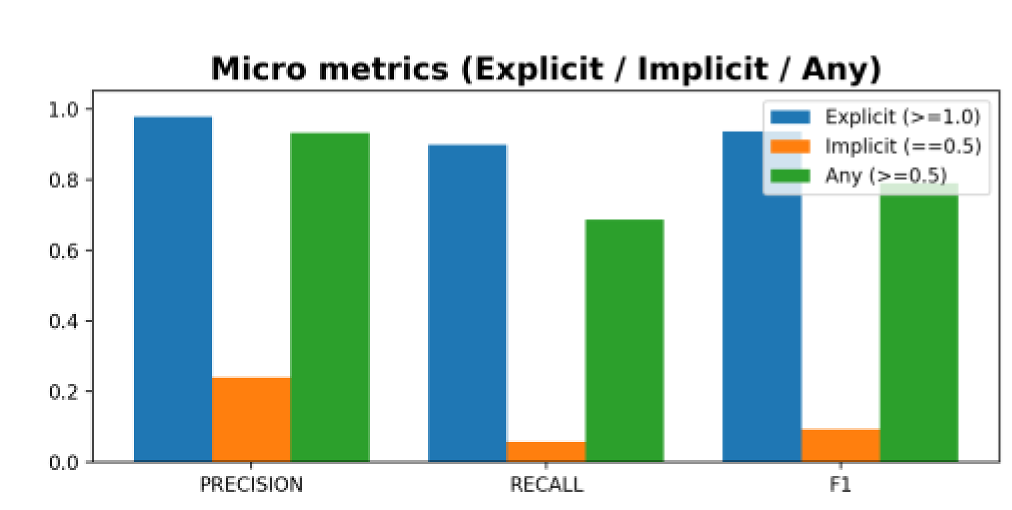
Micro F1:

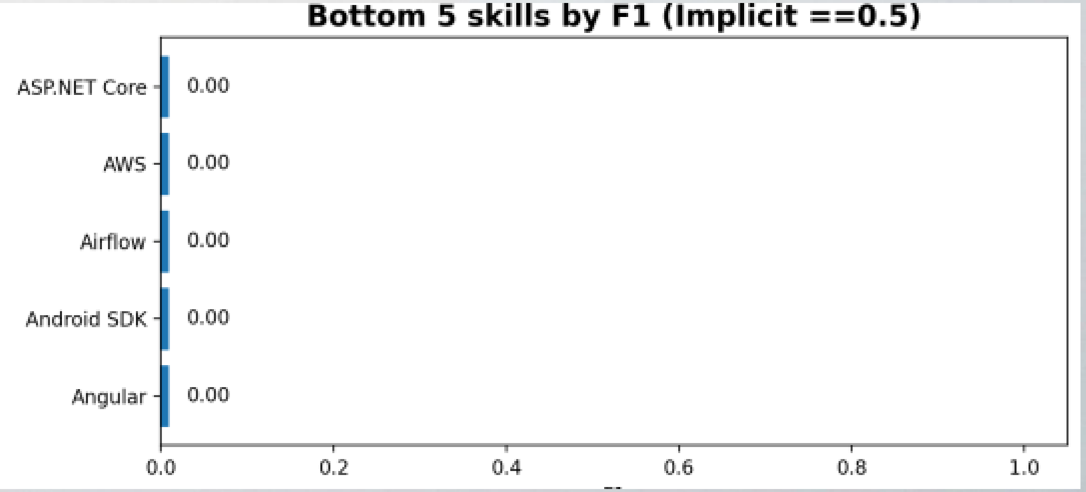
- Implicit (==0.5): 0.094 (P=0.240, R=0.059)

Key takeaway: Strong on Explicit, very weak on Implicit (low recall).

FN #1 — Load Balancing (gt=0.5 → pred=0)
Trigger: “peak traffic”, “rate limiting”
Reason: Implicit-only cue (no explicit mention)

FP #1 — Tailwind CSS (gt=1.0 → pred=0.5)
Trigger: “using Tailwind …”
Reason: Explicit mention, but model downgraded to implicit (uncertainty/noisy phrasing)





Plan

Step	When	What we will do (Scope)	Expected outcome
1	Week 10	Generate 1,500–2,000 synthetic samples; strengthen implicit cues; reduce label leakage; rebalance 0/0.5/1	Larger dataset (1,500–2,000) + cleaner implicit ground truth
2	Week 10	Re-run EDA + baselines on updated data	Updated metrics + identified failure modes
3	Week 10	Stronger baseline inference (zero-shot with stronger LLM)	Better reference baseline for implicit
4	Week 11	Build and fine-tune a classifier for 0/0.5/1 using BERT/RoBERTa	Build a dedicated model for explicit vs implicit labeling
5	Week 11	Error analysis iteration (top errors + skill-specific fixes)	Targeted improvements + final evaluation
6	Before deadline	Prepare final presentation + report	Final submission package