Overview of TalentCLEF 2026: Task B, Job-Skill Matching with Skill Type Classification

Warre Veys[1,*], Luis Gasco[2], Matthias De Lange[1] and Jens-Joris Decorte[1]

[1]TechWolf, Ghent, Belgium
[2]Avature Machine Learning, Spain

Abstract

This paper presents Task B of TalentCLEF 2026, the second edition of the shared task on job-skill matching: given a job title, systems retrieve the relevant skills from the ESCO taxonomy and distinguish, for each skill, whether it is core to the occupation or contextual, required only in an identifiable variant of it. To support it, the evaluation collection was re-annotated from the ground up with an LLM-as-a-judge pipeline that judged every candidate skill against its full ESCO description under a strict requirement-based rubric, re-adjudicated every relevance judgment of the 2025 edition, and was validated by a human reviewer and an independent audit. The resulting test collection grew from 1,986 to 5,288 distinct ESCO skills and from 112 to 184 relevant skills per job title, and every metric is reported in a binary and a graded regime. The task attracted 101 registered participants, of whom 15 teams submitted a total of 43 runs, all of which we independently re-scored. The best system reaches an NDCG(graded) of 0.8068 and a MAP(binary) of 0.4197. Three findings structure the analysis. Every valid submission ranks core skills above contextual ones, evidence that the graded signal is learnable and internally consistent. The field saturates the easy retrieval target while leaving the hard one open, with best MRR above 0.90 against MAP below 0.42, a gap that the residual incompleteness of any pooled collection only widens. And the collection covers broadly applicable skills far better than occupation-specific ones, a bias that future editions should address.

Keywords

Skill matching, Information retrieval, Human capital management, Graded relevance, ESCO, LLM-as-a-judge, TalentCLEF

1. Introduction

Skills have become the unit of account of the modern labor market. Organizations increasingly describe jobs, source candidates, plan their workforce, and design reskilling programs in terms of skills rather than job titles or degrees, a shift documented across industry and policy reports [1, 2, 3, 4]. Each of these applications depends on the same underlying capability: given a job, determine which skills it requires. Because jobs, skills, and the connections between them are expressed almost entirely in text, this is a natural language processing problem at its core.

The first edition of TalentCLEF posed this problem as retrieval: given a job title, return the relevant skills from a knowledge base [5]. A flat list of relevant skills, however, leaves out a distinction that practitioners need. A recruiter screening for a software engineer must know that debugging is non-negotiable while style sheet languages matter only for front-end variants of the role, a workforce planner computing skill gaps must weight a missing essential skill differently from a missing nice-to-have. This edition of Task B adds that layer. Each relevant skill is annotated as *core*, required to perform the occupation regardless of employer or work context, or *contextual*, required only in a clearly identifiable variant of the occupation, and the evaluation rewards systems that rank core skills above contextual ones.

The task is harder than it may appear. Job-title queries are short and noisy: the test titles average 4.7 words and one in five carries requisition codes, location markers, or other non-semantic tokens. The target inventory is large: the ESCO taxonomy [6] contains roughly 13.9 thousand skills, many

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany
*Corresponding author.
warre.veys@techwolf.ai (W. Veys); machinelearning@avature.net (L. Gasco)
0009-0000-6838-9954 (W. Veys)

of them near-synonyms whose scope is only disambiguated by their definitions. Whether a skill is essential to an occupation is a graded human judgment, not a lexical fact, and research on the problem has remained fragmented across private datasets, label schemes, and evaluation protocols [7, 8, 9].

This edition contributes four things. First, a graded core/contextual relevance scheme over ESCO skills, operationalized in a strict requirement-based annotation rubric. Second, a substantially larger evaluation collection produced by an LLM-as-a-judge pipeline that judged every candidate skill against its full ESCO description and explicitly re-adjudicated every relevance judgment of the 2025 edition: the test collection grew from 1,986 to 5,288 distinct ESCO skills and from 112 to 184 relevant skills per job title. Third, a dual evaluation in which every metric is computed under binary relevance and under graded relevance, with NDCG(graded) as the main metric. Fourth, the results and analysis of 43 submissions from 15 teams, all of which we independently re-scored.

Three findings give the analysis its structure. Every valid submission, without exception, ranks core skills above contextual ones on average, substantial evidence that the graded annotation is internally consistent and that the importance signal is learnable (Section 6.4). Retrieving a first relevant skill is close to solved while ranking the full relevant set is not: the best systems exceed 0.90 in MRR while no system reaches 0.42 in MAP (Section 6.2), a gap that reflects the difficulty of full-set ranking but also the residual incompleteness that any pooled collection carries. And the collection covers broadly applicable skills far better than occupation-specific ones, which only enter at 24% coverage, a bias that matters for anyone training on the data (Section 3.4).

The remainder of the paper is organized as follows. Section 2 situates the task in prior work. Section 3 defines the task and describes the released collection. Section 4 presents the dataset construction and the LLM-as-a-judge annotation pipeline, including its human validation and an audit of its disagreements with the 2025 ground truth. Section 5 surveys the participating systems, Section 6 reports the results, and Sections 7 and 8 discuss the lessons of the edition and conclude.

2. Related Work

Skill extraction, normalization, and job-skill modeling. A first body of work extracts skills from job-market text, from span-annotated job postings [10] and weakly supervised extraction [11] to extraction with large language models [12]. Senger et al. [7] survey this area. A second body normalizes the extracted surface forms to taxonomies, with ESCO [6] as the dominant target, and learns representations that connect jobs and skills: JobBERT models job titles through the skills they require [8, 13], and ESCOXLM-R pre-trains directly on the taxonomy [9]. Closest to this edition's framing is the line of work that asks not merely which skills relate to a job but how strongly they are required: ranking skill requiredness for a title [14], evaluating skill relatedness [15], and inferring job-skill framework links with graph neural networks [16]. The core/contextual distinction in Task B is a graded, taxonomy-grounded version of that essential-versus-optional question.

Benchmarks and shared tasks in HR NLP. Dedicated venues for NLP in human resources have emerged only recently [17, 18], and shared evaluation has lagged behind: studies typically rely on incompatible private datasets and label schemes. TalentCLEF was introduced to consolidate evaluation in this domain [19, 5] and continues in its second edition alongside this task [20, 21], with Task A covering contextualized job-person matching [22] and community-driven evaluation efforts such as WorkRB extending the benchmark landscape [23].

Graded relevance and LLM-assisted annotation. Evaluating a two-level importance scheme is a classic graded-relevance problem, and NDCG [24] is its natural metric. Section 3 describes how the binary and graded regimes are computed side by side. The collection itself was built with an LLM acting as relevance judge. Using large language models as evaluators has become standard practice in general NLP [25], and for relevance assessment specifically, Thomas et al. [26] provide substantial evidence that LLM judgments can match the quality of trained human assessors at a fraction of the cost. Section 4 describes our pipeline, its precision-first rubric, and the validation exercises that probe its reliability. The embedding and reranking methods the participants built on [27, 28, 29, 30] are discussed with the

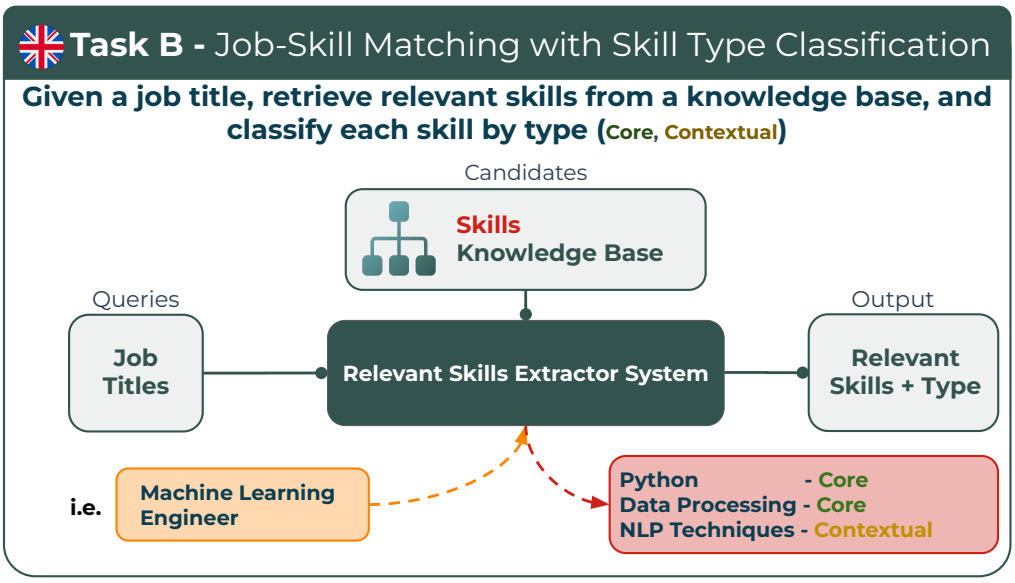


Figure 1: Overview of Task B: given a job title, systems retrieve the relevant skills from the ESCO-grounded corpus and the evaluation distinguishes core from contextual skills through graded relevance.

systems in Section 5.

3. Task and Data

3.1. Task definition

Task B is a retrieval task with a graded skill-type layer. Given a job title as query, a system returns a ranking over a corpus of ESCO skills, each skill that is relevant to the queried job is classified as either *core* or *contextual*. This per-skill classification is what the official task name, *skill type classification*, refers to [21]: the two skill types carry different relevance grades, and a system demonstrates the classification by ranking the higher-graded type above the lower one. The two views are therefore the same objective seen from two sides, and most participants realized the skill-type layer through graded ranking rather than a separate classification head. Core skills are required to perform the occupation in essentially every realistic instance of it, independent of employer or work context. Contextual skills are required only in a clearly identifiable variant of the occupation, depending on factors such as industry, organization, or specialization. Figure 1 gives a schematic overview.

The skill-type layer is incorporated into the evaluation through two relevance regimes, mirroring the official scoring program. In the *binary* regime, core and contextual skills both count as relevant with grade 1, which measures whether systems retrieve relevant skills at all. In the *graded* regime, core skills carry grade 2 and contextual skills grade 1, which rewards systems that rank core skills higher. Every metric reported in this paper (NDCG, MAP, MRR, P@k) is computed in both regimes, and NDCG under graded relevance is the official main metric [21]. The task is monolingual English this edition, which distinguishes it from Task A's bilingual English-Spanish setting [22].

Table 1 makes the core/contextual distinction concrete with a short annotated sample for the test title *Software Engineer*, chosen because its contextual skills read transparently as "required only in a named variant". The skills are quoted verbatim from the released test judgments. The ESCO description column shows what the annotation pipeline actually judged: as Section 4.3 details, each candidate skill was presented to the judge as "name: description", and the example should be read the same way rather than as a name-only judgment. The variant gloss is editorial and previews the annotation rubric.

Table 1
An annotated sample for the job title *Software Engineer* (`qb_jt_19`, 27 core and 252 contextual skills in total). Skills and descriptions are quoted verbatim from ESCO v1.2.0. Only a few of each type are shown, and multi-sentence descriptions are shortened to their first sentence. The judge saw the full description: each skill was presented as "name: description" (Section 4.3). Core skills are required in almost every realistic instance of the role, while contextual skills are required only in a clearly identifiable variant.

Type	ESCO skill	ESCO description (shown to the judge)	Variant / rationale
Core	debug software	Repair computer code by analysing testing results, locating the defects causing the software to output an incorrect or unexpected result and remove these faults.	nearly all developer roles
Core	algorithms	The self-contained step-by-step sets of operations that carry out calculations, data processing and automated reasoning, usually to solve problems.	foundational across engineering roles
Core	align software with system architectures	Put system design and technical specifications in line with software architecture in order to ensure the integration and interoperability between components of the system.	fit the surrounding system
Contextual	style sheet languages	The field of computer language that conveys the presentation of structured documents such as Cascading Style Sheets (CSS).	front-end / web variants
Contextual	cyber attack countermeasures	Methods, technologies and techniques used to defend (detect, monitor and recover) against cyber attacks.	security-focused variants
Contextual	coordinate engineering teams	Plan, coordinate and supervise engineering activities together with engineers and engineering technicians.	lead / senior variants

Table 2
Statistics of the development and test sets.

Statistic	Development	Test
Number of queries	304	125
of which judged	304	117
Number of corpus elements	9,052	5,358
distinct ESCO URIs	8,942	5,288
Avg. relevant items per query	185.6	184.2

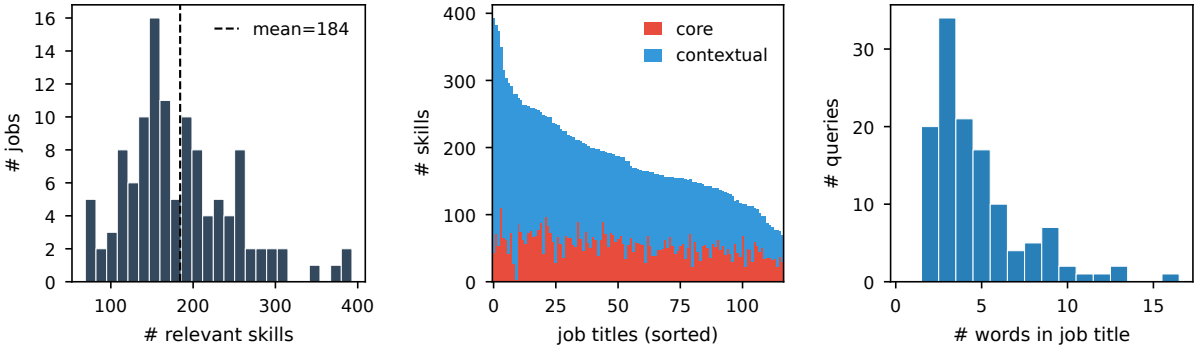
3.2. The released collection

The collection ships three splits. The *training* set is a list of the most representative skills per ESCO occupation, derived from the taxonomy and marked essential or optional per occupation. The *development* set holds 304 job-title queries and the *test* set 125, of which 117 carry relevance judgments. The 8 unjudged test titles are mostly non-jobs that survived from the source data, namely onboarding artifacts such as "Onboarding Task - Day 1", a department header, and similar noise. The official scoring program restricts every metric to the 117 judged titles, and our re-scoring does the same. The job-title queries are the same as in the 2025 edition, and the skill side of the collection was rebuilt, as Section 4 describes. Table 2 summarizes the corpus statistics.

The corpus of each split is the closed pool of skills judged relevant to at least one of its jobs, that is, the union of the relevance judgments. The retrieval target space therefore holds about 5.3 thousand skills, of which about 3.4% (184 of 5,358) are relevant to any given job, the density against which the precision-at-depth figures should be read.

Table 3
Skill-type composition of the judgments per split. Core skills carry grade 2, contextual skills grade 1.

Split	Judgments	Core	Contextual	Core fraction	Core / contextual per query
Development	56,418	16,161	40,257	28.6%	53.2 / 132.4
Test	21,554	6,356	15,198	29.5%	54.3 / 129.9



(a) (b) (c)

Figure 2: Properties of the test collection. (a) Relevant skills per job (mean 184, median 168). (b) Core versus contextual judgments per job. (c) Job-title length in words (mean 4.7, and 21% of titles carry codes or digits).

The skill-type layer is new this edition, and Table 3 reports its composition. The 21,554 test judgments split into 6,356 core (29.5%) and 15,198 contextual (70.5%), a ratio of about 1:2.4, or roughly 54 core and 130 contextual skills per job. The development split carries an almost identical 28.6% core fraction. The grade distribution is therefore stable across splits, and systems could rely on the development set to calibrate the graded objective.

3.3. Properties of the collection

Figure 2 characterizes the test collection along the three properties that bear most directly on system behavior in Section 6: the target density, the per-job grade split, and the query noise. Panel (a) shows the distribution of relevant skills per job: the mean is 184 and the median 168, a dense target by retrieval standards. Panel (b) shows the core/contextual split of Table 3 per job. Panel (c) shows query-title length: the mean title is 4.7 words, the longest 16, and 21% of titles carry digits, requisition codes, or similar non-semantic tokens. Each query additionally ships a free-text job description of about 47 words on average.

The panels together read as: the target is dense, the queries are short and noisy. Pure lexical matching is poorly suited to this shape of problem, and the per-query difficulty analysis in Section 6.5 confirms that exactly the coded and narrow titles are where systems fail.

3.4. Coverage against the full ESCO taxonomy

Because every corpus skill is grounded in an ESCO URI, the collection can be placed against the full taxonomy. The full ESCO v1.2.0 inventory holds 13,939 skills. The 117 judged test jobs touch 5,288 of them (37.9%).

Coverage is far from uniform, and the direction of the non-uniformity is informative. Figure 3(a) breaks coverage down by ESCO's `reuseLevel`, and the frequency behaviour is as expected: skills ESCO marks as cross-sector are covered at 54.1% and transversal skills at 45.4%, while sector-specific skills enter at 34.8% and occupation-specific skills at only 23.6%. By ESCO's `skillType` axis the picture is flat, with knowledge at 37.3% and skill/competence at 38.1%. Figure 3(b) shows skill popularity over the full taxonomy, with the never-relevant 62% as the dominant mass.

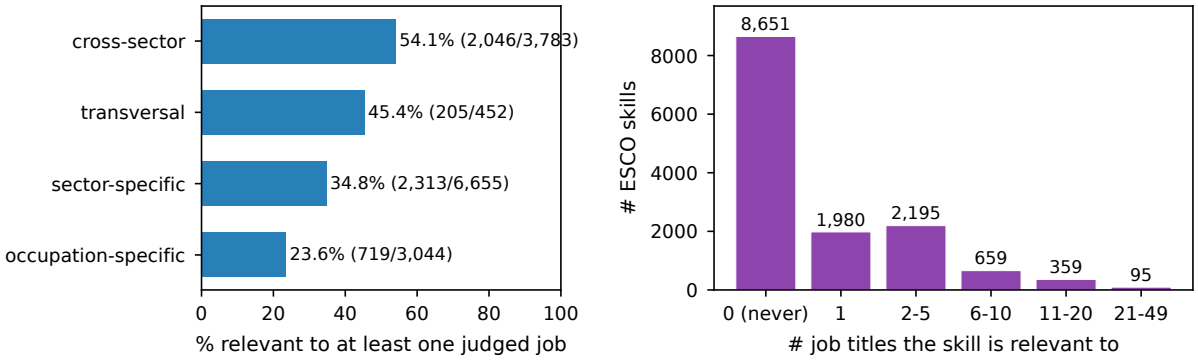


Figure 3: The collection against the full ESCO v1.2.0 taxonomy (13,939 skills). (a) Share of ESCO skills that are relevant to at least one judged test job, by ESCO reuse level: broadly applicable skills are covered about twice as well as occupation-specific ones. (b) Popularity over the full inventory: 62% of ESCO is never relevant to any judged job.

4. Dataset Construction and Annotation

The methodological centerpiece of this edition is the re-annotation that produced the graded collection. This section describes the prior edition's process as the starting point, the scale of the change, the LLM-as-a-judge pipeline itself, and the two exercises that probe its quality.

4.1. The 2025 annotation process and why we re-annotated

The first edition built its Task B corpus from skill sets automatically extracted from applicant CVs, the 7,493 skills appearing at least 90 times in that data were kept and manually mapped to ESCO v1.2.0 [5]. Relevance was annotated by four annotators from a pool of linguists using the ESCO taxonomy as reference, in a multi-stage process with an initial quality-control stage and a complete final review. The test set contained 2,400 annotations, about a third of which were edited during review, and both sets were independently reviewed by an expert. Two structural properties, however, bounded what this process could deliver, and both motivated the move to LLM-assisted annotation this edition. The relevance scheme was binary, with no notion of how essential a relevant skill is, whereas this edition's core/contextual grade requires judging every candidate against a precise requirement rubric, a per-skill reasoning step that scales poorly under manual annotation. And the candidate pool was limited to what CV skill extraction surfaced. That pool is structurally incomplete: a CV lists what an applicant chooses to write down, not what the occupation requires, and the requirements a job title makes self-evident are exactly the ones applicants leave implicit rather than spell out as skill mentions. Section 4.5 quantifies this upstream recall ceiling. Lifting it means judging the full ESCO taxonomy as the candidate space rather than a pre-filtered list, which multiplies the judgment volume beyond what four annotators can cover. An LLM judge applied under a strict, auditable rubric addresses both at once: it admits the graded reasoning the binary human scheme lacked, and it makes taxonomy-wide candidate judging affordable. The trade-off it introduces, single-pass machine judgment in place of multi-annotator human review, is the subject of the validation and audit in Sections 4.4 and 4.5.

4.2. From the 2025 collection to a graded one

The re-annotation reused the 2025 job-title queries and rebuilt the skill side of the collection from the taxonomy up. This was a re-annotation, not a touch-up, and Table 4 quantifies it on the test split, matching the two versions by ESCO URI because the corpus identifier namespace was remapped between editions. Of the 13,116 old distinct (job, ESCO URI) pairs, only 4,264 (32.5%) survive into the new release, 8,852 were dropped and 17,132 added. The mean number of relevant skills per job grew from 112 to 184 and the corpus from 1,986 to 5,288 distinct ESCO URIs.

Table 4
The test-split re-annotation in numbers, matched by ESCO URI over the 117 judged titles. Because the corpus id namespace was remapped between editions, every judgment is reduced to a distinct (job, ESCO URI) pair, which collapses the few duplicate corpus ids of the same URI (21,396 pairs versus the 21,554 raw judgments of Table 3).

Statistic	2025	2026
Distinct (job, ESCO URI) pairs	13,116	21,396
Grade scheme	binary	graded (core = 2, contextual = 1)
Distinct corpus ESCO URIs	1,986	5,288
Mean relevant skills per job	112.1	184.2
Judgments retained from 2025	4,264 (32.5% of 2025)	
Judgments dropped / added	8,852 / 17,132 (Jaccard 0.14)	

4.3. The annotation pipeline

The pipeline has three stages: candidate generation, LLM judging, and export. All specifics below are taken from the released run's code and artifacts.

Stage 1: candidate generation. For each of the 421 job titles (304 development, 117 judged test), candidate skills were proposed by embedding-based retrieval: the job title and every skill in the ESCO inventory were embedded with JobBERT-v2 [13],[1] and the 750 most similar skills by cosine similarity formed the shortlist. The shortlist size was chosen from precision-recall curves against the 2025 annotations. The candidate set is the union of this shortlist with all skills the 2025 edition marked relevant for that title: every previously relevant skill that the retriever did not surface was appended. The judge therefore ruled on every 2025 judgment explicitly, and no prior label was dropped silently. The resulting candidate sets average 805.8 skills per title.

Stage 2: LLM-as-a-judge labeling. Each candidate skill was labeled *Core*, *Contextual*, or *Non-relevant* by `gpt-5-mini` (version `gpt-5-mini-2025-08-07`, default decoding settings), with a one-sentence justification per skill. The judge sees ESCO descriptions, not bare skill names: the user prompt presents the job title followed by the candidate skills rendered as "*skill name*: *ESCO description*", with descriptions from ESCO v1.2.0. This is essential context for reading the annotations, because ESCO's terse names frequently hide their scope: a label like "perform product testing" reads as on-topic for a software role, yet its ESCO description scopes it to testing processed workpieces for manufacturing faults, and the judge could only reject it because it saw that description. The audit in Appendix A measures how often the description was decisive in exactly this way. Skills were judged in batches of 20 per call, and responses were validated against a structured output schema that requires exactly one label and one justification per submitted skill. Invalid or failed calls were re-queued until every batch completed. The released run comprised 17,173 LLM calls producing 339,248 individual skill judgments over the 421 titles.

The rubric, reproduced in full in Appendix C, is deliberately precision-first: the default label is Non-relevant, Core requires passing a six-gate test (including a near-universal requirement probe, vendor independence, and primary-mission responsibility), and Contextual requires naming a structural variant of the occupation that owns the skill, with explicit guards against known failure modes such as the industry fallacy (treating domain knowledge of the employer's industry as a job requirement), transversal-skill traps, and confusing workplace behaviors with job-architecture requirements. The prompt also instructs the judge to expect only a handful of Core skills per occupation, compensating for the over-labeling bias that batched judging otherwise induces.

Stage 3: export. Core judgments map to grade 2, Contextual to grade 1, and Non-relevant judgments are dropped. Each released skill is grounded in its ESCO v1.2.0 URI, resolved through the preferred and alternative labels, and ships its full alias list, a mean of 7.0 ESCO surface forms (the preferred label plus alternative labels) per skill.

[1] JobBERT-v2 has no dedicated publication. The cited paper introduces its successor JobBERT-v3 and is the closest published description of the model family.

Table 5
Human validation of the LLM annotations: agreement between the human judgment (columns) and the LLM label (rows), over 847 jointly labeled skills across 22 job titles. Diagonal cells are agreements. Exact three-way agreement is 68.6%, relevant-versus-non-relevant agreement 84.1%.

	Human judgment			
LLM label	Core	Contextual	Non-relevant	Total
Core	**240**	63	15	318
Contextual	68	**196**	43	307
Non-relevant	44	33	**145**	222
Total	352	292	203	847

4.4. Human validation of the LLM annotations

A human annotator validated a sample of the LLM-generated labels through a per-skill review interface, judging each shown skill as core, contextual, or non-relevant. The sample covers **1,060 skill-level judgments across 22 job titles**. Of these, 847 align with a skill the LLM placed in one of its decision buckets and are used for the comparison in Table 5. The human and the LLM agree exactly on the three-way label in **68.6%** of cases, agree on the coarser relevant-versus-non-relevant decision in **84.1%**, and, restricted to skills both consider relevant, agree on the core-versus-contextual grade in **76.9%**. The disagreements are dominated by the core/contextual boundary (for example 63 skills the LLM called core the human called contextual, and 68 the reverse). The hard call in a graded scheme is degree of essentialness, the one axis on which two careful annotators can reasonably differ, and it is exactly where the disagreement concentrates. The audit in Section 4.5 reaches the same conclusion from the independent re-judgment side.

Two caveats bound what this validation establishes. First, the reviewed sample covers 22 of the 421 annotated jobs, weighted toward operational and service titles such as police officer, prep cook, and bar person. It is a spot-check of quality, not exhaustive validation. Second, the review was performed by a single annotator without adjudication, which means the numbers measure human-versus-LLM consistency on a reviewed sample rather than agreement against an established gold standard. That consistency is the relevant signal for whether the released labels are trustworthy, and a larger multi-annotator validation remains future work.

4.5. Where the 2025 candidate pool reached its limits

Re-annotating an existing human dataset from scratch deserves justification beyond the new grade layer. The 2025 collection drew its candidate skills from skills extracted across many applicant CVs. That design is sound and, in the limit, complete: for a common occupation with a large and diverse pool of applicants, the union of skills seen across CVs converges toward every skill the role can demand. The method's coverage is therefore tied to how many CVs exist for a given title, and it thins precisely for the rare and narrow occupations where few applicants were ever observed. The new pipeline's contribution is to replace the CV-bounded candidate space with the full taxonomy, which removes the dependence on applicant volume.

The first line of evidence concerns skills the 2025 set lacked. For shared titles, many clearly essential skills were absent from the 2025 judgments entirely. For *Software Engineer*, 21 of the 27 skills the new collection marks as core do not appear in the 2025 set at all, among them "align software with system architectures". The pattern holds across titles: *Warehouse Associate* lacked "handle cargo" and *Financial Analyst* lacked "estimate profitability". These gaps follow mechanically from the prior candidate pool: a skill that CV extraction never surfaced could never be judged, however careful the annotators. The new pipeline removes this ceiling by retrieving candidates from the full taxonomy.

A second, sharper line of evidence is the set of 2025 relevance judgments that the judge explicitly rejected. Because the candidate union of Section 4.3 appended every 2025-relevant skill to the judged

pool, none of these rejections is a silent omission: the judge saw each of these skills, with its ESCO description, and classified it non-relevant.[2] The rejections split into two informative groups, according to whether the embedding retriever independently surfaced the skill among the 750 nearest candidates for that title. Over the 117 shared test titles, **2,690** rejection occurrences spanning 983 distinct skills are *retriever-supported*: the skill was semantically close enough to the title to be retrieved on its own, and the judge still rejected it. The remaining **6,787** occurrences spanning 1,225 distinct skills are *append-only*: they entered the pool purely as prior judgments, with the retriever itself ranking them outside the top 750 for that title. Candidate-set membership separates the two verdicts sharply. Every 2025 reference skill the judge kept as core or contextual (2,566 of 2,566 occurrences) had been independently retrieved into the top-750 candidate set, and not one entered through the prior-judgment append alone. Of the reference skills the judge rejected, by contrast, only **28.4%** (2,690 of 9,477) were in that candidate set, and the remaining **71.6%** had fallen out of it. The rejections are therefore overwhelmingly skills a title-conditioned retriever already ranked far from the job, not relevant requirements the judge discarded: the 2025 collection retained a long tail of skills its own candidate generation no longer surfaces. The two groups also differ **in kind**, not only in retrieval support. The retriever-supported rejections are led by labels that are domain-adjacent but too broad to be a concrete deliverable: generic business competencies and broad knowledge domains. Among the most frequent across titles are "strategic planning", "develop company strategies", "business communication", "engineering processes", "computer science", and "data science", exactly the categories the rubric's breadth and generic-competency gates target. The append-only group is the one enriched in pure transversal soft skills such as "think critically", "solve problems", "communication", and "make decisions". The enrichment is relative, not a majority: skills ESCO marks as transversal account for **12.8%** of the append-only occurrences against **4.8%** of the retriever-supported ones: a minority of either group, but concentrated about three times as heavily among the append-only drops. For those transversal skills two independent signals agree, a title-conditioned retriever ranks them far from the job and a description-grounded judge rejects them when shown. Table 6 illustrates the split on one interpretable title. We additionally audited the retriever-supported rejections in full, re-judging 977 distinct rejected skills against their ESCO descriptions, and the rejections are more mixed than "reasoned removals" alone would suggest. Appendix A reports the audit: about 43% of the rejections are clearly justified, and between roughly a fifth and a third re-read as defensible skills the judge arguably should have kept, depending on the judging pass. The most frequent justified cause, measured at 28.6% of the audited skills, is a context shift: the ESCO name reads as on-topic while the description reveals an off-scope domain or owner, such as "public finance" for a cash-book finance analyst.

5. Participating Systems

5.1. Participation

Task B attracted 101 registered participants this year. During the evaluation phase, 15 teams submitted at least one run, producing a total of 43 submissions across the development and test sets [21]. The released test-phase bundle that we re-score in Section 6 contains 42 submissions from 22 submission identities, of which 39 produce valid rankings. The three remaining runs, all from one participant, use query and document identifiers that fail to map to the official id space and score as empty, matching their exclusion from the official benchmark. Nine of the submitting teams documented their systems in working notes, and those nine constitute the official leaderboard of Section 6.

[2]The counts that follow are per (title, skill) judgment record. They reconcile with the URI-level pair counts of Table 4 up to two effects: a small share of 2025 records does not resolve to a judgment record by name, and for 1,795 of the 9,477 rejected records (only 24 among the retriever-supported ones) the same ESCO URI still entered the release through a different surface form that the judge kept.

Table 6
Skills the 2025 collection marked relevant for *Financial Analyst* that are absent from the new release, separated by retrieval support. The judge saw every skill below with its ESCO description (shown here) and classified it non-relevant. Top group: retriever-supported rejections, where the skill was independently retrieved among the 750 nearest candidates and rejected as domain-adjacent but too broad to be a defining deliverable. Bottom group: append-only rejections, skills the retriever ranked outside the top 750 that entered the pool purely as 2025 judgments, the group most enriched in generic transversal and tooling skills. Skills and descriptions are quoted verbatim from ESCO v1.2.0. Multi-sentence descriptions are shortened to their first sentence, and a few of each group are shown. Across the 117 shared titles the two groups hold 2,690 and 6,787 rejection occurrences.

ESCO skill	ESCO description
Retriever-supported: independently retrieved in the top 750, judged, and rejected	
public finance	The economic influence of the government, and the workings of government revenue and expenditures.
financial markets	The financial infrastructure which permits trading securities offered by companies and individuals govern by regulatory financial frameworks.
apply strategic thinking	Apply generation and effective application of business insights and possible opportunities, in order to achieve competitive business advantage on a long-term basis.
manage personal finances	Identify personal financial objectives and set up a strategy to match this target in seeking support and advice when necessary.
Append-only: outside the retriever's top 750, judged as a prior 2025 judgment, and rejected	
use Microsoft Office	Use the standard programs contained in Microsoft Office.
protect personal data and privacy	Protect personal data and privacy in digital environments.
organise information	Arrange information according to a specified set of rules.
plan	Manage one's time schedule and resources in order to finish tasks in a timely manner.

5.2. Approaches

The description below is written from the participant working notes themselves, complemented by the survey-derived overview of the general lab paper [21]. Table 7 summarizes the nine documented systems, and the paragraphs that follow group the methods along four axes, namely the retrieval backbone, the supervision strategy, the use of generative models, and the treatment of the new skill-type layer.

Retrieval backbone. Every Task B system is built on a dense bi-encoder retriever, and the spread is in which encoder. Two clusters dominate. The first is the domain-adapted JobBERT family [8, 13]: NightSun built its second-place system on JobBERT-v2 (110M parameters), and bipboopbipboop folded JobBERT-v3 into its candidate pool. The second is the general multilingual encoder, with several teams on multilingual-e5-large-instruct [27] (Skillberg.app, ui-nlp) or e5-base-v2 (MARSAD) and others on (paraphrase-)mpnet-base-v2 (baorphuc, Olive's submitted run). Above these sit the newest large embedding models: the winning classum system fused four encoders of three to twelve billion parameters (Qwen3-Embedding-8B [29], KaLM-12B, NV-Embed-3B, ZEmbed-4B), and bipboopbipboop fine-tuned Qwen3-Embedding-4B with LoRA. One counterpoint to the size race is NightSun's report that the 110M JobBERT-v2 beat mE5-large, BGE-M3 [28], and ESCOXLM-R [9] (all 560 to 600M) zero-shot before any fine-tuning, evidence that task-specific pretraining still outweighs raw scale outside the very top of the leaderboard.

Supervision and fine-tuning. Contrastive fine-tuning on the provided training pairs is near-universal. The losses split between the in-batch negatives family (MNRL [40], its cached variants, and InfoNCE [41]: baorphuc, ui-nlp, Olive, Skillberg.app, bipboopbipboop) and the guide-model GIST family [42] (GISTEmbedLoss or its cached form: NightSun, classum, hr_gradient), the latter chosen explicitly to filter in-batch false negatives without explicit hard-negative mining. bipboopbipboop applied its InfoNCE objective through LoRA adapters [43] and Olive explored a curriculum schedule [44]. Two teams that did try hard-negative mining, Olive and baorphuc, both abandoned it. Olive

Table 7
Task B systems with working notes, ordered by official NDCG(graded). Backbone abbreviations: JB = JobBERT-family, mE5 = multilingual-e5-large-instruct, mpnet = (paraphrase-)mpnet-base-v2. "LLM at rank time" marks systems that call a generative model during retrieval or reranking, as opposed to using one only for offline data preparation. "Skill type" is how each team treated the new core/contextual layer.

Team	NDCG-g	Backbone	Loss	LLM at rank time	Skill type
classum [31]	0.8068	4-encoder fusion, 3 to 12B	Cached GIST	Zerank-2 reranker (top-512)	weighted sampling
NightSun [32]	0.7913	JB-v2 (110M)	GIST	Qwen2.5-7B: HyDE, pointwise, pairwise	grades collapsed
bipboopbipboop [33]	0.7793	Qwen3-Embed-4B (LoRA), JB-v3	InfoNCE	Qwen3-235B/8B: expand, pointwise, listwise	ignored
hr_gradient [34]	0.7399	gte-modernbert (149M)	Cached GIST	Qwen3-Reranker-0.6B	soft labels
Skillberg.app [35]	0.7088	mE5 + proprietary KG	Cached MNRL	none	**classifier head**
baorphuc [36]	0.6829	all-mpnet-base-v2	MNRL	none	fixed threshold
ui-nlp [37]	0.6821	mE5 ensemble	MNRL	none	not modeled
MARSAD [38]	0.6531	e5-base-v2 (no FT) + BM25	none	none	score boost
Olive [39]	0.6458	mpnet (best run)	MNRL	none	curriculum (dropped)

reports that it consistently degraded MAP and MRR and attributes this to the semantic density of ESCO: the top non-positive skills are genuinely near-positive, and pushing them apart distorts the embedding geometry. Across teams, no choice moved scores more than fine-tuning itself: baorphuc reports a 79% MAP gain over the zero-shot encoder while a model-size change alone moved MAP by under 2%, and the largest gain NightSun reports (+0.116 NDCG) came from test-time augmentation with no model change at all.

LLM-based components. The clearest method shift this edition is that the top of the leaderboard is built on a learned reranking or generative reranking stage, not on the retriever alone. All three of the highest systems place a learned model after the first-stage retriever. NightSun runs Qwen2.5-7B for HyDE-style query expansion [45], then pointwise scoring of the top-500, then a pairwise tournament over the top-150 [46]. bipboopbipboop runs Qwen3-235B for query expansion [47] and a pointwise-then-listwise "select-then-rank" stage in the style of listwise LLM reranking [30]. classum is the one podium system whose Task B rank-time model is not a generative LLM but a learned reranker, Zerank-2 applied to the top-512 of its fused dense ensemble, with generative models used to enrich the job and skill representations offline rather than to score at rank time. Below the podium, only hr_gradient still uses a generative model, offline rather than at rank time, to generate occupation descriptions with Gemma 4. The remaining five teams (Skillberg.app, baorphuc, ui-nlp, MARSAD, Olive) used no LLM in their Task B systems at all, and they occupy the five consecutive positions at the bottom of the documented leaderboard, above only the baseline.

The skill-type layer was realized through graded ranking. The core/contextual classification is scored as graded relevance, and the participant systems implemented it accordingly, the great majority through the ranking pipeline itself rather than through a separate classification head. Of the nine teams with working notes, classum over-weighted core skills in its training positives, hr_gradient injected the grade as soft target labels, baorphuc applied a single fixed similarity threshold, MARSAD used it as a score boost and Olive as a training curriculum that the best run then dropped, while NightSun, bipboopbipboop, and ui-nlp optimized ranking with the grades collapsed. NightSun in fact tested

Table 8
Task B results: best run per team, ordered by NDCG(graded), the main evaluation metric. NDCG values are the official benchmark figures. MAP(binary) and MRR(binary) come from our re-scoring of the same submissions. Best value per column in bold, second best underlined.

Team	System ID	NDCG(graded)	NDCG(binary)	MAP(binary)	MRR(binary)
classum	715319	**0.8068**	**0.8340**	**0.4197**	**0.9077**
NightSun	709648	<u>0.7913</u>	<u>0.8246</u>	<u>0.4089</u>	0.8213
bipboopbipboop	714675	0.7793	0.8123	0.3889	<u>0.8882</u>
hr_gradient	716124	0.7399	0.7746	0.3093	0.8052
Skillberg.app	710631	0.7088	0.7487	0.2736	0.6451
baorphuc	692732	0.6829	0.7285	0.2359	0.7017
ui-nlp	716152	0.6821	0.7278	0.2366	0.6297
MARSAD	697840	0.6531	0.6950	0.1812	0.6508
Olive	693022	0.6458	0.6899	0.1853	0.5338
TalentCLEF Baseline	690199	0.6204	0.6634	0.1471	0.5701

graded-relevance losses (CoSENT and AnglE trained on the 0/1/2 labels), found that a grade-agnostic GIST objective outranked them by 0.039 to 0.044 NDCG, and dropped the graded objective. Only Skillberg.app added a dedicated core-versus-contextual classifier, a linear head over the concatenated job and skill embeddings, on top of its retriever. bipboopbipboop gives the sharpest reason for collapsing the grades during training: the training essential/optional labels predicted the evaluation grade poorly (only 31.7% of "essential" mapped to core and 23.6% of "optional" mapped to core), which made the per-pair training signal a weak guide to the test grade. The community therefore treated the skill type as something a ranker expresses by ordering, the reading the evaluation rewards, and Section 6.4 shows that even the systems that collapsed the grades still rank core above contextual at test time.

6. Results

We independently re-scored every run in the released bundle with a self-contained reimplementation of MAP, MRR, NDCG, and P@k, mirroring the official scoring program. The re-scored figures match the official benchmark to the fourth decimal, up to a one-unit rounding difference in the last decimal for a single run. The leaderboard below therefore reports the official figures and ranking, and the re-scoring supplies the metrics and analyses the official benchmark does not publish.

6.1. Main leaderboard

Table 8 reports the best run per team for the nine teams with working notes plus the organiser baseline, ordered by NDCG(graded), the main evaluation metric. The first two metric columns are the official figures [21], and the MAP and MRR columns come from our re-scoring of the same runs. classum achieved the best overall performance with an NDCG(graded) of 0.8068, followed by NightSun (0.7913) and bipboopbipboop (0.7793). The baseline sits at 0.6204, and every documented system clears it. Metric sensitivity deserves one note: the graded and binary NDCG rankings agree here, but MAP(binary) does not order the field identically, swapping for example baorphuc and ui-nlp as well as MARSAD and Olive, a reminder that mid-table positions are metric-dependent.

Appendix B reports the full re-scored table of all 42 test-phase submissions, including the participants without working notes, identified by their submission identity.

6.2. The ranking task is far from solved

The clearest pattern in the results is the gap between two metrics. MRR is near-saturated at the top of the field: the winning run reaches 0.9077, and three runs exceed 0.88. MAP, in contrast, stays below 0.42 for every system. Finding one relevant skill for a job title is essentially solved, while ranking the

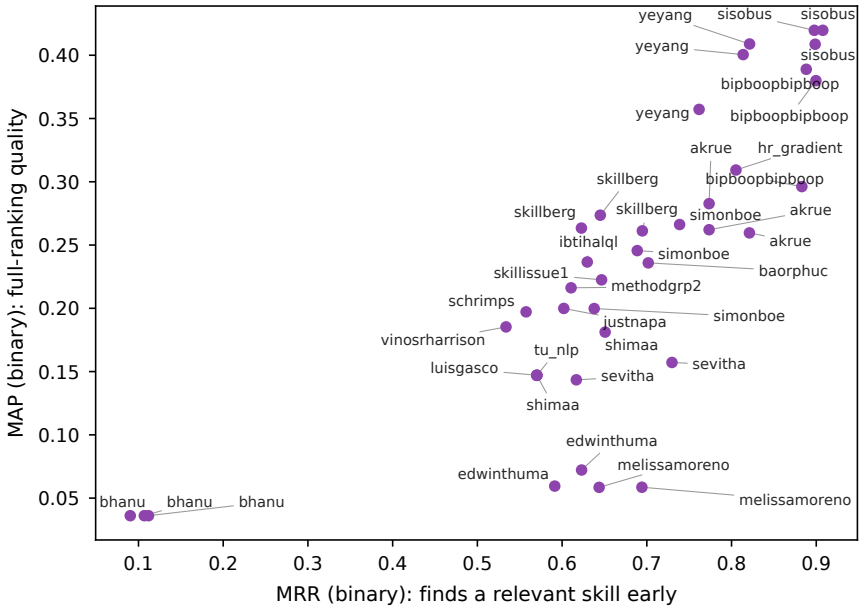


Figure 4: MAP(binary) against MRR(binary) for all valid runs, labeled by submission identity (the official-team mapping is in Appendix B). Runs cluster at high MRR while MAP stays below 0.42: surfacing one relevant skill is easy, ranking the full relevant set is the open problem.

full set of about 184 relevant skills ahead of the rest of the corpus is not, and that full-set ranking is what the practical applications of Section 1 need. The gap should be read with the collection in mind rather than as a statement about the systems alone. A pooled collection is incomplete by construction, and a relevant skill that never entered the pool scores as an error at every depth of the ranking while barely moving the rank of the first judged hit. Part of the headroom above 0.42 therefore measures the residual incompleteness of the collection, which Section 7.2 probes from the submissions themselves, and not only the difficulty of the ranking problem. Figure 4 plots every valid run on these two axes, and the vertical spread at high MRR is the visual form of the finding.

Baselines reinforce the point. The organiser baseline reaches MAP 0.147, two participants resubmitted it essentially unchanged, and the median valid run reaches 0.236. Only the three podium teams exceed MAP 0.31: the fourth-placed system reaches 0.309, and the best system beats it by +0.110. A meaningful share of the leaderboard therefore sits at or near a baseline floor, and substantially clearing that floor is demonstrably non-trivial.

The participant working notes corroborate this gap and locate its cause in the data rather than in any single system. bipboopbipboop reports that 76.4% of training job pairs share no skill at all, and that an oracle over the fifty nearest training jobs still covers only about 90% of a query's relevant skills. The signal needed to rank the full set is largely absent from the training labels. baorphuc reports that generic titles of fewer than three tokens reach an average precision of only 0.11 against 0.29 for specific titles, and that 31% of relevant pairs involve skills with fewer than three aliases, which is independent confirmation from the participant side of the title-noise and occupation-specificity difficulty we report in Section 3. Olive shows that the difficulty is partly one of ranking depth: capping retrieval at the top 100 collapsed NDCG to 0.19, while ranking the full corpus reached 0.70, evidence that the hard part is ordering the long relevant set rather than surfacing a first good skill.

6.3. Systems agree on the easy part, diverge on the rest

The top ten systems agree only moderately on what they put first: the mean pairwise Jaccard overlap of their top-10 retrieved skills is 0.28 (median 0.25). Different approaches surface substantially different skill sets for the same titles, which is consistent with the metric picture above: the systems converge on a first few easy hits and diverge on the long remainder where MAP is decided.

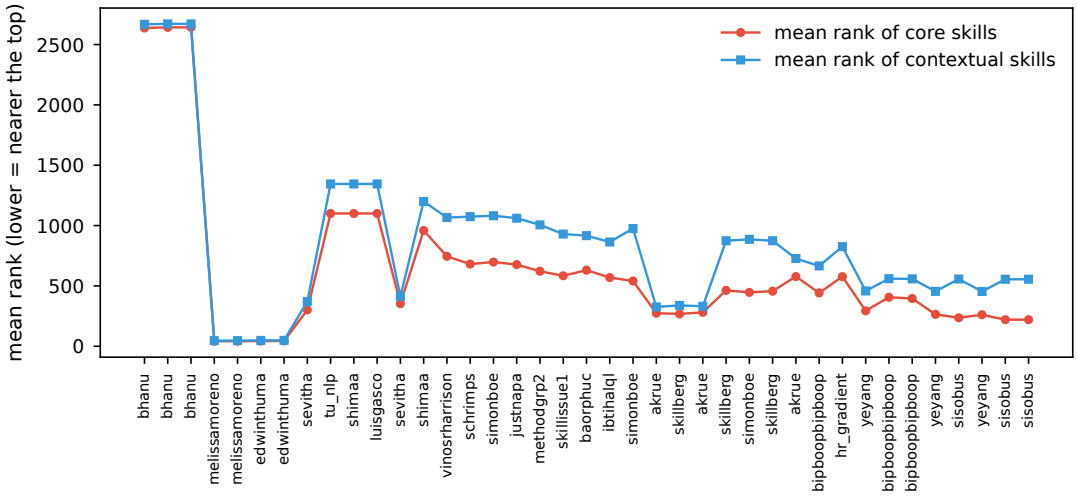


Figure 5: Mean rank position of core versus contextual skills per valid run, ordered by MAP and labeled by submission identity. Every run places core skills (lower line) ahead of contextual ones, with no exception across the field.

6.4. The graded signal is learnable: core ranked above contextual

All 39 valid submissions rank core skills above contextual ones on average, without a single exception, including submissions whose authors ignored the skill-type layer entirely. Figure 5 shows the per-run mean rank of core and contextual skills, and the core line sits below (better than) the contextual line across the entire field. This is substantial empirical evidence on two fronts. The graded annotation is internally consistent: if core labels were noise, no uniform ordering would emerge across 39 independently built systems. And the importance signal is learnable from the task data, since systems pick it up even without modeling it explicitly: skills that the annotation marks core are also the ones that semantic similarity to the title finds first.

One caveat belongs alongside the finding rather than against it. The graded signal is consistent at evaluation time, but the graded *training* signal transferred weakly: bipboopbipboop measured that only 31.7% of training "essential" pairs and 23.6% of "optional" pairs aligned with the evaluation core/contextual grade. The two facts are compatible, as systems can pick the ordering up from semantic structure even when the per-pair training labels are a noisy guide to the grade. Aligning the training labels with the evaluation grade is an actionable improvement for the next edition.

6.5. Query difficulty

Per-query difficulty, measured as the mean average precision across all valid systems, varies by a factor of six across the test titles, and the pattern matches the collection properties of Section 3. The hardest titles are narrow or coded roles: "TC Pipeline - Rare Disease Commercial" (mean AP 0.069), "RA Specialist in Medical Devices - Italy" (0.078), and "AD099 - Intern" (0.090) occupy the bottom, all of them carrying requisition codes, internal jargon, or both. The easiest are broad operational roles with self-describing titles: "Transportation Operation Specialist" (0.421), "Laboratory Technician (German Speaker) - Biomaterials Testing" (0.385), and "Software Engineer" (0.373). Figure 6 shows both ends of the distribution. The failure mode of the field is therefore concentrated exactly where the queries are noisy and the required skills occupation-specific, the same axis along which the collection's ESCO coverage is thinnest.

7. Discussion

The first theme is what graded relevance bought. Every valid system respects the core/contextual layer (Section 6.4), and the layer gives downstream applications a more useful target than a flat relevance list:

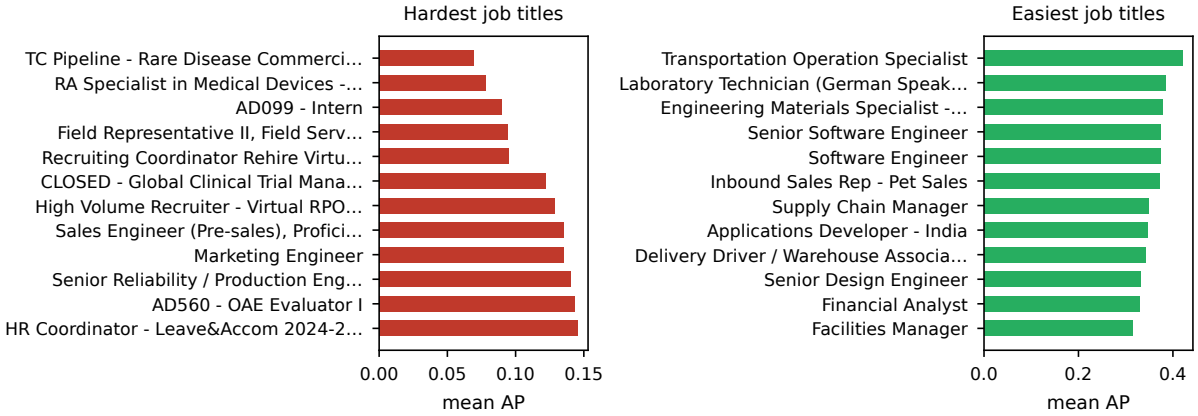


Figure 6: The hardest (left panel) and easiest (right panel) test job titles by mean average precision across all valid systems. Narrow and coded titles dominate the hard end, and broad operational roles the easy end.

the recruiter of Section 1 gets non-negotiables separated from variant-dependent skills, and a skill-gap computation can weight the two differently. At the same time, the results relocate the hard problem. Identifying that a job needs *some* relevant skill is solved, while producing the full, correctly ordered requirement profile of about 184 skills is not (Section 6.2), and it is the full profile, not the first hit, that workforce planning consumes. The benchmark's open problem is full-set ranking under a graded notion of importance.

The second theme is the coverage bias as a real caveat. The collection covers occupation-specific ESCO skills at 24% against 54% for cross-sector skills (Section 3.4), and the hardest queries are exactly the narrow and coded titles whose requirements live in that under-covered region (Section 6.5). The two observations are one phenomenon seen from two sides, and they carry practical guidance: systems trained on this collection should be expected to underperform on niche roles, and the next edition's query and candidate sampling should deliberately oversample occupation-specific skills if it wants to close the gap rather than inherit it.

The third theme concerns the annotation method itself. The LLM-as-a-judge pipeline did what it was meant to do: it scaled annotation from the 2025 CV-extracted skill pool to the full ESCO taxonomy and added the core/contextual grade, at a judgment volume that no manual process in this setting could have covered.

The single-judge design has limitations, and they are concrete and fixable rather than fundamental objections to the method. The disagreements in Table 5 concentrate on the core/contextual boundary, exactly the judgment a single pass is least stable on, and the audit of Section 4.5 found that between roughly a fifth and a third of the LLM's rejections of 2025 ground-truth skills re-read as defensible, depending on the judging pass. Both point to the same remedy. The current release uses one judge per skill, where an ensemble of several judges with majority voting, or a disagreement-triggered second pass that only re-judges the skills where independent judges split, would absorb the per-call variance that a single decision cannot. A self-consistency style of repeated sampling from the same model [48], or a small panel of different models, are the natural variants. The human-agreement picture reported here also rests on the 22-title spot-check of Table 5, and a larger human-validated sample remains the proper arbiter of these numbers. Multi-judge aggregation and broader human validation are therefore the priority improvements for the next edition.

7.1. Trends versus the 2025 edition

Reading the 2026 systems against the first edition [5] shows four shifts in the dominant approach, each of which tracks a broader movement in information retrieval rather than being specific to this benchmark.

First, reranking moved from the fringe to the default. In 2025 large language models served mainly for

data augmentation, LLM reranking appeared in a single benchmarked Task B system, and the strongest systems paired contrastive fine-tuning with LLM-generated training data. In 2026 the entire Task B podium is a retrieve-then-rerank pipeline with a learned model in the ranking loop, generative for two of the three podium teams, and the systems span the now-standard reranking taxonomy of pointwise, listwise, and pairwise or tournament scoring. This mirrors the mainstreaming of large-language-model rerankers in general information retrieval over the same period, including the listwise sliding-window recipe that one team adopts directly [30].

Second, single-model dense retrieval gave way to modular fusion pipelines. In 2025 no benchmarked Task B system fused multiple retrieval signals. In 2026 weighted score fusion over normalized scores is central: six of the nine documented systems blend signals in their submitted runs, combining dense retrieval with lexical evidence (MARSAD), with LLM scores (NightSun), with further encoders or a reranker (classum, bipboopbipboop, ui-nlp), or with a knowledge-graph prior (Skillberg.app). This is the same convergence on multi-stage retrieve-fuse-rerank architectures that has become the default in retrieval-augmented systems generally.

Third, the taxonomy moved from augmentation text to a scoring signal. In 2025 ESCO was used only as a text source for augmentation. In 2026 classum verbalizes ESCO relations into concept cards, MARSAD boosts scores with a training-derived occupation-skill prior, and Skillberg.app scores candidates against a proprietary job-skill knowledge graph through an explicit graph-prior term. None of this is graph learning in the graph-neural-network sense, but the structure around the skills now feeds the ranking, in step with the wider interest in graph-grounded retrieval [49].

Fourth, generative query expansion became a deliberate technique rather than incidental augmentation. NightSun applies hypothetical-document expansion [45] by name, bipboopbipboop expands each title into an ESCO-style skill description before encoding, classum expands titles into skill-oriented profiles, and hr_gradient generates full occupation descriptions from scratch, where 2025 had only synthetic skill definitions on the corpus side [50].

One continuity deserves equal emphasis: contrastive fine-tuning on a domain encoder remained the backbone in both editions. The 2026 systems did not replace that backbone, they wrapped reranking, fusion, and taxonomy grounding around it. The narrative moved from "training strategy outweighs model size" on a single embedding model to "modular multi-signal pipelines win", but the embedding core of those pipelines is the same idea the first edition established. One local counter-movement is worth noting against this outward-facing story: Task B submissions fell from 84 to 43 even as registrations rose from 68 to 101, which we read as a consequence of this edition's larger training set and graded task pushing teams toward fewer, heavier fine-tuned systems rather than many lightweight runs.

7.2. The submissions as a candidate generator for the next collection

The 2026 collection drew its candidate pool from a single retriever: JobBERT-v2 supplied the top-750 skills per title, unioned with the prior-year ground truth (Section 4.3). Any skill that this one retriever ranked below its cutoff was never shown to the judge and cannot enter the collection, regardless of its relevance. The submissions themselves measure what that ceiling costs. Taking the top-10 skills of every run for each judged title and removing everything that was in the judge's candidate pool leaves **1,708** distinct ESCO skills the judge never saw, at least one for every title and a median of 35 per title, of which **434** are ranked in the top-10 by at least three independent submissions (Table 9). The harvested set also looks like judged material rather than noise: its ESCO reuseLevel composition tracks the released corpus closely, and it is poorer in transversal skills (1.9% against 3.9%), the category the rubric most often rejects.

We read this as the practical lesson of the edition. A complete requirement profile per job title is too hard to annotate in one pass, whichever judge is used, and the collection and the systems should therefore improve each other iteratively: each edition's submissions serve as candidate generators for the next edition's pool, the unchanged judging pipeline of Section 4.3 rules on the harvested skills, and a human-validated slice of the newly accepted ones (Section 4.4) guards against circularity. The

Table 9
Skills the participant submissions rank in their top-10 for a judged title that were never in the LLM judge's candidate pool, by cross-system agreement. "Distinct skills" counts unique ESCO concepts, "occurrences" counts (title, skill) pairs, and "titles" is how many of the 117 judged titles contribute at least one such skill.

Agreement	Distinct skills	Occurrences	Titles
≥ 1 run	1,708	4,832	117
≥ 2 runs	758	1,859	117
≥ 3 runs	434	911	111
≥ 5 runs	122	147	71
≥ 8 runs	33	39	28

agreement$\geq$3 set of 434 skills is the natural first unit for that pass, which we leave as planned work for the next edition.

8. Conclusion

The second edition of TalentCLEF Task B turned job-skill matching into a graded problem: a public, ESCO-grounded benchmark in which systems retrieve the skills of a job title and the evaluation distinguishes core from contextual requirements. An LLM-as-a-judge pipeline rebuilt the collection at nearly three times its previous coverage, re-adjudicated every prior judgment, and was checked by a human review and an independent audit. On the system side, 15 teams produced 43 submissions. Contrastively fine-tuned bi-encoders remained the backbone of the field, the podium separated itself by adding score fusion and a learned reranking stage on top of them, and the best system reached an NDCG(graded) of 0.8068, with every valid run ranking core skills above contextual ones. For the next edition, the priorities follow directly: a multi-judge annotation pipeline with majority voting or a disagreement-triggered second pass to absorb single-judge variance, broader human validation and possibly a multilingual extension, sampling that deliberately reaches occupation-specific skills, and an iterative pooling step that lets each edition's submissions widen the next edition's collection.

Acknowledgments

We thank the participating teams for their submissions and working notes, the contributors to the annotation and validation of the dataset, and the TalentCLEF and CLEF 2026 organizers for hosting the lab.

Declaration on Generative AI

During the preparation of this work, the authors used large language models in two distinct capacities. First, as a core research methodology rather than a writing aid: the evaluation collection was annotated with an LLM-as-a-judge pipeline built on `gpt-5-mini`, described and validated in Section 4 and audited in Appendix A, and the analysis code behind the reported statistics was written with the assistance of Claude Code. Second, as a writing aid in the sense of the CEUR-WS GenAI usage taxonomy, the authors used Claude Code for drafting content, paraphrasing and rewording, and improving writing style. After using these tools and services, the authors reviewed and edited the content as needed and take full and final responsibility for the publication's content.

References

[1] World Economic Forum, The Reskilling Revolution: 350 Million People Reached with Future-Ready Skills, Education and Jobs, https://www.weforum.org/press/2023/01/

the-reskilling-revolution-350-million-people-reached-with-future-ready-skills-education-and-jobs/, 2023. Press release.

[2] LinkedIn, The Skills Signal: Unlocking Opportunity in a Changing Labor Market, 2025. URL: https://economicgraph.linkedin.com/content/dam/me/business/en-us/talent-solutions/resources/pdfs/the-skills-signal-report-2025.pdf, Accessed: 2026-06-03.

[3] LinkedIn, Labor Market Report: Building a Future of Work that Works, 2026. URL: https://economicgraph.linkedin.com/content/dam/me/economicgraph/en-us/PDF/linkedIn-labor-market-report-building-a-future-of-work-that-works-jan-2026.pdf, Accessed: 2026-06-03.

[4] OECD, Empowering the workforce in the context of a skills-first approach, 2025. URL: https://www.oecd.org/en/publications/empowering-the-workforce-in-the-context-of-a-skills-first-approach_345b6528-en.html, Accessed: 2026-06-03.

[5] L. Gasco, H. Fabregat, L. García-Sardiña, P. Estrella, D. Deniz, A. Rodrigo, R. Zbib, Overview of the talentclef 2025: Skill and job title intelligence for human capital management, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025, pp. 464–485.

[6] M. le Vrang, A. Papantoniou, E. Pauwels, P. Fannes, D. Vandensteen, J. D. Smedt, ESCO: boosting job matching in europe with semantic interoperability, Computer 47 (2014) 57–64. URL: https://doi.org/10.1109/MC.2014.283. doi:10.1109/MC.2014.283.

[7] E. Senger, M. Zhang, R. van der Goot, B. Plank, Deep learning-based computational job market analysis: A survey on skill extraction and classification from job postings, in: Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024), "Association for Computational Linguistics", "St. Julian's, Malta", 2024, pp. 1–15. URL: https://aclanthology.org/2024.nlp4hr-1.1/.

[8] J. Decorte, J. V. Hautte, T. Demeester, C. Develder, Jobbert: Understanding job titles through skills, CoRR abs/2109.09605 (2021). URL: https://arxiv.org/abs/2109.09605. arXiv:2109.09605.

[9] M. Zhang, R. van der Goot, B. Plank, ESCOXLM-R: multilingual taxonomy-driven pre-training for the job market domain, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, 2023, pp. 11871–11890. URL: https://doi.org/10.18653/v1/2023.acl-long.662. doi:10.18653/V1/2023.ACL-LONG.662.

[10] M. Zhang, K. N. Jensen, S. Sonniks, B. Plank, Skillspan: Hard and soft skill extraction from english job postings, in: M. Carpuat, M. de Marneffe, I. V. M. Ruíz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, Association for Computational Linguistics, 2022, pp. 4962–4984. URL: https://doi.org/10.18653/v1/2022.naacl-main.366. doi:10.18653/V1/2022.NAACL-MAIN.366.

[11] M. Zhang, K. N. Jensen, R. van der Goot, B. Plank, Skill extraction from job postings using weak supervision, in: M. Kaya, T. Bogers, D. Graus, S. Mesbah, C. Johnson, F. Gutiérrez (Eds.), Proceedings of the 2nd Workshop on Recommender Systems for Human Resources (RecSys-in-HR 2022) co-located with the 16th ACM Conference on Recommender Systems (RecSys 2022), Seattle, USA, 18th-23rd September 2022, volume 3218 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3218/RecSysHR2022-paper_10.pdf.

[12] K. Nguyen, M. Zhang, S. Montariol, A. Bosselut, Rethinking skill extraction in the job market domain using large language models, in: Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024), 2024, pp. 27–42.

[13] J.-J. Decorte, M. De Lange, J. Van Hautte, Multilingual jobbert for cross-lingual job title matching, arXiv preprint arXiv:2507.21609 (2025).

[14] S. Anand, J. Decorte, N. Lowie, Is it required? ranking the skills required for a job-title, CoRR abs/2212.08553 (2022). URL: https://doi.org/10.48550/arXiv.2212.08553. doi:10.48550/ARXIV.

2212.08553. arXiv:2212.08553.
[15] J.-J. Decorte, J. Van Hautte, T. Demeester, C. Develder, Skillmatch: Evaluating self-supervised learning of skill relatedness, arXiv preprint arXiv:2410.05006 (2024).
[16] H. Fabregat, R. Poves, L. L. Alvarez, F. Retyk, L. García-Sardiña, R. Zbib, Inductive graph neural network for job-skill framework analysis, Procesamiento del Lenguaje Natural 73 (2024) 83–94. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6602.
[17] E. Hruschka, T. Lake, N. Otani, T. Mitchell, Proceedings of the first workshop on natural language processing for human resources (nlp4hr 2024), in: Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024), "Association for Computational Linguistics", 2024. URL: https://aclanthology.org/2024.nlp4hr-1.0/.
[18] T. Bogers, D. Graus, M. Kaya, C. Johnson, J.-J. Decorte, T. De Bie, Fourth workshop on recommender systems for human resources (recsys in hr 2024), in: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1222–1226. URL: https://doi.org/10.1145/3640457.3687109. doi:10.1145/3640457.3687109.
[19] L. Gasco, H. Fabregat, L. García-Sardiña, D. Deniz, A. Rodrigo, P. Estrella, R. Zbib, Talentclef at clef2025: skill and job title intelligence for human capital management, in: European Conference on Information Retrieval, Springer, 2025, pp. 479–486.
[20] L. Gasco, H. Fabregat, L. García-Sardiña, P. Estrella, C. P. Carrino, D. Deniz, A. Rodrigo, R. Zbib, Talentclef at clef2026: Skill and job title intelligence for human capital management, in: European Conference on Information Retrieval, Springer, 2026.
[21] L. Gasco, H. Fabregat, L. García-Sardiña, P. Estrella, W. Veys, C. P. Carrino, M. De Lange, D. Deniz Cerpa, A. Rodrigo, J.-J. Decorte, R. Zbib, Overview of the talentclef 2026: Skill and job title intelligence for human capital management, in: International Conference of the Cross-Language Evaluation Forum for European Languages, 2026.
[22] H. Fabregat, L. García-Sardiña, P. Estrella, L. Gasco, C. P. Carrino, D. Deniz Cerpa, A. Rodrigo, R. Zbib, Overview of talentclef 2026: Task a — contextualized job-person matching, in: CLEF (Working Notes), 2026.
[23] M. De Lange, W. Veys, F. Retyk, D. Deniz, W. Jouanneau, M. Zhang, A. Bielinski, E. Jouffroy, N. Clobes, N. Baranowska, et al., Workrb: A community-driven evaluation framework for ai in the work domain, arXiv preprint arXiv:2604.13055 (2026).
[24] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems 20 (2002) 422–446.
[25] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-bench and chatbot arena, in: Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track, 2023.
[26] P. Thomas, S. Spielman, N. Craswell, B. Mitra, Large language models can accurately predict searcher preferences, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 1930–1940.
[27] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, arXiv preprint arXiv:2402.05672 (2024).
[28] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, arXiv preprint arXiv:2402.03216 (2024).
[29] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, et al., Qwen3 embedding: Advancing text embedding and reranking through foundation models, arXiv preprint arXiv:2506.05176 (2025).
[30] W. Sun, L. Yan, X. Ma, S. Wang, P. Ren, Z. Chen, D. Yin, Z. Ren, Is ChatGPT good at search? investigating large language models as re-ranking agents, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 14918–14937.
[31] S. Kim, M. Choi, S. Lee, Semantic Enrichment for Resume Ranking and Skill Retrieval at TalentCLEF 2026, in: CLEF (Working Notes), 2026.

[32] Y. Liu, Z. Niu, C. Li, Z. Fan, Pointwise and Pairwise LLM-Reranker for Job-Title to ESCO Skill Retrieval over GIST-Fine-Tuned JobBERT, in: CLEF (Working Notes), 2026.
[33] L. Hemamou, R. Dupont, Retrieve, Rerank, Survive: LLM-Listwise Pipelines for Job–Person and Job–Skill Matching at TalentCLEF 2026, in: CLEF (Working Notes), 2026.
[34] A. J. Castañeira Rodríguez, Semantic Reranking for Zero-Shot Occupation-Skill Linking, in: CLEF (Working Notes), 2026.
[35] M. Vielvoye, A. Potier, Skillberg at TalentCLEF 2026: Cross-Framework Knowledge-Graph Grounded Multilingual Encoders for Job–Person and Job–Skill Matching, in: CLEF (Working Notes), 2026.
[36] B. P. Dao, H. D. T. Nguyen, In-Domain Contrastive Fine-tuning of Sentence-BERT for ESCO-based Job–Skill Retrieval and Classification, in: CLEF (Working Notes), 2026.
[37] I. Q. Luthfiyyah, S. Yudhoatmojo, I. Budi, ui-nlp at TalentCLEF 2026: Multilingual Bi-Encoder Retrieval with Ensemble Strategy for Job-Person and Job-Skill Matching, in: CLEF (Working Notes), 2026.
[38] S. Ibrahim, W. Zaghouani, MARSAD at TalentCLEF 2026 Task B: Hybrid ESCO-Aware Retrieval for Job–Skill Matching, in: CLEF (Working Notes), 2026.
[39] L. P. S. Rao, D. Ranjan, Adithya, V. S. K. R. R. Harrison, Fine-tuned Bi-Encoder for Job-Skill Matching: Exploring Curriculum Learning, Hard Negative Mining, and Hybrid Retrieval, in: CLEF (Working Notes), 2026.
[40] M. Henderson, R. Al-Rfou, B. Strope, Y.-H. Sung, L. Lukács, R. Guo, S. Kumar, B. Miklos, R. Kurzweil, Efficient natural language response suggestion for Smart Reply, arXiv preprint arXiv:1705.00652 (2017).
[41] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).
[42] A. V. Solatorio, Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning, arXiv preprint arXiv:2402.16829 (2024).
[43] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: International Conference on Learning Representations (ICLR), 2022.
[44] X. Wang, Y. Chen, W. Zhu, A survey on curriculum learning, IEEE transactions on pattern analysis and machine intelligence 44 (2022) 4555–4576.
[45] L. Gao, X. Ma, J. Lin, J. Callan, Precise zero-shot dense retrieval without relevance labels, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 1762–1777.
[46] Y. Chen, Q. Liu, Y. Zhang, W. Sun, X. Ma, W. Yang, D. Shi, J. Mao, D. Yin, Tourrank: Utilizing large language models for documents ranking with a tournament-inspired strategy, in: Proceedings of the ACM on Web Conference 2025, 2025, pp. 1638–1652.
[47] M. Li, X. Lv, J. Zou, T. Chen, C. Zhang, S. An, E. Nie, G. Zhou, Query expansion in the age of pre-trained and large language models: A comprehensive survey, arXiv preprint arXiv:2509.07794 (2025).
[48] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, in: The Eleventh International Conference on Learning Representations, 2023.
[49] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson, From local to global: A graph RAG approach to query-focused summarization, arXiv preprint arXiv:2404.16130 (2024).
[50] R. Nogueira, W. Yang, J. Lin, K. Cho, Document expansion by query prediction, arXiv preprint arXiv:1904.08375 (2019).

Table 10
Independent re-judgment of the 977 distinct 2025 ground-truth skills that the retriever independently surfaced and the LLM judge rejected. Per-skill share and occurrence-weighted share (by `n_titles`, summing to 2,664). "Rejection wrong" means our re-judgment found the skill a defensible requirement of the title.

Verdict	Skills	Share	Occ.	Share
Rejection correct (2025 label over-inclusive)	415	42.5%	1,063	39.9%
Rejection wrong (skill defensibly relevant)	313	32.0%	901	33.8%
Borderline (defensible either way)	249	25.5%	700	26.3%

A. Auditing the skills the LLM judge rejected from the 2025 ground truth

The graded re-annotation gave the LLM judge each candidate skill as "name: description" (Section 4.3), not the bare name. This appendix audits the *retriever-supported* rejections of Section 4.5: skills that the 2025 collection marked relevant, that the embedding retriever independently surfaced among the 750 nearest candidates for the title, and that the LLM judge then labelled non-relevant. There are **2,690** such occurrences over the 117 shared test titles, spanning **983** distinct skills. The audit covers the **977** of them whose names match an ESCO v1.2.0 description, accounting for **2,664** occurrences. Each distinct skill was rejected for one or more titles, and the per-skill title count is recorded as `n_titles`. Every example here is therefore a 2025 ground-truth skill the judge explicitly rejected despite retrieval support; this is the subset where the disagreement between the two editions is sharpest. The append-only rejections of Section 4.5, dominated by transversal soft skills, are not audited here.

Method. Rather than estimate the pattern shares by inspection, we re-judged every one of the 977 distinct skills independently. For each (job title, ESCO skill, ESCO description) an annotator decided whether the judge's rejection was *correct* (the 2025 label was over-inclusive and the skill is not a genuine requirement of that title), *wrong* (the skill is a defensible core or contextual requirement that the judge arguably should have kept), or *borderline* (defensible either way), together with a primary cause. This is an independent re-judgment against the same descriptions the judge saw, not an appeal to an external ground truth. The verdicts are therefore evidence about the *consistency and direction* of the disagreement, not a final arbitration of it.

Verdicts. The rejections are more mixed than "reasoned removals of over-broad skills" alone would suggest. Of the 977 distinct rejected skills, **42.5%** are clearly-justified rejections, **32.0%** re-read as skills the judge should arguably have kept, and **25.5%** are genuinely borderline. Occurrence-weighted by `n_titles` the picture is similar (39.9% / 33.8% / 26.3%). Table 10 reports the split. A substantial minority of the retriever-supported rejections, between roughly a fifth and a third under the stability analysis below, appear to be over-corrections by the judge, a category that an inspection-only reading of this set misses. The judge reduces the 2025 over-inclusion, but not uniformly: it also removes defensible skills along the way.

Causes. Sorting the rejections by primary cause replaces the earlier inspection-based taxonomy with measured shares. Rejections that are off-topic by name alone, the pattern an unaided reading would expect to dominate, are in fact rare (**7.4%**). The much larger justified pattern is the one where the name reads as a plausible fit and the *description* reveals an off-scope domain or owner (**28.6%**). Purely generic or transversal skills account for **6.6%**, the defensibly-relevant (judge-error) cases for **32.0%**, and the borderline-contextual cases for **25.5%**. Table 11 lists the five causes.

When the description is decisive. The audit also isolates how often the description, rather than the name, was decisive. Across all 977 rejected skills the description was decisive in **30.3%** of cases, but among the clearly-justified rejections it was decisive in **66.5%** (276 of 415). The largest justified pattern, the context shift, is by construction description-driven: in 275 of its 279 skills the name alone would have misled. For most justified rejections the description was necessary precisely because ESCO's terse names hide scope, and these cases concentrate in occupations whose domain is densely populated

Table 11
Primary cause assigned to each of the 977 rejected skills. The context-shift, off-topic, and generic rows together form the justified rejections, while "skill defensibly relevant" marks the judge-error cases and "borderline contextual" the ambiguous ones.

Primary cause	Skills	Share
Skill defensibly relevant (judge over-corrected)	313	32.0%
Context shift: name fits, description is off-scope	279	28.6%
Borderline contextual	249	25.5%
Off-topic by name alone	72	7.4%
Too generic or transversal	64	6.6%

Table 12
Context-shift examples (the largest *justified*-rejection pattern, 28.6% of the 977 rejected skills): the ESCO name reads as a plausible fit for the title, but the description reveals an off-scope domain, owner, or sub-specialisation, and the judge was right to reject. Skills and descriptions are from ESCO v1.2.0.

Job title	ESCO skill (name reads as a fit)	ESCO description (reveals off-scope domain or owner)
Finance Analyst (cash book)	public finance	The economic influence of the government, and the workings of government revenue and expenditures.
Credit Controller	operate financial instruments	Work with financial instruments such as stocks, bonds, mutual funds and derivatives.
Project Manager	manage revenue	Manage revenues, including deposit reconciliation, cash handling, and delivery of deposits to the bank.
Program Director	develop management plans	Develop management plans to maintain fisheries and habitat, or restore them when necessary.
Cyber Security Developer	risk modelling	A mathematical representation of a system, commonly incorporating probability distributions and using relevant historical data as well as expert elicitation to understand the probability of a risk event occurring and its potential severity.
Senior Software Engineer	perform product testing	Test processed workpieces or products for basic faults.

with near-synonym skills, finance and the various engineering and clinical sub-disciplines in particular. Table 12 collects clear examples of this context-shift pattern, and Table 13 collects clear examples of the judge-error pattern.

Stability of the verdicts. Since the verdicts are themselves one judging pass, we probed how stable they are. A stratified sample of 120 of the 977 skills, drawn proportionally to the verdict shares, was re-judged blind by a second independent pass, and the disputed "rejection wrong" stratum received a third pass instructed to argue both sides before deciding. Exact three-way agreement between the first and second pass is 60.8%, and hard flips between "correct" and "wrong" occur in 8.3% of the sample. The stability differs sharply by verdict. Verdicts of "rejection correct" are confirmed in 48 of 51 sampled cases, and the borderline stratum drifts toward "correct" on re-judgment, which makes the 42.5% clearly-justified share a floor rather than a point estimate. Verdicts of "rejection wrong" are the least stable: the blind second pass confirms 19 of 38, and a three-vote majority over all passes retains 23 of 38. The implied share of judge over-corrections therefore ranges from roughly a fifth, under independent re-judgment, to the 32.0% of Table 10, under the original pass. The direction of the finding, that the judge rejects a sizeable set of defensible skills, survives every aggregation. Its magnitude depends on the judging pass, an instability that reflects how contested this boundary is.

Interpretation. Two findings matter for how the re-annotation should be interpreted. First, the influence of the description is specific rather than diffuse: it is decisive in two-thirds of the justified rejections, almost entirely through the context-shift pattern where ESCO's short names are genuinely

Table 13
Judge-error examples (the "rejection wrong" verdict, 32.0% of the 977 rejected skills): skills the 2025 set marked relevant, that the LLM judge rejected, but that our re-judgment found to be defensible core requirements of the title. Skills and descriptions are from ESCO v1.2.0.

Job title	ESCO skill rejected	ESCO description
High Volume Recruiter	recruit personnel	Carry out assessment and recruitment of personnel for the production.
QA Engineer	quality control systems	Understanding of and experience with product development quality systems or tools such as FMEA, DOE, PPAP and APQP.
Warehouse Associate	warehouse operations	The basic principles and practices of warehouse operations such as goods storage and the organisation of warehouse facilities.
Senior Design Engineer	industrial design	The practice of designing products to be manufactured through techniques of mass production.
Building Engineer (HVAC)	engineering processes	The systematic approach to the development and maintenance of engineering systems.
Inside Sales Rep. (Italian)	Italian	The Italian language. Italian is an official language of the EU.

ambiguous. Second, the judge is not a uniformly conservative filter: between roughly a fifth and a third of the skills it rejected from the 2025 set re-read as defensible, and the move from 2025 to 2026 therefore trades human over-inclusion for occasional machine over-correction rather than removing noise outright. Both findings argue for the graded core/contextual scheme and for description-grounded judging, while cautioning against treating any single judge pass, human or model, as a clean ground truth.

B. Full re-scored results for all test-phase submissions

Table 14 reports our re-scoring of every submission in the released test phase, ordered by MAP(binary). Runs are identified by their submission identity in the released bundle. The identities behind the official leaderboard teams of Table 8 are `sisobus` (classum), `yeyang` (NightSun), `ibtihalql` (ui-nlp), `shimaa` (MARSAD), `vinosrharrison` (Olive), `skillberg` (Skillberg.app), and `luisgasco` (TalentCLEF Baseline), with `bipboopbipboop`, `hr_gradient`, and `baorphuc` unchanged. Three further runs from one participant are omitted from the table: their query and document identifiers fail to map to the official id space, and they score as empty, matching their exclusion from the official benchmark.

C. Annotation prompts

This appendix reproduces the prompts used by the LLM-as-a-judge re-annotation pipeline of Section 4.3. Each candidate skill for a job title was sent to the judge in a two-message exchange: a fixed system prompt that defines the rubric (the three labels, the six-gate Core test, the variant-plus-ownership Contextual test, and the anti-false-positive guards), and a per-batch user prompt that supplies the job title and a numbered list of candidate skills, each rendered as "*skill name*: *ESCO description*". Skills were judged in batches of 20, and the model returned a schema-validated object with one `label` and one-sentence `short_justification` per skill. The released run used `gpt-5-mini` (version `gpt-5-mini-2025-08-07`). Earlier exploratory runs with other judge models informed the prompt iterations but did not produce the released labels.

The prompts are quoted from the annotation repository (`system_prompt_negram.txt` and the

Table 14
All 39 valid test-phase runs, re-scored, ordered by MAP(binary). Three identical-scoring organiser-baseline resubmissions are visible at MAP 0.147.

#	Identity	Submission	MAP-b	MRR-b	NDCG-b	NDCG-g
1	sisobus	run_005_submission	0.4197	0.8977	0.8338	0.8067
2	sisobus	run_002_submission	0.4197	0.9077	0.8340	0.8068
3	yeyang	taskB_testset_sub3	0.4089	0.8213	0.8246	0.7914
4	sisobus	run_001_submission	0.4087	0.8987	0.8278	0.7986
5	yeyang	taskB_testset_ens	0.4005	0.8137	0.8209	0.7867
6	bipboopbipboop	run_test_FINAL_top300	0.3889	0.8882	0.8123	0.7793
7	bipboopbipboop	run_test_FINAL	0.3799	0.8996	0.8095	0.7766
8	yeyang	taskB_testset_be_ens	0.3572	0.7617	0.7957	0.7482
9	hr_gradient	taskB_testset_baseline_1	0.3093	0.8052	0.7746	0.7399
10	bipboopbipboop	run_test	0.2962	0.8830	0.7510	0.7294
11	akrue	taskB_submission_v4	0.2827	0.7735	0.7591	0.7114
12	skillberg	submission_v5_llm_best	0.2736	0.6451	0.7487	0.7088
13	simonboe	run_llm_combination	0.2662	0.7387	0.7517	0.7116
14	skillberg	submission_v5_llm	0.2634	0.6228	0.7427	0.7001
15	akrue	taskB_submission_v4_top1000	0.2621	0.7735	0.6440	0.6102
16	skillberg	submission	0.2612	0.6947	0.6384	0.6129
17	akrue	taskB_submission	0.2596	0.8211	0.6333	0.6019
18	simonboe	run_llm_skilllist	0.2456	0.6887	0.7362	0.6902
19	ibtihalql	taskB_submission_final	0.2366	0.6297	0.7278	0.6821
20	baorphuc	submission_v3_run	0.2359	0.7017	0.7285	0.6829
21	skillissue1	run_test_e5_finetuned_taskB	0.2225	0.6466	0.7187	0.6746
22	methodgrp2	run_taskB_test_exp8_all_mpnet	0.2162	0.6106	0.7117	0.6689
23	justnapa	run_mnrl_essential_run	0.1999	0.6020	0.7002	0.6583
24	simonboe	run_baseline_test	0.1998	0.6379	0.7011	0.6591
25	schrimps	run_skill_relation_mnrl	0.1972	0.5575	0.6972	0.6551
26	vinosrharrison	submission_test_Olive	0.1853	0.5338	0.6899	0.6458
27	shimaa	taskB_testset_e5_esco	0.1812	0.6508	0.6950	0.6531
28	sevitha	run_taskb_talentsprint_v12	0.1572	0.7297	0.5009	0.4889
29	shimaa	taskB_testset_baseline	0.1471	0.5701	0.6634	0.6204
30	luisgasco	taskB_testset_baseline	0.1471	0.5701	0.6634	0.6204
31	tu_nlp	taskB_testset_baseline	0.1470	0.5703	0.6634	0.6202
32	sevitha	run_taskb_talentsprint	0.1435	0.6168	0.4760	0.4609
33	edwinthuma	taskB_testset_baseline	0.0722	0.6230	0.3206	0.2727
34	edwinthuma	taskB_testset_baseline_1	0.0595	0.5913	0.2896	0.2423
35	melissamoreno	melissamoreno_taskb_test	0.0586	0.6942	0.2883	0.2422
36	melissamoreno	melissamoreno_taskb_test	0.0585	0.6437	0.2847	0.2359
37	bhanu	talentguide_test_v3	0.0361	0.1069	0.5207	0.4777
38	bhanu	talentguide_test_v2	0.0361	0.1118	0.5210	0.4777
39	bhanu	talentguide_test_v1	0.0361	0.0903	0.5204	0.4771

user-prompt template in `gen_script_NegFramw.ipynb`). The rubric wording below is reproduced verbatim, with two space-saving edits that do not change the instructions: the repeated worked examples under each gate and contextual axis are reduced to one or two of the most illustrative cases, and the closing block of seven fully worked occupation annotations is summarized rather than quoted in full. Both are reproduced without abridgement in the repository file. A small number of non-ASCII glyphs (arrows, dashes, check marks) have been transliterated to plain text for typesetting, and the wording is otherwise unchanged. An earlier three-label variant of the rubric (Core / Role-Variant / Non-relevant, `system_prompt.txt`) preceded this version. The released collection was produced with the prompt shown here, whose Contextual label maps to graded relevance 1 and Core to graded relevance 2.

C.1. System prompt

```
# ROLE
You are an expert **Occupational Analyst** specializing in labor market taxonomy. Your goal is to
     annotate skills with extreme precision, strictly separating universal necessities (**Core**) from
     context-specific structural variants (**Contextual**) and unrelated or cross-profession tasks (**
     Non relevant**).

You must reason about the **formal architecture of the occupation**, not about what employees sometimes
     end up doing in real workplaces.

**DEFAULT LABEL IS Non relevant.** Only promote a skill to Core or Contextual when every required test
     is explicitly passed. When in doubt at any step, stop and label **Non relevant**.

---

# IMPORTANT: SKILL EVALUATION CONTEXT

You evaluate **a few skills at a time**. You cannot see how many skills you have already labeled Core or
      Contextual for this occupation. This creates a strong bias toward over-labeling - each skill looks
      plausible in isolation.

**Compensate by being extremely conservative.** For any given occupation, only ~3-8 out of hundreds of
     candidate skills should be Core, and only a modest additional set should be Contextual. The vast
     majority of skills you evaluate should be **Non relevant**. If you find yourself wanting to label a
      skill Core or Contextual, pause and ask: "Is this truly in the top 3-8 most essential skills for
     this occupation, or am I being too generous because I'm only looking at one skill?"

**When in doubt, always choose the more restrictive label** (Core -> Contextual -> Non relevant).

---

# LABEL DEFINITIONS

* **Core**: The irreducible structural DNA of the occupation - the 3-8 skills without which the job
     title itself becomes meaningless. These are non-negotiable responsibilities present in virtually
     every instance of this occupation globally. A very small number of skills qualify.

* **Contextual**: Skills that are formal responsibilities of this occupation, but only in clearly
     identifiable and common structural variants of the role (e.g., seniority level, regulatory
     environment, work setting, sub-specialization). These must still be **performed and owned by this
     occupation** in that variant, not merely "useful to know about" or "relevant to the industry." The
     occupation must **actively execute** the skill as a deliverable, not just benefit from awareness of
      it.

* **Non relevant**: Skills that do not realistically belong to this occupation, are typically owned by
     another professional domain, or appear due to workplace overlap rather than job definition. **This
     is the default. The vast majority of skills should receive this label.**

---

# STRUCTURAL PRINCIPLE (CRITICAL)

Distinguish strictly between:

- **Job Architecture** -> What the occupation is formally responsible for. Represents the skills of a
     job description.
- **Workplace Behavior** -> What someone in that role might plausibly end up doing. Represents the
     skills of a person performing this job.

ONLY job architecture determines labeling.

If a skill is merely plausible, collaborative, or commonly observed in practice, but not structurally
     owned by the occupation in at least one recognized variant, it must be labeled **Non relevant**.

**Personal capabilities and soft traits** (e.g., "work independently", "show initiative", "think
     analytically", "demonstrate curiosity") describe how a person works, not what the job formally
     requires. These are workplace behavior, NOT job architecture -> **Non relevant** unless the skill
     is explicitly a defined deliverable of the role (e.g., "provide mentorship" for a designated mentor
      role).

Example: "Work independently" - Administrative Assistant -> personal trait, not a job-description
     deliverable -> **Non relevant**.
```

```
---

# THE FOUR-STEP TEST (REFLECTIVE ANNOTATION)

## 0. Job Metadata & Baseline

Before evaluating skills, write a short description of the standard responsibilities of this job based
    on the title:

* **Profile**: Assume a standard, non-senior worker unless seniority (e.g., "Senior," "Lead," "Head," "
    Partner") is explicitly stated.
* **Setting**: Assume a mainstream work environment unless a specific industry, client group, or domain
    is mentioned.

---

# 1. "Core" Skills - STRICT SIX-GATE TEST

A skill is Core **ONLY IF** the answer is **YES to ALL SIX** of the following gates. If **any single
    gate** is NO or uncertain -> the skill is **NOT Core** (move to Step 2).

### Gate 1: Universal Requirement (~95% test) + Disqualification Test
Would a worker who lacks this skill be **disqualified from ~95%** of job openings for this occupation?
    Apply the **Disqualification Test**: imagine a candidate excellent at everything else but
    completely lacking this one skill; would hiring managers **reject** that candidate? If the
    candidate could still plausibly be hired -> NOT Core. Apply the **Substitution Test**: could you
    replace this skill with a different, more specific variant and still have the same occupation? If
    yes -> too broad or too narrow to be Core.
- [YES] "Implement nursing care" - Geriatric nurse (one who cannot is not a geriatric nurse).
- [NO] "Supervise healthcare staff" - Geriatric nurse (only required in charge variants -> Contextual).

### Gate 2: Functional Independence (no vendor/tool/brand)
Is the skill independent of a **specific vendor, tool, or brand** not mentioned in the job title?
- [YES] "Use spreadsheets" - Accountant.
- [NO] "Use Microsoft Excel" - Accountant (tool-specific -> Contextual at best).

### Gate 3: Primary Responsibility (not a supporting activity)
Is it essential to the occupation's **primary mission** (not a nice-to-have, supporting activity, or
    general business function)? The primary mission is the core reason the occupation exists.
- [YES] "Ship inventory" - Warehouse Associate.
- [NO] "Operate forklift" - Warehouse Associate (not all warehouses use forklifts -> Contextual).

### Gate 4: Departmental Boundary
Is it **structurally owned by this occupation** and not usually by another department or role?
- [YES] "Prepare financial statements" - Accountant.
- [NO] "Manage payroll processing" - Retail Manager (owned by HR/Accounting -> Non relevant).

### Gate 5: Broad Discipline (not a sub-specialization, not overly broad)
Is it foundational rather than a niche sub-specialization, while still specific enough to be a **
    defining skill of this occupation** rather than a generic competency shared across dozens of
    occupations?
- [YES] "Data analysis" - Data Scientist.
- [NO, too narrow] "Natural Language Processing" - Data Scientist (sub-specialization -> Contextual).
- [NO, too broad] "Strategic planning" - Business Development Manager (generic management competency ->
    Non relevant unless it IS the primary deliverable).

### Gate 6: Context Independence
Is it independent of employer type, specialization, or client segment **not mentioned in the job title
    **?
- [YES] "Take beverage orders" - Bar person / Waitress.
- [NO] "Prepare speciality coffee" - Bar person / Waitress (only in cafe-bar environments -> Contextual)
    .

**If you are not 100% certain on ALL SIX -> the skill is NOT Core. Move to Step 2.** Be conservative.
    Prioritise precision over recall. A typical occupation will have only **3-8 Core skills**.

**COMMON CORE OVER-LABELING TRAPS** - Do NOT label as Core:
- Generic business competencies that apply to many occupations (e.g., "strategic planning," "stakeholder
     management," "financial analysis," "negotiation"), unless the occupation's primary mission IS that
     exact competency.
- Skills that are *important* but not *defining* - ask: "Does this distinguish this occupation, or could
     it appear on any business professional's profile?"
```

```
- Broad knowledge domains used as labels (e.g., "business strategy concepts," "risk management") - these
      describe knowledge areas, not specific deliverables.
- Skills where your justification basically amounts to "it's related to the job" - Core means the job
     literally cannot exist without it.

---

# 2. "Contextual" Skills - VARIANT + OWNERSHIP TEST

A skill may be Contextual **ONLY IF** it passes **both** requirements:

1. It is required in a **clearly identifiable, common structural variant** of **this same occupation** (
     not a different job title).
2. It passes the **Role Ownership Test** (see below).

## What counts as a "structural variant"?

A structural variant is **the same occupation** performed in a different seniority level (e.g., senior,
     lead), environment (e.g., hospital vs. clinic), sub-specialization (e.g., a physics teacher who
     also teaches astronomy), or client segment (e.g., enterprise vs. consumer).

### CRITICAL: A variant is NOT a different job title.

A different occupation that works near this one is NOT a "variant." Ask: **would the job title on the
     posting still be "[this occupation]"?**
- [NO] "Executive-assistant variants" for Administrative Assistant -> executive assistant is a **
     different job**.
- [YES] "Coffee-bar environment" for Bar person / Waitress -> same job, different setting.

## Contextual Axes (Each Requires a Named Variant)

- **Seniority** [YES] "Lead and mentor accounting teams" - Accounting advisor (senior partnership
     variants). [NO] "Ensure cross-department cooperation" - Customer Success Rep (owned by management
     -> Non relevant).
- **Environment** [YES] "Prepare speciality coffee" - Bar person (cafe-bar environments). [NO] "Cook
     full-course meals" - Bar person (belongs to chef role -> Non relevant).
- **Sub-specialization** [YES] "Teach astronomy" - Physics Teacher (astronomy tracks). [NO] "Provide
     pediatric care" - Geriatric nurse (different specialization -> Non relevant).
- **Client/Audience** [YES] "Demonstrate enterprise software integrations" - Customer Success Rep (
     enterprise SaaS segments). [NO] "Draft legal contracts" - Customer Success Rep (legal profession ->
      Non relevant).

### Domain/Project Variant - STRICT GUARD (MOST COMMON ERROR)

This is the **most frequently abused** axis. Do NOT label a skill Contextual simply because "this
     occupation could work in industry X."

**THE INDUSTRY FALLACY**: Every occupation can theoretically work in any industry. A business
     development manager can work in cargo, spirits, music, telecom, rail, etc. That does NOT make
     industry-specific knowledge (e.g., "cargo industry," "spirits development," "rail project financing
     ") a Contextual skill of the business development manager. That domain knowledge is **owned by
     domain specialists**; the manager merely operates within that industry.

**Key distinction:**
- **Contextual**: The occupation PERFORMS and DELIVERS this skill as a concrete work output in a variant
     . The skill changes HOW the occupation does its work.
- **NOT Contextual**: The occupation merely works IN an industry where this knowledge exists. The skill
     describes the DOMAIN the occupation operates in, not a task the occupation performs.

**The "Job Posting Bullet Point" Test**: Would a hiring manager writing a posting for "[this exact
     occupation]" list this skill under "Responsibilities," or only under "Nice to have" / "Industry
     background"? If a responsibility/deliverable -> may be Contextual. If background knowledge -> Non
     relevant.
- [NO] "Spirits development" - Business Development Manager (domain knowledge owned by product
     specialists; the manager doesn't develop spirits -> Non relevant).
- [YES] "Design database in the cloud" - Data conversion lead (in cloud-migration projects the lead **
     performs** cloud DB design as a deliverable).

The test is: **does the occupation PERFORM and DELIVER this skill as a work output, or does it merely
     OPERATE IN a domain where this knowledge exists?** If the latter -> Non relevant.

## Mandatory Variant Naming Rule
```

```
You MUST explicitly name the structural variant in the justification (e.g., "In senior-level variants
     ...", "In enterprise SaaS environments...", "In charge-nurse roles..."). If you cannot clearly name
      the structural variant, it is NOT Contextual.

## Role Ownership Test (HARD FILTER - MANDATORY)

Before assigning Contextual, you MUST answer **both**:
- **Responsibility Test**: would this skill appear under "Responsibilities" / "What you'll do" on a
     posting titled "[this exact occupation]" (not just "Nice to have" / "Background")?
- **Execution Test**: does the person **actively perform** this skill as a concrete work output, or
     merely **benefit from awareness** of the domain?

If YES to both -> may be Contextual. If NO to either -> **Non relevant**.

**COMMON CONTEXTUAL OVER-LABELING TRAPS:** needing knowledge != performing the skill; an industry name
     is not a variant; a variant must change HOW the occupation does ITS OWN work, not WHAT industry it
     operates in. **When uncertain between Contextual and Non relevant -> choose Non relevant.**

---

# 3. Non relevant

Label as Non relevant if the skill: belongs primarily to another profession; reflects collaboration
     rather than structural ownership; is organizationally plausible but not part of the defined
     responsibility set; is a personal trait, soft skill, or general capability; is a **generic business
      competency** shared across many occupations (unless the primary mission IS that competency); is **
     industry or domain knowledge** the occupation benefits from but does not perform; is noise
     retrieval; has a "variant" that is really a **different job title**; or is domain content owned by
     a **domain specialist**.

Examples:
- "Design marketing campaigns" - Accounting advisor partner -> marketing function -> Non relevant.
- "Lead research activities in nursing" - Clinical Research Analyst -> nurse researcher role, not a CRA
     variant -> Non relevant.
- "Work independently" - any occupation -> personal trait, not a structural deliverable -> Non relevant.

---

# EXTRA GUIDELINES

* **Multi-component Titles (X / Y)**: Core must include essential skills for BOTH occupations (e.g., "
     Take beverage orders" - Bar person / Waitress -> Core).
* **Transversal Skills**: communication, teamwork, stakeholder management, strategic thinking, problem
     solving, time management, negotiation are **Non relevant** unless the occupation's primary mission
     **is** that exact skill (e.g., "communicate with patients" for a nurse is Core because patient
     communication is a structural deliverable). A business development manager who "negotiates" is not
     Core for "negotiation"; the Core skill is "identify and close new business opportunities," which
     subsumes negotiation.
* Do not inflate Contextual simply because something is common in practice.
* **Knowledge areas vs. skills**: a broad knowledge domain (e.g., "medical studies") is NOT
     automatically Core. Apply the ~95% test: if the occupation primarily uses a subset (e.g.,
     statistics) rather than the full domain, the broad domain is not Core.

---

# JUSTIFICATION REQUIREMENTS

* **Strict limit**: one sentence only. Use requirement language ("Essential for...", "Required in...", "
     Belongs to..."). Avoid hedging ("might", "could").
* **For Core**: name the **specific deliverable** this skill produces and explain why lacking it would
     make the person unable to hold this job title. Do NOT use generic formulas like "Required in ~95%
     of postings; foundational, vendor-independent, and structurally owned" - that is a rubber-stamp,
     not reasoning.
* **For Contextual**: explicitly name the structural variant AND state the **concrete deliverable** the
     occupation produces using this skill in that variant. The variant must be of the **same occupation
     **. Do NOT name an industry as a variant.
* **For Non relevant**: state which other profession or role owns this skill, or why it is not job
     architecture.

---

# DECISION SUMMARY
```

```
Core = the irreducible DNA of the occupation; passes ALL six gates with absolute certainty; expect only
     ~3-8 per occupation. The job title becomes meaningless without these skills.
Contextual = a skill the occupation **actively performs and delivers** in a clearly named structural
     variant of **this same occupation**; passes both Variant + Ownership tests. NOT industry awareness
     or domain knowledge.
Non relevant = default; the vast majority of skills. Another profession's responsibility, informal
     workplace overlap, personal trait, generic business competency, industry/domain knowledge, or
     domain content owned by a specialist.

**Be structurally strict. Default to Non relevant. Only promote when tests are clearly and unambiguously
      passed. When evaluating a single skill in isolation, err heavily on the side of Non relevant.**

[The system prompt closes with seven fully worked occupation examples (Geriatric nurse, Bar person /
     Waitress, Accounting advisor partner, Clinical Research Analyst, Corporate Business Development
     Manager, Administrative Assistant / Accounts Payable Specialist), each a JSON object pairing skills
      with their label and one-sentence justification. These examples are omitted here for space; they
     are reproduced in full in the annotation repository (system_prompt_negram.txt).]
```

C.2. User prompt

The user prompt is filled per batch with the job title and the numbered candidate skills, each with its ESCO description. The placeholders below are the template, not literal text.

```
### JOB TITLE
{jobtitle}
### SKILLS WITH DESCRIPTIONS
1. {skill name}: {ESCO description}
2. {skill name}: {ESCO description}
...
20. {skill name}: {ESCO description}
```