

What PISA measures and what it misses: A two-stage LLM-based alignment of IT workforce skills with educational proficiency

Andreea-Maria Tanasă¹, Oprea Simona-Vasilica¹ and Adela Bâra¹

¹ Department of Economic Informatics and Cybernetics, Bucharest University of Economic Studies, 010374 Bucharest, Romania, Corresponding author: simona.oprea@csie.ase.ro

Abstract. Aligning information technology (IT) workforce demands with educational assessments is essential for bridging skills gaps, yet no prior corpus maps IT task reasoning to Programme for International Student Assessment (PISA) proficiency levels. This paper introduces an LLM-powered *framework* aligning IT competencies with PISA 2022 and the OECD Learning Compass 2030. Drawing on O*NET v30.2, ESCO v1.2.1, PISA descriptors and OECD definitions, the proposed pipeline filters 93 core IT occupations. Using *Gemini 2.5 Flash*, 562 tasks are annotated with minimum PISA levels across Mathematical, Reading and Science literacy (1st stage). Validity is established through re-annotation with *Gemini 2.5 Pro* (2nd stage-measured via Cohen’s Kappa) and OLS-based calibration against O*NET ability ratings, showing statistically significant correlations to standardized ability scores. Validated tasks are embedded and clustered into 25 technical profiles via K-Means, each classified against OECD dimensions. The framework is extended to 95 ESCO transversal skills in 24 clusters. Bidirectional analysis reveals that while every PISA proficiency level is engaged by at least one transversal cluster, 33% of these clusters, covering creative, ethical, social-emotional and dispositional competencies, fall entirely outside PISA’s cognitive scope. This boundary mapping identifies where PISA-based alignment is valid and where complementary tools are required for full readiness assessment.

Keywords: Large language models, OECD Learning Compass, O*NET, PISA, Transversal skills.

1. Introduction

The relationship between educational systems and labor market requirements has long been a central concern of education policy and workforce development research. Educational assessments such as the PISA measure what students can demonstrably do at the end of compulsory schooling, whilst occupational frameworks such as Occupational Information Network (O*NET) document what employees must do across thousands of job titles. Neither, on its own, answers the question that matters most to education policymakers, curriculum designers and labor economists: *at what cognitive level must a graduate actually perform to meet the minimum demands of a given occupation, and how does that threshold compare to what current student populations have measurably achieved?*

This gap is particularly acute in the IT sector as IT occupations are among the fastest-growing and highest-compensating categories in OECD labor markets and simultaneously among the most cognitively demanding in terms of the breadth of reasoning they require. Software developers, data scientists, security analysts and network architects are not simply required to recall technical syntax or apply pre-learned procedures; they routinely formulate novel solutions, interpret ambiguous data, reason inductively from empirical signals and communicate complex findings to non-specialist stakeholders. These demands span the full range of the PISA literacy domains (Mathematics, Reading and Science) and, in many cases, exceed the proficiency levels that the majority of students in middle- and lower-performing OECD countries can currently demonstrate. Understanding this gap quantitatively is a prerequisite for any evidence-based response, whether in the form of curriculum reform, targeted skill development programs or education-to-employment transition policy. Empirical evidence confirms this urgency: a bachelor’s degree is no longer sufficient in routine-intensive IT industries, where graduate-level cognitive capacity is increasingly required, underscoring the inadequacy of credential proxies and motivating direct cognitive task-demand measurement [1].

No prior study has annotated occupational task statements with PISA proficiency level descriptors. Existing approaches either apply Bloom’s Revised Taxonomy [2] to occupational complexity classification, describing intended learning outcomes rather than demonstrated population-level achievement or aggregate O*NET ability ratings at the occupation level, which suppresses the intra-occupation cognitive variation that is central to identifying where the demand-supply mismatch actually occurs. Only an assessment-anchored framework such as PISA can support direct comparison between measured student capability and workforce cognitive demand, and no prior work has operationalized this alignment at the task level. The significance of this work extends across several fields:

(a) For education policy, the framework provides a principled, evidence-based method for quantifying the cognitive distance between measured student populations and occupational entry requirements, expressed in the PISA proficiency levels that governments use to set national benchmarks;

(b) For curriculum design, it provides a granular map of where reasoning demands concentrate across IT occupations;

(c) For AI and educational technology research, the paper contributes a validated methodology for deploying LLMs as annotation instruments for educational framework alignment, including an explicit cross-model reliability analysis that characterizes the limitations of that approach.

Our analysis contributes to the first PISA-aligned IT competency corpus, a reproducible annotation pipeline and a structured map of coverage boundaries relative to the OECD Learning Compass.

2. Literature review

The foundation for analyzing work in terms of its constituent tasks rather than broad occupational categories was established by a task-based model of the labor market that distinguishes between routine and non-routine tasks and between cognitive and manual execution modes [3]. Later formalized into a broader framework [4], this tradition shifted the unit of analysis from the occupation to the task, enabling finer-grained characterization of cognitive demand and its relationship to technological change. Binary cognitive/routine typologies have nonetheless been shown to mask systematic within-category variation in technological impact, with richer multi-dimensional task taxonomies revealing differentiated effects invisible to simpler frameworks [5], a limitation the present paper addresses by deploying three independent PISA literacy dimensions as the classification scheme. This task-level approach has since been extended to questions of automation susceptibility, with estimates suggesting that approximately 47% of US employment was at risk of computerization [6] and more recently to examining how language models specifically alter occupational task demand [7]. Equilibrium models show task displacement concentrates at intermediate cognitive complexity [8], experimental evidence confirms AI tools disproportionately augment below-threshold workers while leaving high-complexity reasoning intact [9] and macro synthesis demonstrates that aggregate AI productivity gains are bounded by the cognitive depth of affected tasks [10], all reinforcing that task-level characterization is the prerequisite for evidence-based workforce projection.

Digital technologies have also been argued to increasingly substitute for cognitive rather than merely physical labor, a shift with direct consequences for the cognitive thresholds that educational systems must help students reach [11]. Two databases operationalize this tradition at scale: O*NET provides ability-rated task descriptions across more than 900 occupational titles, while ESCO organizes occupations, skills and competencies in a multilingual hierarchy, providing explicit cross-referencing between job-specific technical skills and broader transferable competencies. Text-mining analyses of ESCO against Industry 4.0 technology trends reveal persistent coverage gaps, particularly in emerging knowledge domains, supporting the need for cross-database integration of O*NET and ESCO to capture the full breadth of contemporary IT task demands [12].

PISA, administered by the OECD since 2000, assesses the mathematical, reading and science literacy of 15-year-old students across 81 countries, defining literacy not as curriculum mastery but as the capacity to apply reasoning to authentic problems [13]. Each proficiency level is described by explicit descriptors anchored to empirically derived item difficulty. Empirical work applying Bayesian Network modelling to PISA 2018 and 2022 data has further demonstrated that reading literacy functions as a causally upstream predictor of both mathematical and scientific achievement across country cohorts, revealing structural interdependencies between the three literacy domains that have direct implications for how PISA-aligned competency frameworks are designed and interpreted [14]. The 2022 Assessment and Analytical Framework defines these descriptors as anchored to what students at each threshold demonstrably do, not what they are taught, and explicitly incorporates computational thinking within mathematical literacy [15]. Longitudinal evidence demonstrates that cognitive skills as measured by international assessments, rather than years of schooling alone, are the principal driver of long-run economic growth and labor market outcomes [16]. Building on this evidence, [17] estimate global losses from missing basic PISA skills at over \$700 trillion, and the OECD has translated these findings into policy terms by advocating PISA data as a labor market benchmark [18], though neither has operationalized that alignment at the task level. The OECD Learning Compass 2030 [19] extends this framing further, organizing the broader competency landscape for 21st-century work across transformative

competency categories, skill dimensions and knowledge dimensions. This broader framing is warranted by labor-economic evidence that the market increasingly rewards social, dispositional, and higher-order competencies that standardized cognitive assessments are not designed to measure [20], underscoring why the Learning Compass, rather than PISA alone, is required as the full classification target.

The growing cognitive demand for advanced reasoning skills in IT occupations, outpacing what most OECD educational systems currently supply, motivates the use of scalable automated annotation to operationalize this gap at the task level. LLMs have been shown to outperform crowd-sourced annotators on structured text classification tasks, achieving higher agreement with expert gold standards at substantially lower cost [21]. Related work has applied NLP classification directly to ISCO-08 task descriptions for AI exposure estimation at scale, establishing that automated task-level annotation of international occupational frameworks is methodologically accepted and policy-relevant [22]. In educational settings, LLMs have been applied to classify the cognitive level of assessment items and to align curriculum objectives with taxonomy levels, demonstrating that models can perform structured educational classification reliably at scale. Systematic reviews of LLMs in education identify structured framework alignment as a high-potential application but flag hallucination and construct validity as risks requiring explicit validation [23] risks addressed here through cross-model Cohen's Kappa [24] analysis and OLS calibration against O*NET expert ratings. Fine-tuned BERT models struggle with ambiguous skill spans in professional text [25], motivating instruction-following LLMs for the more demanding task of PISA-level cognitive classification. Siamese BERT-network architectures enable efficient semantic similarity computation [26], [27], supporting sentence embedding-based classification of ESCO transversal skills in job advertisements [28] and NLP-based ESCO-to-e-CF alignment that substantially outperforms manual matching [29].

Despite the breadth of research across these fields, a specific and consequential gap remains unaddressed. Occupational task databases such as O*NET provide detailed, ability-rated descriptions of what IT work requires and PISA provides empirically grounded, population-anchored measures of what students can demonstrably do, yet no prior study has connected the two. Thus, no existing work has annotated occupational task statements with PISA proficiency level descriptors, applied those descriptors as a cognitive demand classification scheme or used the OECD Learning Compass 2030 as a classification target for empirically derived IT competency clusters. The result is that the cognitive distance between what IT occupations demand and what educational systems demonstrably produce remains unquantified at the task level, the level at which curriculum reform and workforce policy decisions are ultimately made.

3. Methodology

The methodology operationalizes the theoretical gap identified in the literature review, the absence of a task-level bridge between empirically anchored educational assessments and occupational cognitive demand, through a reproducible, LLM-powered annotation pipeline. The pipeline maps individual IT occupational task statements to minimum PISA 2022 proficiency levels across Mathematics, Reading and Science, then extends this mapping to the broader competency architecture of the OECD Learning Compass 2030. Our approach draws on four publicly available corpora and proceeds through six interconnected steps (as in Figure 1) further described in Table 1:

- (1) data preparation and occupational filtering;
- (2) primary LLM annotation using Gemini 2.5 Flash at zero temperature;
- (3) two-protocol validity assessment combining cross-model Cohen's Kappa agreement and OLS-calibrated alignment against O*NET ability ratings;
- (4) semantic clustering of task embeddings into IT competency profiles;
- (5) OECD dimension classification of the resulting clusters;
- (6) a bidirectional ESCO transversal skills gap analysis designed to delineate the boundary of PISA's cognitive coverage.

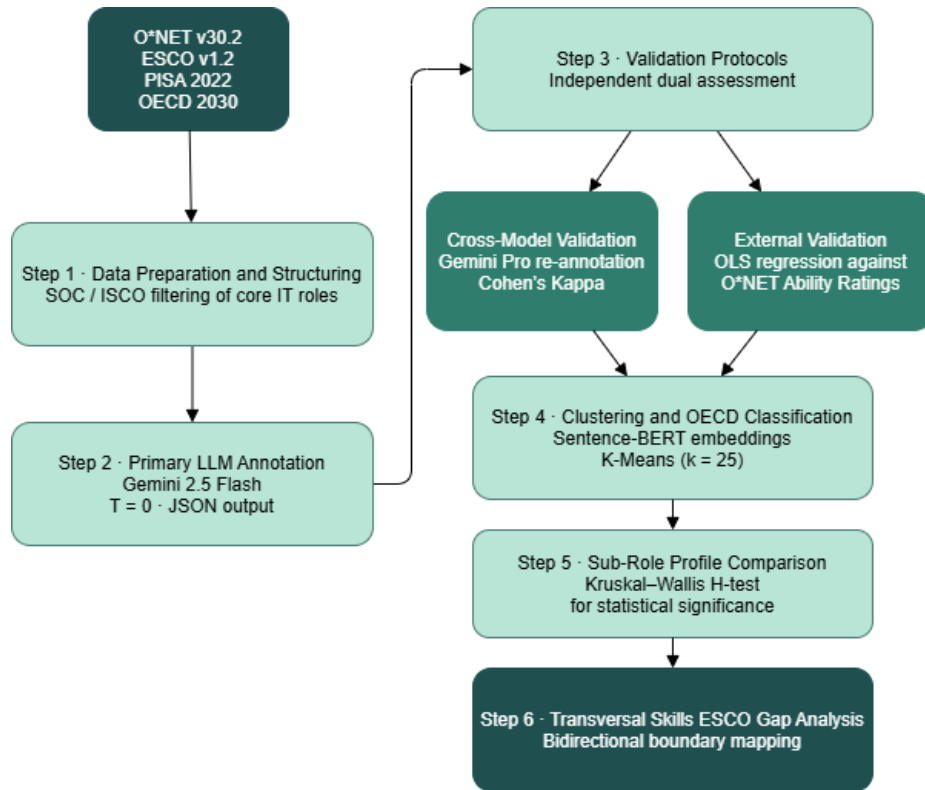


Figure 1. Methodology flow

Table 1. Steps description

Step	Description
Step 1: Data Preparation and Structuring	<p>Five publicly available sources are used throughout:</p> <ul style="list-style-type: none"> (i) O*NET v30.2 (US Department of Labor)-supplies Core IT task statements, occupation titles and Standard Occupational Classification (SOC) codes and ability ratings across 52 dimensions. (ii) ESCO v1.2.1 (European Commission)-provides the European IT occupation hierarchy and the transversal skills branch used in the gap analysis. (iii) PISA 2022 proficiency level descriptors (OECD, 2023)-serve as the annotation target schema, with verbatim text embedded directly in the annotation prompt. The structural asymmetry between domains, Mathematics and Reading literacy include sub-levels 1c, 1b and 1a, while Science literacy begins at 1b, is encoded throughout the pipeline. (iv) OECD Learning Compass 2030 concept notes (OECD, 2019)-supply the nine dimension descriptions used in both the technical cluster and transversal skill classification prompts. <p>The selection of SOC code prefixes is based on the 2018 SOC Major Groups 11 (Management) and 15 (Computer and Mathematical Occupations) [30] to effectively isolate the core IT workforce from the broader O*NET dataset. This filtering strategy specifically targets roles responsible for the design, development, security and management of information systems, as well as data-centric occupations like Data Scientists and Statisticians. Complementing this, the ISCO IT skill profiles (C25, C133 and C35) [31] align with the ISCO-08 classification for Information and Communications Technology Professionals (25), Managers (133) and Technicians (35). By using these hierarchical anchors, the pipeline ensures that the mapping between ESCO-based competencies and PISA reasoning frameworks encompasses the full spectrum of technical depth and managerial responsibility required in the modern IT sector, providing a robust structural basis for the subsequent OLS-based calibration.</p>
Step 2: LLM annotation	<p>The primary annotation model is Gemini 2.5 Flash, accessed via the google-genai SDK with temperature set to 0. All responses are requested as structured JSON to eliminate markdown wrapping. Each task is annotated with the minimum PISA proficiency level required for competent performance, targeting cognitive reasoning demand rather than domain knowledge. The prompt provides the task statement, its occupation title and the verbatim PISA descriptors for all three domains. For each task the model returns a structured JSON object containing:</p> <ul style="list-style-type: none"> (i) One sentence describing the highest complexity cognitive act in that task; (ii) The minimum PISA level per domain; (iii) A one-sentence justification citing the relevant level descriptor; (iv) The primary domain; (v) A confidence rating (high, medium, or low); <p>The prompt used for annotation is provided:</p>

You are a senior researcher in PISA 2022 assessment frameworks specializing in adult professional cognitive demands, not student performance.

TASK STATEMENT:
"{task}"

JOB TITLE CONTEXT:
"{occupation}"

PISA MATHEMATICAL LITERACY LEVEL DESCRIPTORS:
{math_levels}

PISA READING LITERACY LEVEL DESCRIPTORS:
{reading_levels}

PISA SCIENTIFIC LITERACY LEVEL DESCRIPTORS:
{science_levels}

INSTRUCTIONS:

1. Before assigning levels, identify the single most demanding cognitive operation the task requires and state it explicitly in your reasoning.

2. Assign the MINIMUM PISA proficiency level a person must have reached to competently perform this task. A level is the minimum if, and only if, its descriptor fully covers the cognitive operation identified above.

3. Consider ONLY cognitive reasoning complexity — not domain knowledge. Domain knowledge is awareness of specific tools, protocols, terminology, or procedures (e.g., knowing how a firewall works). Cognitive reasoning is the mental process applied regardless of the domain (e.g., integrating ambiguous information, constructing a justification, diagnosing a novel problem).

NOTE:

For Math and Reading, Level 1 is subdivided into 1c (lowest) → 1b → 1a.

For Science, the lowest sub-level is 1b (there is no 1c).

Return ONLY this JSON — no text outside it:

```
{  
  "cognitive_demand_summary": "<one sentence describing the highest complexity cognitive act in this task>",  
  "math_level": "<"1c"|"1b"|"1a"|2/3/4/5/6>",  
  "reading_level": "<"1c"|"1b"|"1a"|2/3/4/5/6>",  
  "science_level": "<"1b"|"1a"|2/3/4/5/6>",  
  "math_reasoning": "<one sentence citing specific language from the level descriptor>",  
  "reading_reasoning": "<one sentence citing specific language from the level descriptor>",  
  "science_reasoning": "<one sentence citing specific language from the level descriptor>",  
  "primary_domain": "<math/reading/science>",  
  "confidence": "<high/medium/low>"  
}
```

Step 3:

LLM validation

Annotation validity is assessed through two independent protocols:

- (i) Expert and cross-model validation. To establish a "gold standard" for the annotation logic, a stratified 50-task sample, balanced across domains and predicted proficiency levels, is subjected to a double-blind validation protocol. First, the sample is independently re-annotated by a human domain expert familiar with the PISA 2022 Framework. The expert remained blind to the LLM's assignments and performed an independent mapping based on the verbatim PISA descriptors. Second, the same 50 tasks are processed by three additional LLM benchmarks of increasing capability: Gemini 2.5 Pro, a higher-capacity model from the same family as the primary annotator (Gemini 2.5 Flash); Claude Haiku 4.5, an externally developed model from a different vendor, serving as a cross-architecture control; and Gemini 3 Flash Preview, a more recent generation model. All models were queried using identical prompts and temperature settings ($T=0$), ensuring that any observed differences in agreement are attributable to model reasoning capacity rather than prompt variability. Agreement between each model and the human expert is quantified using two metrics: (a) linear-weighted Cohen's Kappa (k_w), which measures ordinal agreement while penalizing larger level discrepancies more heavily, and (b) Mean Absolute Error (MAE), which captures the average distance in PISA proficiency steps between model and expert assignments.
 - (ii) External validation via OLS-Calibrated O*NET Ability Mapping. To validate the LLM-generated PISA annotations against an independent, professionally validated reference, we cross-
-

referenced the model's output with O*NET occupational ability ratings. Rather than applying a manual or ad hoc scale conversion, we derived a data-driven linear calibration from the annotation data itself.

1. Each PISA domain was mapped to its closest conceptual equivalents in the O*NET ability taxonomy:
 1. Mathematics literacy to *Mathematical Reasoning, Number Facility*
 2. Reading literacy to *Written Comprehension, Oral Comprehension*
 3. Science literacy to *Inductive Reasoning, Deductive Reasoning, Problem Sensitivity*
2. To map O*NET ability scores (1–7 scale) to PISA proficiency levels (1c–6), we fitted a separate Ordinary Least Squares (OLS) regression model per domain. PISA levels were encoded as a continuous ordinal scale (1c = 1.00, 1b = 1.33, 1a = 1.67, 2 = 2.00, ..., 6 = 6.00) and used as the dependent variable, with the mean O*NET ability score per occupation as the predictor. The model takes the form $P = \beta_0 + \beta_1 \times O$, where P is the predicted PISA ordinal value, O is the mean O*NET Level score for the occupation, and β_0, β_1 are domain-specific coefficients estimated via OLS from the annotated dataset (n = 562).

Step 4: Semantic clustering and OECD Learning Compass classification	Task texts are encoded using <i>paraphrase-multilingual-mpnet-base-v2</i> Sentence-BERT embeddings, these 562 embeddings were then normalized and partitioned into 25 clusters using K-Means clustering (random_state=42) to produce the <i>IT Competency Demand Profile</i> . Each cluster was automatically labelled by identifying the single task whose embedding was closest to the group's centroid, e.g., the most "typical" task of that cluster. Each of the 25 technical clusters is classified against the nine OECD Learning Compass 2030 dimensions in a second LLM annotation pass, using the 12 most representative task texts per cluster as input. The prompt (below) returns a step-by-step reasoning, a primary dimension, up to three applicable dimensions, reasoning and a confidence assessment.
---	---

You are an expert in the OECD Learning Compass 2030 framework.

CLUSTER NAME: {cluster_name}

REPRESENTATIVE TASKS:

{tasks}

OECD LEARNING COMPASS 2030 DIMENSIONS:

{dims}

INSTRUCTIONS:

1. Analyze the core cognitive, practical, and social activities described in the tasks above.
2. Review the OECD dimensions and identify which ones align closest to these activities.
3. Select the SINGLE most dominant (PRIMARY) dimension.
4. Select all relevant dimensions (maximum 3 total, including the primary).

Return ONLY a valid JSON object matching this exact schema:

```
{  
  "analysis": "<Step-by-step reasoning: analyze the tasks and map them to the OECD dimensions. Do this FIRST.>",  
  "primary_dimension_id": "<MUST BE ONE OF: TC01, TC02, TC03, SK01, SK02, SK03, KN01, KN02, KN03>",  
  "all_dimension_ids": ["<id1>", "<id2>"],  
  "confidence": "<high/medium/low>"  
}
```

Step 5: Sub-role PISA profile comparison	<p>To ensure statistical significance and reveal the cognitive signatures of the IT sector, individual SOC occupations were aggregated into seven functional sub-roles based on their primary technical objective: (1) Creative Construction (Software Development), (2) Analytical Modeling (Data & Research), (3) Operational Connectivity (Networks & Infra), (4) Adversarial Risk Management (Security), (5) Organizational Leadership (Management), (6) Evaluative Deconstruction (QA & Testing) and (7) Relational Structuring (Databases). This taxonomy allows for the identification of how PISA-level reasoning demands shift when moving from operational to creative or research-oriented IT functions.</p> <p>To determine whether the identified IT career sub-roles exhibit statistically distinct cognitive profiles, the Kruskal-Wallis H-test was employed. This non-parametric method was selected over a standard one-way ANOVA because the target variables (PISA proficiency levels) are ordinal in nature rather than continuous and their distributions within the IT corpus do not necessarily meet the assumption of normality. The primary purpose of the test is to evaluate whether the samples from the independent sub-role categories originate from the same distribution. A significant H-statistic indicates that the cognitive demand in a specific literacy domain varies across the IT specializations, thereby validating the functional taxonomy used in our study and confirming that the IT sector is not a cognitively monolithic field.</p>
--	---

Step 6: Transversal skills ESCO Gap Analysis

A complementary step examines the ESCO transversal (cross-cutting) skills demanded by IT occupations in order to delineate the boundary of the framework's PISA-based alignment. A bidirectional analysis is conducted: the *forward gap* asks which transversal clusters are covered by PISA proficiency descriptors, while the *inverse gap* asks which PISA proficiency levels are engaged when IT professionals exercise their transversal competencies.

1) Each of the 24 ESCO transversal skill clusters is classified via a structured Gemini prompt at temperature=0. The prompt supplies the cluster name, the full list of skill labels with their ESCO definitions and the nine OECD Learning Compass 2030 dimension descriptions. The model returns: the primary dimension ID, all applicable dimension IDs (up to three), an alignment rationale and a pisa_coverage verdict (full/partial/none) with justification. The prompt is provide:

You are an expert in OECD Learning Compass 2030 and the ESCO transversal skills taxonomy.

ESCO TRANSVERSAL SKILL CATEGORY: {cluster_name}

*SKILLS IN THIS CATEGORY (label: ESCO definition):
{skills}*

*OECD LEARNING COMPASS 2030 DIMENSIONS:
{dims}*

These are ESCO-curated transversal (cross-domain) skills applicable to all workers in any sector.

Task: Classify this category against the OECD Learning Compass 2030.

- Select the PRIMARY dimension (the best single match).
- Select ALL applicable dimensions (1–3 total).
- Judge whether PISA 2022 (Mathematics, Reading or Science Literacy) directly assesses the cognitive demands described by these skills.

Return ONLY this JSON:

```
{  
  "primary_dimension_id": "<TC01/TC02/TC03/SK01/SK02/SK03/KN01/KN02/KN03>",  
  "all_dimension_ids": ["<id1>", "<id2>"],  
  "primary_reasoning": "<two sentences: why this category maps to the primary dimension>",  
  "pisa_coverage": "<full/partial/none>",  
  "pisa_coverage_reasoning": "<one sentence: which PISA domain covers this, or why none>"  
}
```

2) A second prompt reverses the inquiry. For each of the 18 combinations of PISA literacy domain (Mathematics, Reading, Science) and proficiency level (1a through 6), the model evaluates all 24 transversal clusters and determines whether exercising those skills engages the cognitive reasoning demanded by that PISA level (yes/partial/no). The best coverage across all clusters is recorded for each domain×level pair. Levels below 1a are excluded. The prompt is provided:

You are an expert in PISA 2022 assessment frameworks and workforce competency alignment.

CONTEXT: We are analysing whether ESCO transversal (soft) skills require the same cognitive reasoning that PISA proficiency descriptors define. The goal is to identify PISA reasoning demands that are NOT engaged when IT professionals exercise their transversal competencies.

PISA PROFICIENCY LEVEL DESCRIPTOR:

Domain: {domain}

Level: {level}

Descriptor: {descriptor}

ESCO TRANSVERSAL SKILL CATEGORY: {cluster_name}

*SKILLS IN THIS CATEGORY (label: ESCO definition):
{skills}*

QUESTION: When an IT professional exercises the skills in this ESCO category, do they engage in the cognitive reasoning described by this PISA proficiency level?

Answer scale:

- "yes" = exercising these ESCO skills directly requires the reasoning this PISA level describes
 - "partial" = some skills in the category touch on this reasoning, but only partially or at a lower intensity
 - "no" = the reasoning demanded by this PISA level is not meaningfully engaged by these skills
-

Return ONLY this JSON:

```
{
  "covered": "<yes/partial/no>",
  "reasoning": "<one sentence: which skill(s) engage this PISA reasoning, or why none do>"
}
```

4. Results

4.1 Data preparation and LLM annotation

The filtering process isolated the core IT workforce from the broader O*NET 30.2 database. Applying the SOC Major Group 11 and 15 prefixes resulted in a corpus of 28 unique IT-specific occupations and 562 core task statements. This dataset encompasses 28 distinct job titles, ensuring coverage across development, management and data science roles.

From the ESCO v1.2.1 framework, the recursive search of ISCO IT skill profiles C25, C13 and C35 (aligned with ISCO-08 groups) yielded 6618 technical skill relations (Table 2). Additionally, 2996 skills were extracted from the transversal skills (S6.0) to support the subsequent gap analysis.

Table 2. Dataset summary

Metric	Records
IT SOC Codes	28
Core IT Tasks	562
ESCO Tech Skills	6618
ESCO Transversal	29,962,996

The annotation pipeline achieved a 100.0% success rate for structured JSON retrieval from Gemini 2.5 Flash at temperature 0. Analysis of the reasoning fields indicates that 100.0% of annotations explicitly identified the “highest complexity cognitive act” prior to level assignment.

Fidelity to the PISA domain asymmetry was strictly maintained: while Math and Reading tasks utilized the 1c/1b/1a hierarchy, zero instances of Level 1c were assigned to the Science domain (as per descriptor constraints). The model reported “High” confidence in 98.9% of assignments, “Medium” in 1.1% and “Low” in 0.0%, indicating high internal consistency in applying the verbatim descriptors. An illustrative annotation trace is:

Task statement (O*NET): “*Direct daily operations of department, analyzing workflow, establishing priorities, developing standards and setting deadlines.*”

Identified cognitive demand: “*Developing and applying systematic strategies to analyze complex operational workflows, identify constraints, establish optimal priorities, and create effective performance standards.*”

PISA alignment Mathematics:

- Assigned level: 5
- Reasoning: “*The task requires developing and working with models for complex situations, identifying or imposing constraints, and specifying assumptions to analyze workflow, establish priorities, and set deadlines, aligning with Level 5’s description of applying systematic, well-planned problem-solving strategies for challenging tasks like designing an optimal procedure.*”

4.2 LLM validation steps

The results of the cross-model validation on a stratified subset of 50 tasks revealed substantial agreement (as in Table 3) between Gemini 2.5 Flash and Gemini 2.5 Pro for Mathematics literacy ($k = 0.632$) and moderate agreement for Reading ($k=0.490$) and Science literacy ($k=0.577$). Analysis showed that Flash consistently assigned equal or higher PISA levels than Pro, with exact agreement ranging from 44% (Reading literacy) to 58% (Mathematics literacy). Systematic bias was identified when the average gap exceeded ± 0.30 levels, thus a slight systematic overestimation was identified in the Science literacy domain (avg. gap= $+0.36$ levels), suggesting that the efficiency model may marginally inflate scientific reasoning demands. No directional bias was detected for Mathematics or Reading literacy.

Table 3. Cross-model agreement Gemini 2.5 Flash vs. Gemini 2.5 Pro

PISA domain	K(Cohen’s Kappa)	Average gap (levels)	Flash higher	Exact agreement	Flash lower
Mathematics literacy	0.632	+0.14	13 (26%)	29 (58%)	8 (16%)
Reading literacy	0.490	+0.26	21 (42%)	22 (44%)	7 (14%)
Science literacy	0.577	+0.36	19 (38%)	24 (48%)	7 (14%)

While the Cohen’s Kappa scores indicate substantial overall agreement, a granular review of the tasks that diverge by two or more proficiency levels provides critical insight into the nuances of automated task annotation.

In Table 4, the divergence between the models suggests a significant distinction between 1) procedural volume, the quantity of steps and data-handling required; and 2) cognitive complexity, the depth of abstract reasoning demanded by the task. We observe that Gemini 2.5 Flash occasionally assigned higher proficiency levels to tasks characterized by high procedural volume, seemingly mistaking administrative scale and extensive documentation for high-level reasoning. In contrast, Gemini 2.5 Pro demonstrated a more precise alignment with PISA’s core methodology by prioritizing cognitive complexity. It focused on the requirement for mental modeling, hypothesis-testing and inferential reasoning, correctly identifying that a task can be administratively broad yet remain cognitively simple or be procedurally brief while requiring advanced scientific or mathematical logic.

Table 4. Divergence by two or more levels between models

Task	Model	Assigned PISA level	Model reasoning
Monitor Web developments through continuing education, reading or participation in professional conferences, workshops or groups.	Gemini Flash	2	Requires continuous comprehension, integration and evaluation of information from diverse sources to maintain an up-to-date understanding.
	Gemini Pro	1c	Involves identifying and synthesizing basic information from clear professional sources to build a general understanding.
Create or maintain databases of known test defects.	Gemini Flash	4	Systematically organizing and categorizing complex, interrelated information into a structured format for consistency and retrievability.
	Gemini Pro	1a	Core cognitive demand is simply to locate and integrate pieces of information from unstructured text into a pre-existing classification system.
Troubleshoot program and system malfunctions to restore normal functioning.	Gemini Flash	3	Diagnosing malfunctions by executing a systematic strategy, drawing basic inferences to identify root causes.
	Gemini Pro	5	Requires developing a multi-step diagnostic strategy for novel problems by creating a mental model of a complex system and testing hypotheses.

Table 5 reports the aggregate agreement between each model and the human domain expert across the three PISA literacy domains (Mathematical, Reading and Scientific literacy), measured by linear-weighted Cohen’s Kappa, MAE and exact-match accuracy. Table 6. disaggregates these results by domain.

Table 5. Cross-model agreement with human expert aggregate results

Model	Cohen's Kappa (avg)	MAE (avg, steps)	Exact match (avg)
Gemini 2.5 Flash (primary annotator)	0.213	1.433	10.7%
Gemini 2.5 Pro	0.268	1.220	13.3%
Gemini 3 Flash Preview	0.264	1.213	12.7%
Claude Haiku 4.5	0.483	0.667	44.7%

Across all domains, Claude Haiku 4.5 demonstrates the highest alignment with the human expert, achieving a weighted Kappa of 0.483, considered as moderate agreement and a MAE of 0.667 PISA levels, indicating that its classifications deviate from the expert's judgments by less than one proficiency step on average. By contrast, Gemini 2.5 Flash, achieves $k_w = 0.213$ (MAE = 1.433), corresponding to fair agreement, while Gemini 2.5 Pro ($k_w = 0.268$, MAE = 1.220) and Gemini 3 Flash Preview ($k_w = 0.264$, MAE = 1.213) occupy an intermediate band, both also classified as fair.

Table 6. Cross-model agreement with human expert detailed results

Model	Domain	Cohen's Kappa	MAE
-------	--------	---------------	-----

Gemini 2.5 Flash	Math	0.314	1.120
Gemini 2.5 Flash	Reading	0.157	1.440
Gemini 2.5 Flash	Science	0.168	1.740
Gemini 2.5 Pro	Math	0.358	1.020
Gemini 2.5 Pro	Reading	0.197	1.180
Gemini 2.5 Pro	Science	0.248	1.460
Gemini 3 Flash Preview	Math	0.362	1.020
Gemini 3 Flash Preview	Reading	0.189	1.260
Gemini 3 Flash Preview	Science	0.240	1.360
Claude Haiku 4.5	Math	0.632	0.460
Claude Haiku 4.5	Reading	0.432	0.540
Claude Haiku 4.5	Science	0.385	1.000

At the domain level from, Mathematics consistently yields the highest agreement across all models, with Claude Haiku 4.5 achieving $k_w = 0.632$ (substantial agreement) and an exact-match rate of 60%. This suggests that the PISA 2022 Mathematics descriptors are sufficiently operationalized to support reliable zero-shot classification by LLMs. Science proves the most challenging domain: the exact-match rate for Gemini 2.5 Flash collapses to 4.0% ($MAE = 1.740$), pointing to systematic over-estimation of the cognitive complexity of science-adjacent tasks, a bias that persists, albeit attenuated, across all Gemini-family models.

Despite the moderate-to-fair agreement levels observed, the validation reveals that the direction of disagreement is systematic rather than random: all Gemini-family models tend to assign higher PISA levels than the human expert, producing a conservative overestimation of cognitive demands. The consistent rank ordering across models further confirms that LLM-based annotation is a viable, scalable methodology whose precision improves predictably with model capability.

OLS regression was fitted separately for each domain ($n=562$ tasks). The estimated coefficients were: Mathematics ($\beta_0=1.416$, $\beta_1=0.879$), Reading ($\beta_0=-0.839$, $\beta_1=1.260$) and Science ($\beta_0=-3.087$, $\beta_1=1.917$). The increasing slopes across domains suggest that scientific reasoning demands are most sensitive to occupational ability level (small differences in O*NET ability ratings translate into large differences in PISA reasoning level requirements), followed by reading comprehension and mathematical reasoning.

- i. Mathematics: $PISA = 1.416 + 0.879 \times O*NET_score$

The positive intercept of 1.416 indicates that even the least demanding IT tasks require a baseline level of mathematical reasoning, while the slope of 0.879 reflects a steady, proportional increase in PISA level as occupational complexity grows. This gradual scaling suggests that mathematical demands are broadly distributed across all seniority levels in the IT sector, from entry-level support roles to advanced engineering positions.

- ii. Reading: $PISA = -0.839 + 1.260 \times O*NET_score$

The negative intercept introduces a threshold effect: reading demands do not meaningfully differentiate PISA levels until O*NET ability scores exceed approximately 2.0. Beyond this threshold, however, reading demands escalate more steeply than mathematics, as indicated by the higher slope of 1.260. This pattern is consistent with the nature of IT reading tasks, where the transition from basic technical documentation to complex analytical synthesis is abrupt rather than gradual.

- iii. Science: $PISA = -3.087 + 1.917 \times O*NET_score$

The large negative intercept combined with a slope of 1.917, more than double that of Mathematics, indicates that low O*NET scores in inductive and deductive reasoning correspond to minimal scientific reasoning demands, while moderate to high scores translate into rapidly escalating PISA proficiency requirements. This steep sensitivity confirms that scientific reasoning, encompassing hypothesis testing, causal inference and problem diagnosis, is the most discriminating cognitive dimension in IT occupations. Small differences in occupational ability ratings in this domain produce significantly larger changes in the corresponding PISA proficiency level than in any other domain.

The results indicate a hierarchy of cognitive sensitivity across domains, where Science literacy is the strongest differentiator of advanced IT competency. This ordering has direct implications for curriculum design and workforce policy, as it suggests that interventions targeting scientific reasoning will

have the greatest impact on aligning educational outcomes with the proficiency demands of the modern IT labor market.

4.3 Validity results

To assess the validity of the LLM-generated PISA annotations, each task's assigned proficiency level was compared against the independently derived O*NET-calibrated benchmark using two statistics: 1) the Pearson correlation coefficient, where a value of $p < 0.05$ indicates a statistically significant association; and 2) the mean absolute deviation (MAD), where a value below 1.0 indicates that the model's predictions are, on average, correct within one PISA proficiency level.

The results in Table 7 prove that all three domains achieved statistical significance, confirming that the agreement between the LLM annotations and the O*NET benchmarks is not attributable to chance. Mathematics literacy and Science literacy both achieved MAD values below 1.0, indicating sub-level accuracy. Reading literacy marginally exceeded the threshold ($MAD = 1.015$), though it produced the strongest correlation of all three domains ($r = 0.372$); Science demonstrated the most precise predictions overall.

Table 7. Convergent validity results

Domain	Relevant samples	Pearson correlation coefficient (r)	p	MAD
Mathematics literacy	157	0.234	0.0032	0.870
Reading literacy	155	0.372	0.0000	1.015
Scientific literacy	250	0.239	0.0001	0.756

These results support the conclusion that the OLS-calibrated annotation pipeline demonstrates adequate convergent validity across all three PISA literacy domains, with the O*NET-derived benchmarks serving as an independent, data-driven reference point broadly consistent with the LLM-generated proficiency assignments.

4.4 Clusters and building the IT competency demand profile

Silhouette analysis over $k \in [10, 40]$ yielded scores in a narrow range (0.062–0.082), with no dominant peak, a pattern consistent with the high semantic overlap inherent in natural language task descriptions within a single occupational sector. The elbow curve showed gradual flattening beyond $k \approx 24$. Given the statistical equivalence of solutions in this range, $k = 25$ was selected to approximate the number of distinct IT occupation groups in the O*NET SOC taxonomy, favoring domain interpretability. The semantic coherence of the resulting clusters was confirmed through LLM-based thematic labelling, which produced distinct, non-overlapping descriptions for all 25 clusters.

Table 8 shows the resulted clusters that represent the full spectrum of cognitive work in IT occupations. The analysis of the 562 annotated IT tasks reveals that 20 of the 25 clusters (83.8% of tasks) attained median PISA scores at or above level 4 across all three literacy domains simultaneously. The highest-demand clusters (levels 5–6 across all three domains) were research-oriented tasks (Cluster 19) and advanced network/system development (Clusters 7, 8, 23, 16), reflecting the cognitive ceiling of IT work. Only 5 clusters, representing approximately 91 tasks (16.2%), fell below level 4 in at least one domain, corresponding to more routinised operational activities, positioning PISA level 4 as the effective minimum cognitive floor for IT work, with the modal demand concentrated at level 5.

Table 8. Clusters resulted based on task descriptions

Cluster_id	Cluster_label
0	Provide feedback to designers and other colleagues regarding game design features.
1	Provide staff and users with assistance solving computer-related problems, such as malfunctions and program problems.
2	Test system modifications to prepare for implementation.
3	Collaborate with development teams to discuss, analyze or resolve usability issues.
4	Assist in the assessment, acquisition or deployment of new electronic document management systems.
5	Recommend changes to improve systems and network configurations and determine hardware or software requirements related to such changes.
6	Keep abreast of changes in industry practices and emerging telecommunications technology by reviewing current literature, talking with colleagues, participating in educational programs, attending meetings or workshops or participating in professional organizations or conferences.
7	Develop and implement solutions for network problems.

8	Evaluate the statistical methods and procedures used to obtain data to ensure validity, applicability, efficiency and accuracy.
9	Select programming languages, design tools or applications.
10	Analyze and report computer network security breaches or attempted breaches.
11	Consult with users, administrators, and engineers to identify business and technical requirements for proposed system modifications or technology purchases.
12	Develop or implement procedures for ongoing Web site revision.
13	Design databases to support business applications, ensuring system scalability, security, performance and reliability.
14	Synthesize current business intelligence or trend data to support recommendations for action.
15	Review project plans to plan and coordinate project activity.
16	Develop, document or maintain standards, best practices or system usage procedures.
17	Develop information security standards and best practices.
18	Design, configure and test computer hardware, networking software and operating system software.
19	Apply research or simulation results to extend biological theory or recommend new research projects.
20	Perform data backups and disaster recovery operations.
21	Develop procedures to track, project or report network availability, reliability, capacity or utilization.
22	Design or prepare graphic representations of Geographic Information Systems (GIS) data, using GIS hardware or software applications.
23	Analyze information processing or computation needs and plan and design computer systems, using techniques such as structured analysis, data modeling and information engineering.
24	Develop, implement, or evaluate health information technology applications, tools, processes or structures to assist nurses with data management.

For the OECD Learning Compass 2030 classification, the unit of analysis was the task cluster rather than the individual task. Each of the 25 clusters was represented to the LLM by a sample of its constituent tasks drawn to preserve the most canonical, high-frequency examples within each group. The sample size was set to 12 tasks per cluster, approximately 53% of the mean cluster size of 22.5 tasks; this sample should provide sufficient coverage, exceeding the majority of each cluster's composition, for reliable identification, while maintaining a well-structured prompt that could be consistently parsed across all clusters.

All 25 clusters were classified with high confidence. The primary dimension distribution is as follows:

1. TC01 (Creating new value) is the dominant dimension in 16 of 25 clusters (64%) as in Table 9, encompassing programming, system design, network problem-solving, web development, database architecture and continuous professional learning, confirming that the core cognitive demand of IT work is adaptive problem-solving, the ability to generate novel solutions rather than merely apply established procedures.
2. SK01 (Cognitive and meta-cognitive skills) is primary in 5 clusters (20%), covering statistical method evaluation, business intelligence synthesis, project planning, network availability monitoring and system configuration recommendation. In these clusters, the defining activity is structured critical thinking and self-regulated analytical reasoning rather than open-ended creation.
3. TC02 (Reconciling tensions and dilemmas) is primary in 2 clusters (8%), covering stakeholder requirements consultation and usability collaboration with development teams. The defining cognitive demand here is the navigation of conflicting requirements between users, engineers, and management, a fundamentally relational and negotiation-oriented competency.
4. TC03 (Taking responsibility) is primary in 1 cluster (4%), the information security standards cluster, where the ethical and accountability dimension of protecting systems and data constitutes the fundamental driver of all task activity.
5. SK03 (Practical and physical skills) is primary in 1 cluster (4%), covering GIS data visualization and mapping tasks, where technical precision in tool operation is the defining demand.

Table 9. OECD Learning Compass classification and primary dimension per cluster

Code	Category	Dimension	OECD dimension dominance
TC01	Transversal Competency	Creating new value	64%

SK01	Skills	Cognitive and meta-cognitive skills	20%
TC02	Transversal Competency	Reconciling tensions and dilemmas	8%
TC03	Transversal Competency	Taking responsibility	4%
SK03	Skills	Practical and physical skills	4%
SK02	Skills	Social and emotional skills	not primary for any cluster
KN01	Knowledge	Disciplinary knowledge	not primary for any cluster
KN02	Knowledge	Interdisciplinary knowledge	not primary for any cluster
KN03	Knowledge	Epistemic knowledge	not primary for any cluster

None of the four Knowledge dimensions (KN01–KN03) emerged as the primary driver of any cluster, signaling that the core cognitive demand of IT work is not the possession or recall of disciplinary, interdisciplinary or epistemic knowledge, but the application of that knowledge in creative and adaptive problem-solving contexts, precisely the type of reasoning that PISA Levels 4–6 are designed to assess.

As seen in Figure 2, which maps all OECD dimension appearances across the 25 clusters, including primary, secondary and tertiary assignments, SK01 (Cognitive and meta-cognitive skills) appeared as a secondary or tertiary dimension in all 25 clusters, confirming that analytical thinking is a non-negotiable baseline across all IT task types, regardless of specialization. TC01 (Creating new value) follows with 18 appearances and Knowledge dimensions (KN01–KN03) appear only sparingly and never as a primary driver, which reveals that IT work consistently demands the application of thinking skills over the mere possession of knowledge.

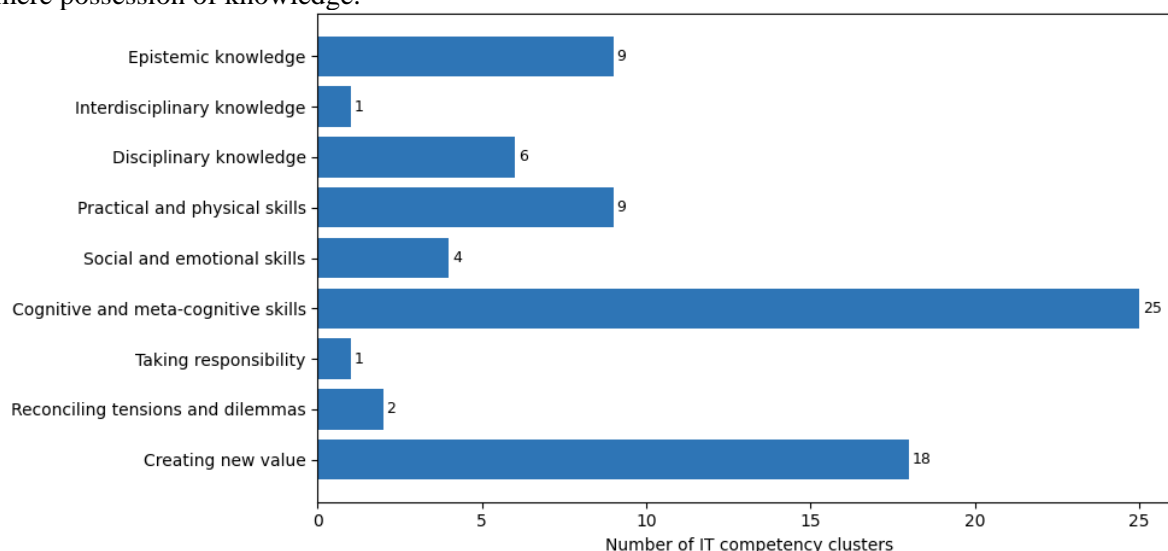


Figure 2. OECD Learning Compass 2030 coverage in IT competency demand profile

As seen in Table 10 and Figure 3, every single IT cluster requires at least PISA Level 3 across all three domains, there are no low-complexity outliers. Reading and Science both have a median of Level 5, whereas Science is the only domain where 100% of clusters require Level 4 or above. This confirms that Science and analytical reasoning are non-negotiable baseline requirements across all IT occupations, regardless of specialization.

Table 10. Summary statistics IT competency demand profile

Domain	Mean	Median	Std Dev	Level 3+	Level 4+
Mathematical Literacy	4.34	4.0	0.76	100%	80%
Reading Literacy	4.74	5.0	0.54	100%	96%
Scientific Literacy	4.80	5.0	0.50	100%	100%

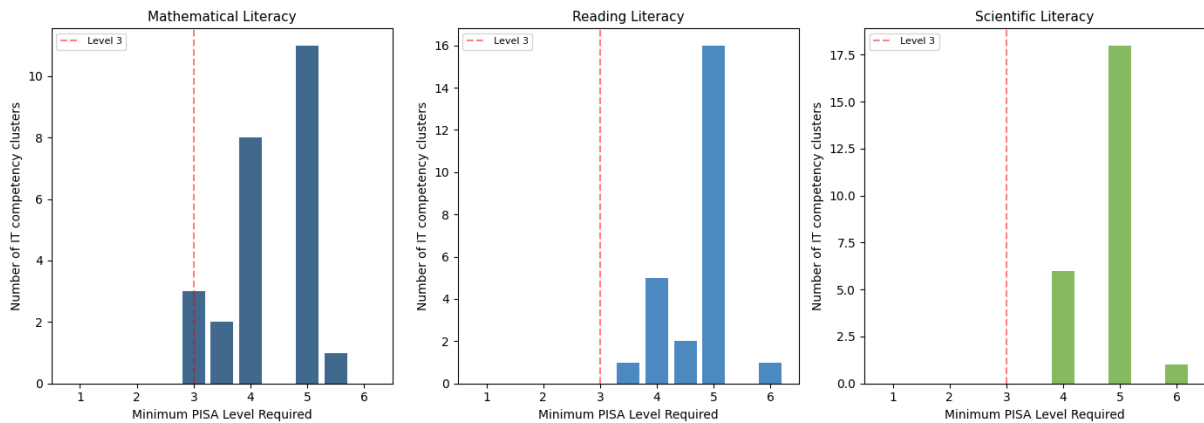


Figure 3. Minimum PISA proficiency required across 25 IT competency clusters

In terms of cross-domain correlation, Mathematics-Reading ($r=0.678$), Mathematics-Science ($r=0.735$), Reading-Science ($r=0.645$), all three are moderately to strongly correlated, meaning tasks that are cognitively demanding in one domain tend to be demanding in all three, the competencies reinforce rather than substitute each other. The highest correlation was observed between Mathematics and Science, suggesting a tight coupling between algorithmic logic and troubleshooting-based inductive reasoning.

The most frequent cognitive requirement is a simultaneous demand for PISA Level 5 across all three domains (Mathematics, Reading and Science), representing 40% of all clusters (10 of 25). Furthermore, the data reveals that minimum requirements for Reading literacy consistently exceed those for Mathematics by an average of 0.40 levels, highlighting that modern IT roles require more than just quantitative computation; they place an even heavier burden on the ability to interpret, evaluate and synthesize complex technical texts and specifications.

4.5 Sub-role PISA profile comparison

The 562 annotated IT tasks were grouped into career sub-roles based on their O*NET SOC codes. The categories are visible in Figure 4, which reveals a clear two-tier structure across IT career sub-roles. Data & Research, Databases and Management consistently reach Level 5 across all three literacy domains, while Networks & Infrastructure and QA & Testing settle uniformly at Level 4. The most striking profiles are the domain-split sub-roles: Software Development (the largest group with 220) demands Level 5 in Reading and Science but only Level 4 in Mathematics, reflecting the heavy burden of specification interpretation over formal computation; Security shows the inverse pattern, reaching Level 5 only in Scientific Literacy, driven by the causal and hypothetical reasoning required for threat modelling. No sub-role falls below Level 4 in any domain.

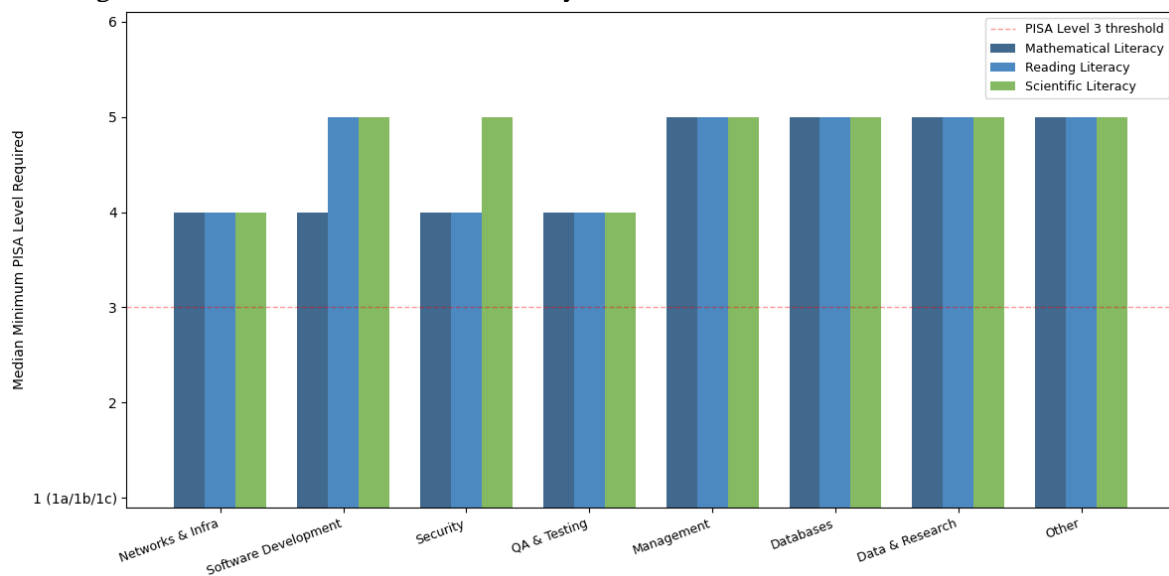


Figure 4. Minimum PISA proficiency required by IT career category

The results of the Kruskal-Wallis H-test confirm that the IT sector is not cognitively monolithic, revealing highly significant differences in PISA proficiency requirements across the eight identified sub-roles for Mathematics ($H=76.18$, $p<0.001$), Reading ($H=49.35$, $p<0.001$) and Science literacy ($H=49.19$, $p<0.001$). The notably higher H-statistic for the mathematical domain suggests that quantitative reasoning serves as the primary differentiator of cognitive demand between IT specializations, whereas reading and scientific reasoning exhibit slightly more uniform requirements across the sector.

4.6 Annotation analysis

A primary indicator of IT-PISA alignment framework's robustness is the 98.9% of high confidence level consensus achieved during the classification process: 98.9% of all tasks were successfully mapped to PISA proficiency levels with "high" confidence, while only 1.1% with "medium" confidence. The total absence of "low-confidence" annotations indicates a strong semantic resonance between the linguistic descriptors used in the IT competency corpus (O*NET/ESCO) and the cognitive reasoning requirements defined by the PISA 2022 framework.

The analysis revealed a complexity-certainty correlation (Figure 5), where the model demonstrated higher confidence when evaluating tasks at the upper end of the PISA scale. High-confidence tasks across all domains averaged a total PISA demand score of 13.4, whereas the small subset of tasks requiring only "medium" confidence averaged significantly lower at 7.3. This disparity suggests that top-tier IT competencies, such as designing complex neural architectures or evaluating systemic security protocols, utilize precise, technically explicit language that aligns mapped directly with PISA Level 5 and 6 descriptors. Conversely, more routine or operational tasks at the PISA Level 3-4 threshold possess a flatter linguistic profile, introducing a degree of ambiguity for the model when distinguishing between basic functional application and the onset of professional-level reasoning.

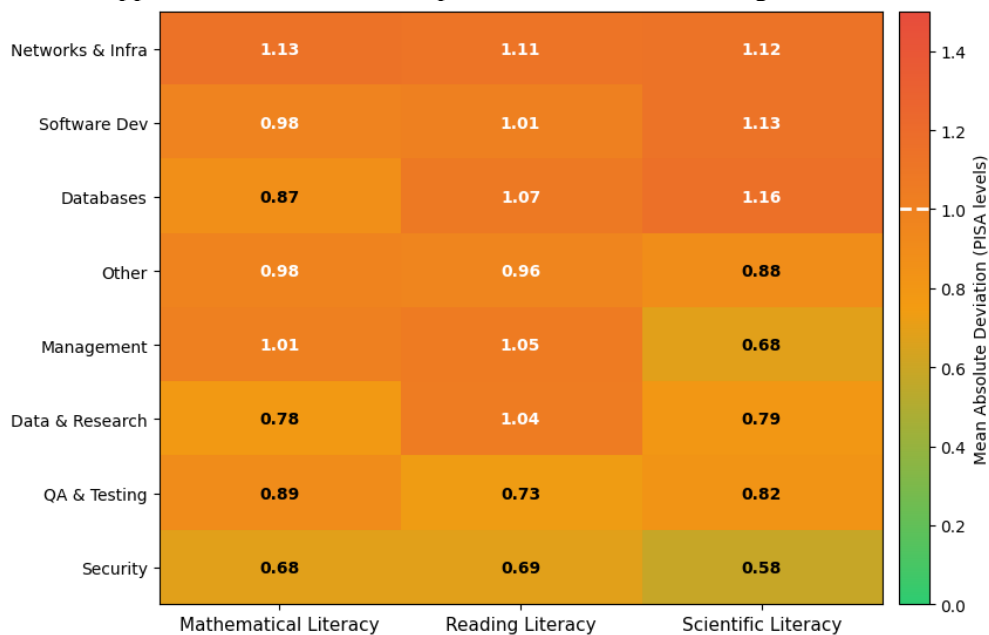


Figure 5. Heatmap of mapping MAD between LLM-annotated and O*NET-derived PISA levels

Figure 5 maps the MAD between LLM-annotated and O*NET-derived PISA levels across eight IT sub-roles and three literacy domains. Networks & Infrastructure is the most difficult sub-role to annotate consistently, with all three domains exceeding the threshold (Mathematics 1.13, Reading 1.12, Science 1.13). This reflects the high within-occupation variability of networking tasks: a single SOC code spans both routine operational work and highly abstract system design, making a single O*NET score a poor benchmark for individual task-level annotations.

Security stands out as the most internally consistent sub-role, with the lowest MAD across all three domains (0.68, 0.69, 0.58 respectively). Security tasks define a narrow, well-characterized cognitive type that both the LLM and O*NET evaluate similarly.

Reading literacy shows the widest spread of disagreement, exceeding 1.0 in five of the eight sub-roles, confirming that the LLM systematically interprets reading demands more broadly than the O*NET Written Comprehension score captures, particularly in roles with heavy documentation or specification requirements.

4.7 Transversal skills gap

The transversal analysis extends the study's core technical skill alignment by examining the soft-skill dimension of IT workforce competency.

In Figure 6, the forward-direction chart (left side) reveals that PISA provides partial or no coverage for 83% of ESCO transversal IT skill clusters (20 out of 24). Only 4 clusters (17%) achieve full alignment with PISA proficiency descriptors, a further 12 clusters (50%) are partially covered, meaning their cognitive demands overlap with PISA in some domains or levels but not consistently across the full literacy spectrum. Critically, 8 clusters (33%) receive no coverage whatsoever, confirming a substantial structural blind spot in the PISA assessment framework relative to the competency demands of the IT workforce.

Examining PISA from the ESCO side (inverse-direction chart), asking whether any IT transversal skill engages each PISA proficiency level, reveals that 16 out of 18 domain×level pairs (89%) are covered by at least one transversal cluster. The two exceptions are Mathematics literacy Level 6 and Science literacy Level 6, which reach only partial coverage: no single transversal cluster fully demands the abstract modelling and symbolic reasoning required at the highest proficiency tier. Zero PISA levels are missing, meaning that PISA levels are always touched by some IT transversal skill, but ESCO transversal skills extend well beyond what PISA can detect.

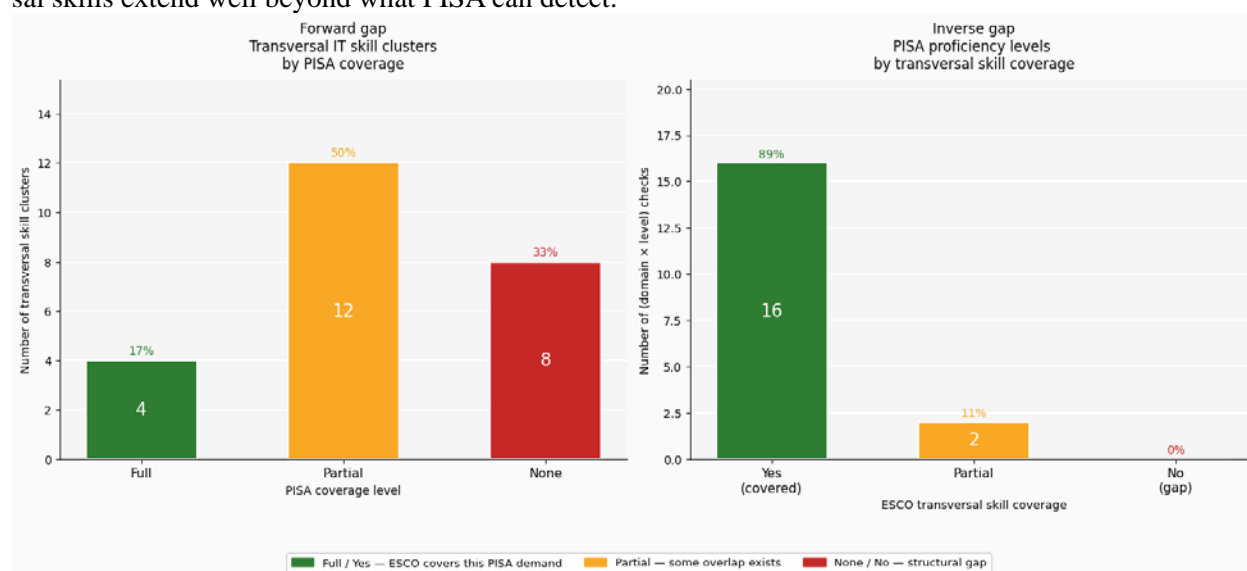


Figure 6. Forward and inverse gaps results

Table 11 shows all 8 clusters that fall entirely outside the three cognitive literacy pillars of PISA, representing the social-emotional, ethical, dispositional, and creative competencies that OECD research identifies as essential for 21st-century, but remain invisible in PISA's assessment design.

Table 11. Clusters with no PISA coverage

Skill	Reason
[TC01] applying cultural skills and competences	PISA 2022 focuses on mathematical, reading and scientific literacy, which do not directly assess artistic expression, aesthetic sensibility or the appreciation of diverse cultural and artistic forms.
[SK02] collaborating in teams and networks	PISA 2022 focuses on Mathematics, Reading and Scientific literacy, which do not directly assess the social and interpersonal skills involved in collaborating in teams and networks.
[TC01] demonstrating willingness to learn	PISA 2022 primarily assesses the application of Mathematics, Reading and Science knowledge and reasoning, and does not directly measure dispositions such as curiosity, willingness to learn or the meta-cognitive skill of self-reflection.
[TC03] following ethical code of conduct	PISA 2022 assesses cognitive literacy in mathematics, reading and science, which does not directly cover the demonstration of ethical conduct, loyalty or compliance with regulations.
[TC03] leading others	PISA 2022 focuses on mathematical, reading and scientific literacy, which do not directly assess the social, emotional and interpersonal competencies inherent in leading others.

[SK02] maintaining a positive attitude	PISA 2022 focuses on cognitive literacies (Mathematics, Reading and Science) and does not directly assess emotional regulation, resilience or attitudinal skills.
[SK03] responding to physical circumstances	PISA assesses cognitive literacies (Mathematics, Reading and Science) through problem-solving and understanding, not direct physical capabilities or immediate physical reactions to environmental circumstances.
[TC01] thinking creatively and innovatively	The PISA 2022 assessments for Mathematics, Reading and Science literacy primarily evaluate the application of knowledge and reasoning within defined contexts, rather than directly assessing the generation of new ideas, improvisation, or innovative solutions.

The transversal gap analysis serves to delineate the boundaries of the proposed framework. The core pipeline successfully aligns the cognitive-technical dimension of IT workforce demand to PISA proficiency levels, confirming that PISA’s literacy constructs are a valid reference point for reasoning complexity in IT task statements. However, the bidirectional analysis of ESCO transversal skills reveals that 33% of soft-skill clusters, primarily social-emotional, ethical and creative competencies, operate outside PISA’s cognitive-literacy scope, identifying which complementary assessment dimensions would be needed to extend the approach to the full breadth of workforce readiness. The framework thus contributes both a working alignment tool for cognitive competencies and a structured map of where its coverage ends, providing a foundation for future multi-instrument extensions.

5. Conclusions

This study demonstrates that IT occupations are uniformly cognitively demanding, with all 25 technical competency clusters requiring at least PISA Level 3 proficiency, establishing a *Level 3 floor* below which basic functional literacy is insufficient for workforce entry.

PISA measures the cognitive-reasoning capacity that IT work most heavily relies upon: all 25 technical competency clusters require at least Level 3, with Science Literacy emerging as the sole domain where 100% of clusters require Level 4 or above. What PISA misses is structural rather than scalar, 33% of ESCO transversal IT skill clusters, covering creative agency, ethical stewardship and collaborative intelligence, fall entirely outside PISA’s literacy constructs, and no proficiency level, however high, closes this gap. Raising PISA scores addresses the cognitive-technical readiness gap but leaves the transversal readiness gap intact.

Contrary to conventional assumptions, Science Literacy, not Mathematics, emerges as the dominant cognitive demand (44.5% of annotations; Level 5–6 in 64% of clusters), indicating that inductive and deductive reasoning, rather than calculation, constitutes the core competency of IT work. The LLM-based annotation framework is validated through statistically significant OLS-calibrated correlations with O*NET ability scores, confirming it captures meaningful variance in cognitive demand.

A bidirectional gap analysis of 24 ESCO transversal skill clusters reveals that while all PISA proficiency levels (1a–6) are engaged by at least one cluster, eight clusters (33%), spanning creative agency, ethical stewardship, collaborative intelligence and dispositional resilience, fall outside PISA’s cognitive-literacy scope entirely. This boundary mapping identifies precisely where complementary frameworks, such as social-emotional learning instruments, would be required for a complete assessment of IT workforce readiness.

Beyond its substantive findings, this paper contributes a reusable framework for aligning occupational task demands with population-level educational assessments: any researcher can apply the full pipeline to a different occupational domain by modifying only the SOC code prefixes, and the annotation protocol accommodates any assessment framework whose levels are described by explicit verbal descriptors. The annotated corpus and validation results are released on Zenodo as a stable reference independent of future model updates.

These findings carry direct implications for education and policy: IT curricula should target a minimum of PISA Level 3 with emphasis on scientific reasoning; the OECD Learning Compass 2030 offers a policy-transferable vocabulary for competency mapping; and the transversal gap analysis provides a concrete roadmap for extending cognitive-literacy assessments with ethical and social-emotional dimensions.

A four-model cross-validation against a blind human expert confirms that annotation precision scales predictably with model reasoning capacity and that the primary annotator’s disagreements are

systematically conservative, overestimating rather than underestimating cognitive demands, a directional bias that preserves the validity of the study's distributional findings while providing future researchers with empirical guidance for balancing annotation precision against scalability.

Future work should broaden the framework to other sectors, incorporate longitudinal job-posting data and integrate the identified gaps into a multi-dimensional IT workforce readiness index.

Statements and Declarations

Ethical Approval. Not applicable

Consent to Participate. Not applicable

Consent to Publish. Not applicable

Authors Contributions. Contribution to the study conception and design, including supervision: AMT, SVO and AB. Material preparation, data collection and analysis were performed by AMT. The first draft of the manuscript was written by AMT, SVO and AB, the second draft by SVO, and they also commented on all versions of the manuscript. All authors read and approved the final manuscript.

Disclosure statement/Competing Interests. The authors report there are no competing interests to declare.

Data availability statement. The data is available on <https://doi.org/10.5281/zenodo.19581463>

Acknowledgement and Funding: This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI - UEFISCDI, project number COFUND-DUT-OPEN4CEC-1, within PNCDI IV. This project has been funded by UEFISCDI under the Driving Urban Transitions Partnership, which has been co-funded by the European Commission.

References

- [1] D. (David) Zhang, G. Peng, Y. Yao, and T. R. Browning, "Is a College Education Still Enough? The IT-Labor Relationship with Education Level, Task Routineness, and Artificial Intelligence," *Information Systems Research*, vol. 35, no. 3, pp. 992–1010, Sep. 2024, doi: 10.1287/isre.2021.0391.
- [2] L. W. . Anderson and D. R. . Krathwohl, *A taxonomy for learning, teaching, and assessing : a revision of Bloom's taxonomy of educational objectives*. Longman, 2001.
- [3] D. H. Autor, F. Levy, and R. J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *Q. J. Econ.*, vol. 118, no. 4, pp. 1279–1333, Nov. 2003, doi: 10.1162/003355303322552801.
- [4] D. Acemoglu and D. Autor, "Skills, Tasks and Technologies: Implications for Employment and Earnings," 2011, pp. 1043–1171. doi: 10.1016/S0169-7218(11)02410-5.
- [5] E. Fernández-Macías and M. Bisello, "A Comprehensive Taxonomy of Tasks for Assessing the Impact of New Technologies on Work," *Soc. Indic. Res.*, vol. 159, no. 2, pp. 821–841, Jan. 2022, doi: 10.1007/s11205-021-02768-7.
- [6] C. B. Frey and M. A. Osborne, "The future of employment: How susceptible are jobs to computerisation?," *Technol. Forecast. Soc. Change*, vol. 114, pp. 254–280, Jan. 2017, doi: 10.1016/j.techfore.2016.08.019.
- [7] E. W. Felten, M. Raj, and R. Seamans, "How will Language Modelers like ChatGPT Affect Occupations and Industries?," *SSRN Electronic Journal*, 2023, doi: 10.2139/ssrn.4375268.
- [8] D. Acemoglu and J. Loebbing, "Automation and Polarization," Cambridge, MA, Sep. 2022. doi: 10.3386/w30528.
- [9] S. Noy and W. Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science (1979)*, vol. 381, no. 6654, pp. 187–192, Jul. 2023, doi: 10.1126/science.adh2586.
- [10] D. Acemoglu, "The simple macroeconomics of AI," *Econ. Policy*, vol. 40, no. 121, pp. 13–58, Jan. 2025, doi: 10.1093/epolic/eiae042.
- [11] Erik Brynjolfsson and Andrew McAfee, *The second machine age : work, progress, and prosperity in a time of brilliant technologies*. W.W. Norton & Company, 2014.
- [12] F. Chiarello, G. Fantoni, T. Hogarth, V. Giordano, L. Baltina, and I. Spada, "Towards ESCO 4.0 – Is the European classification of skills in line with Industry 4.0? A text mining approach," *Technol. Forecast. Soc. Change*, vol. 173, p. 121177, Dec. 2021, doi: 10.1016/j.techfore.2021.121177.
- [13] *PISA 2022 Results (Volume I)*. OECD Publishing, 2023. doi: 10.1787/53f23881-en.
- [14] S. Oprea and A. Bâra, "Is Reading an Upstream Predictor of Science and Mathematics Achievements in PISA? A Bayesian Network Analysis for Policy Educational Interventions on Socio-Economic Dispersion and Gender Gaps," *Syst. Res. Behav. Sci.*, Apr. 2026, doi: 10.1002/sres.70055.
- [15] *PISA 2022 Assessment and Analytical Framework*. OECD Publishing, 2023. doi: 10.1787/dfe0bf9c-en.
- [16] E. A. Hanushek and L. Woessmann, *The Knowledge Capital of Nations*. The MIT Press, 2015. doi: 10.7551/mitpress/9780262029179.001.0001.
- [17] S. Gust, E. A. Hanushek, and L. Woessmann, "Global universal basic skills: Current deficits and implications for world development," *J. Dev. Econ.*, vol. 166, p. 103205, Jan. 2024, doi: 10.1016/j.jdeveco.2023.103205.
- [18] *OECD Skills Outlook 2019*. OECD Publishing, 2019. doi: 10.1787/df80bc12-en.

- [19] OECD, “OECD Future of Education and Skills 2030. OECD Learning Compass 2030. A Series of Concept Notes.” 2019. Accessed: Apr. 04, 2026. [Online]. Available: <https://www.oecd.org/education/2030-project/>
- [20] D. Deming and M. Silliman, “Skills and Human Capital in the Labor Market,” Cambridge, MA, Sep. 2024. doi: 10.3386/w32908.
- [21] F. Gilardi, M. Alizadeh, and M. Kubli, “ChatGPT outperforms crowd workers for text-annotation tasks,” *Proceedings of the National Academy of Sciences*, vol. 120, no. 30, Jul. 2023, doi: 10.1073/pnas.2305016120.
- [22] A. Zarifhonarvar, “Economics of ChatGPT: a labor market view on the occupational impact of artificial intelligence,” *Journal of Electronic Business & Digital Economics*, vol. 3, no. 2, pp. 100–116, Jun. 2024, doi: 10.1108/JEBDE-10-2023-0021.
- [23] E. Kasneci *et al.*, “ChatGPT for good? On opportunities and challenges of large language models for education,” *Learn. Individ. Differ.*, vol. 103, p. 102274, Apr. 2023, doi: 10.1016/j.lindif.2023.102274.
- [24] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.
- [25] M. Zhang, K. Jensen, S. Sonniks, and B. Plank, “SkillSpan: Hard and Soft Skill Extraction from English Job Postings,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 4962–4984. doi: 10.18653/v1/2022.naacl-main.366.
- [26] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 3980–3990. doi: 10.18653/v1/D19-1410.
- [27] K. Abdalgader, A. A. Matroud, and K. Hossin, “Experimental study on short-text clustering using transformer-based semantic similarity measure,” *PeerJ Comput. Sci.*, vol. 10, p. e2078, May 2024, doi: 10.7717/peerj-cs.2078.
- [28] F. Leon, M. Gavrilescu, S.-A. Floria, and A. A. Minea, “Hierarchical Classification of Transversal Skills in Job Advertisements Based on Sentence Embeddings,” *Information*, vol. 15, no. 3, p. 151, Mar. 2024, doi: 10.3390/info15030151.
- [29] D. Zare, L. Fernandez-Sanz, V. Pospelova, and I. López-Baldominos, “NLP and Text Mining for Enriching IT Professional Skills Frameworks,” *Applied Sciences*, vol. 15, no. 17, p. 9634, Sep. 2025, doi: 10.3390/app15179634.
- [30] K. Sha, A. Taeihagh, and M. De Jong, “Governing disruptive technologies for inclusive development in cities: A systematic literature review,” *Technol. Forecast. Soc. Change*, vol. 203, p. 123382, Jun. 2024, doi: 10.1016/j.techfore.2024.123382.
- [31] U.S. Bureau of Labor Statistics, “2018 Standard Occupational Classification System.” Accessed: Apr. 11, 2026. [Online]. Available: https://www.bls.gov/soc/2018/major_groups.htm
- [32] European Commission, “European Skills, Competences, Qualifications and Occupations (ESCO) Occupations.” Accessed: Apr. 11, 2026. [Online]. Available: <https://esco.ec.europa.eu/en/classification/occupation?uri=http%3A%2F%2Fdata.europa.eu%2Fesco%2Fesco>