BRIDGING EDUCATION AND THE LABOR MARKET: AN EDGE-AI ARCHITECTURE FOR AUTOMATED STUDENT-TO-CAREER MATCHING

Gorgi Kakasevski
Faculty of Informatics, AUE-FON University, North Macedonia, gorgik@fon.edu.mk
Toni Stojanovski
Faculty of Informatics, AUE-FON University, North Macedonia, toni.stojanovski@fon.edu.mk
Slavcho Chungurski
Faculty of Informatics, AUE-FON University, North Macedonia, slavcho.chungurski@fon.edu.mk
Anamaria Ilieva
FEIT, Ss. Cyril and Methodius University, North Macedonia, anamaria.ilieva@feit.ukim.edu.mk

Abstract: In the contemporary educational landscape, aligning academic curricula with rapid labor market transformations remains a critical challenge. For academic institutions and career advisors, manually bridging the gap between student competencies and industry demands is unscalable due to the highly unstructured, heterogeneous nature of student resumes (CVs) and corporate job postings. Furthermore, while modern Artificial Intelligence (AI) offers robust parsing capabilities, utilizing commercial, cloud-based Large Language Models (LLMs) to process sensitive student data introduces severe privacy, compliance, and data sovereignty risks. The primary purpose of this study is to overcome these limitations by engineering a secure, automated tool that empowers the "teacher of the future" to align student profiles with industry needs without compromising data integrity. The methodology proposes a privacy-preserving Edge-AI architecture designed for automated student-to-career matching. It utilizes a multi-stage pipeline of decentralized, locally hosted LLMs to perform intelligent data extraction. Specifically, the system employs Qwen 3.5:35B to digitize visually complex job offers via Optical Character Recognition (OCR), Gemma 3:27B for semantic structuring and strict JSON schema enforcement, and TranslateGemma:27B for high-fidelity localization into the Macedonian (or other local) language. A student-centric matching algorithm subsequently evaluates these profiles, computing semantic similarity scores across re-calibrated domains: mandatory prerequisites, competencies, formal education, practical exposure, and logistical alignment. Experimental results validate the efficacy of this local architecture. Executed on edge hardware (NVIDIA RTX 5090), the pipeline achieved a 99.4% JSON schema validity rate for CV structuring and a 100% success rate in deterministic matching outputs, with end-to-end processing averaging under 75 seconds, of which 30 seconds for CV-job matching. The conclusions drawn indicate that localized edge-computing architectures can achieve enterprise-grade parsing and matching while keeping all sensitive data strictly air-gapped, neutralizing cloud-based privacy risks and high costs. Based on these findings, it is highly recommended that academic IT departments and career centers adopt on-premises AI inference to maintain General Data Protection Regulation (GDPR) compliance while modernizing advisory services. Additionally, the data suggests that dynamically adjusting algorithmic weights toward foundational academic knowledge rather than longitudinal corporate tenure significantly improves matching fairness for entry-level candidates, paving the way for future integrations directly into university Learning Management Systems (LMS).
Keywords: Edge Computing, Automated Career Matching, Large Language Models (LLMs), Data Sovereignty, Educational Technology.

1. INTRODUCTION

The transition from higher education to the professional workforce is a critical phase in a student's development. Historically, educational institutions have operated somewhat asynchronously from the immediate demands of the labor market, relying on theoretical curricula that may lag behind rapid industry advancements (Suleman, 2018). However, the paradigm of modern education is shifting. The "teacher of the future" is no longer solely a transmitter of academic knowledge, but a proactive facilitator of student employability and career alignment. To fulfill this role effectively, educators and academic advisors require advanced, data-driven tools capable of identifying real-time skills gaps and matching student competencies with evolving labor market opportunities.

The primary technological barrier to this alignment is the unstructured nature of career data. Student portfolios, resumes (CVs), and corporate job postings exist in a multitude of heterogeneous formats. Traditional automated matching systems, such as legacy Applicant Tracking Systems (ATS), rely predominantly on rigid, keyword-based heuristics. These legacy systems are notoriously brittle; they frequently fail to recognize synonymous skills, ignore the contextual depth of practical experience, and heavily penalize candidates who lack precise lexical matches, resulting in high rates of false negatives (Bogen & Rieke, 2018).

Recent breakthroughs in Natural Language Processing (NLP) and Artificial Intelligence (AI) implemented via Large Language Models (LLMs) offer a solution to these semantic limitations. LLMs excel at contextual understanding and can be explicitly prompted to extract and organize unstructured text into highly structured, machine-readable formats, such as JavaScript Object Notation (JSON) (Zhao et al., 2023). However, the integration of these models into the educational sector presents a critical ethical and legal challenge. Transmitting sensitive, personally identifiable student information (PII) to external, often foreign-owned, cloud APIs (e.g., OpenAI, Anthropic) introduces unacceptable data privacy vulnerabilities and violates stringent academic data sovereignty frameworks, such as the General Data Protection Regulation (GDPR) (Feretzakis et al., 2025).

To resolve the tension between advanced AI capabilities and strict data privacy, this paper presents a decentralized Edge-AI architecture for automated student-to-career matching. By deploying highly optimized, open-weight language models on local hardware environments, institutions can harness the semantic reasoning of state-of-the-art AI while maintaining an absolute, air-gapped secure perimeter around student data.

This paper details the end-to-end pipeline of this proactive educational tool. The methodology begins with Optical Character Recognition (OCR) to digitize and parse CVs and job offers, followed by edge-based LLM inference to parse both CVs and job postings into strict, validated JSON schemas. Subsequently, the paper details a multi-variable matching algorithm that computes overall compatibility by scoring specific, weighted data clusters, such as formal education, accumulated experience, and technical skill sets. Through this localized architecture, the proposed system demonstrates how educational institutions can securely automate labor market alignment, providing students with precise, actionable career guidance without compromising data integrity.

2. AUTOMATED EXTRACTION AND STRUCTURING OF RESUME DATA

The fundamental challenge in automated career matching lies in the unstructured, highly variable nature of CVs and student resumes. Traditional parsing systems have historically relied on rigid regular expressions and keyword matching, which often fail to capture the nuanced, multi-dimensional competencies of a candidate (Montuschi et al., 2009). Modern labor market research dictates that candidate profiles must be transformed into standardized, ontology-based representations—structuring data into distinct semantic clusters such as formal education, longitudinal experience, technical proficiencies, and soft skills (Fazel-Zarandi et al., 2009). Furthermore, recent advancements in NLP emphasize that extracting data is no longer sufficient; systems must also perform semantic enrichment, such as standardizing vague job titles and calculating chronological seniority, to ensure accurate algorithmic matching (Ling et al., 2025).

To meet these modern structural requirements, the proposed Edge-AI architecture utilizes a highly constrained, instruction-tuned LLM to perform intelligent data extraction (Manish et al., 2024). Because OCR outputs frequently contain fragmented text, erratic spacing, and pagination artifacts (e.g., "Page 1 of 3"), the LLM is explicitly instructed to perform contextual error correction during the extraction phase. It mentally repairs hyphenation breaks and removes noise before mapping the data.

The second step is structuring the data in JSON format. To ensure the resulting data is perfectly interoperable with the subsequent matching engine, the LLM is guided by a complex prompt architecture designed to enforce strict data normalization and validation. The extraction logic operates on several core principles:

- **Heuristic Quality Gating:** Before deep parsing begins, the model evaluates the document for minimum viable extraction signals. The text must contain either a plausible name and contact method, a clear experience entry, or an education entry. If these signals are missing or the OCR text is overwhelmingly corrupted, the model aborts the extraction, flags the document as "failed," and generates localized, constructive feedback for the user.
- **Strict Schema Enforcement:** The model is constrained to output exclusively in a deterministic JSON format. It standardizes chronological data into strict ISO formats and categorizes skills into distinct arrays (technical, soft, and languages using ISO 639-1 codes).
- **Semantic Enrichment and Synthesis:** Rather than simply copying text, the LLM enriches the data. For instance, if a student lists a vague job title such as "Self-employed" or "Manager," the model synthesizes a highly descriptive title by analyzing the surrounding text (e.g., "Self-Employed: Wooden Interior Production"). Additionally, it calculates total years of experience, estimates a candidate's seniority level, and generates a standardized, recruiter-facing summary in the target language (Macedonian), regardless of the language used in the original CV.

By enforcing these extraction rules at the edge, the system transforms chaotic, unstructured student resumes into highly organized, machine-readable semantic profiles, forming a reliable foundation for the subsequent matching algorithm.

3. PARSING AND FEASIBILITY ANALYSIS OF JOB DESCRIPTIONS

For an automated matching architecture to function symmetrically, the structural integrity of the corporate job description must match the quality of the parsed student CV. Modern human resources literature emphasizes that an effective job offer must transcend a simple list of duties; it must be a highly structured ontology detailing non-negotiable prerequisites (must-haves), preferred qualifications (nice-to-haves), explicit technical tools, subjective soft skills, and clear parameters regarding seniority, compensation, and remote work policies (Khaouja et al., 2021; Spada et al., 2024). However, extracting this data automatically is technically complex due to how corporate entities publish these offers.

While some employers input structured data directly via web forms (e.g., occupation, salary), a vast majority upload job offers as highly designed marketing materials - often in PDF format or as multi-column infographic images. Traditional OCR engines struggle significantly with these designed graphics, frequently merging columns of text horizontally and corrupting the reading order (Holley, 2009). To overcome this, the proposed architecture first utilizes a Vision-Language Model (VLM) running locally to accurately interpret the spatial layout of designed images, extracting a coherent, linear text stream before passing it to the core parsing LLM.

Once the text is secured, the LLM initiates a two-phase process: Feasibility Validation and Semantic Structuring.

3.1. Phase 1: Feasibility Validation (Quality Gating)

Before attempting to structure the job data, the LLM acts as an expert HR validator. It evaluates whether the job post contains sufficient usable text for algorithmic matching. If the text is deemed "unusable," the model categorizes the failure (e.g., too_short, ocr_low_quality, format_noise) and triggers specific administrative actions. Crucially, if the failure originates from the employer's input (e.g., uploading an empty document or providing only two sentences of text), the LLM generates a localized, actionable suggestion (in Macedonian) instructing the company on how to rectify the job post. If the text passes validation, the system flags the extraction status as "ok" and proceeds to Phase 2.

3.2. Phase 2: Semantic Structuring and Echoing

During this phase, the LLM extracts the core requirements into a strict JSON schema. The model is explicitly constrained to extract technical skills, tools, and certifications exactly as written in their native terminology (e.g., "Python," "C#," "AWS") without translating them, ensuring high-fidelity matching against student CVs. Conversely, subjective traits are segregated into a distinct "soft skills" array.

Furthermore, the model is programmed to infer implicit data based on contextual clues. For example, if seniority is not explicitly stated, the LLM calculates it based on the required "yearsOfExperience" (e.g., 0 - 2 years mapping to "Junior"). Alongside this AI-driven extraction, the architecture enforces an "Echo" protocol, ensuring that deterministic parameters explicitly provided by the employer via the platform's user interface - such as the official company name, predefined industry tags, and geographic city - are preserved exactly as entered, overriding any hallucinatory tendencies of the LLM.

By applying rigorous feasibility validation and separating objective technical requirements from subjective soft skills, this pipeline transforms visually complex corporate job offers into standardized digital blueprints, ready for the multi-weighted matching algorithm.

4. MULTI-WEIGHTED SEMANTIC MATCHING ALGORITHM

The core objective of automated candidate screening is to calculate a precise, multi-dimensional alignment score between the candidate's structured profile (from Chapter 2) and the job's requirements (from Chapter 3). Modern algorithmic HR research demonstrates that matching cannot rely on a single, monolithic similarity score; rather, it requires a weighted, multi-parameter approach that independently evaluates distinct semantic clusters, such as non-negotiable prerequisites, contextual experience, and logistical alignment.

This algorithmic necessity is validated by standardized corporate screening protocols. As outlined in standard HR screening rubrics, candidate assessments fundamentally depend on evaluating "Mandatory Requirements," cross-referencing "Skills Match," analyzing "Trajectory and Stability" within the experience, and explicitly identifying "Strengths" and "Gaps". A system that merely counts keyword overlaps fails to replicate this human-level, contextual evaluation.

To achieve this sophisticated evaluation at the edge, the proposed architecture deploys a two-stage LLM pipeline. The first stage executes the complex semantic reasoning, while the second stage manages linguistic localization.

4.1. Stage 1: English-Native Semantic Reasoning

Empirical testing reveals that open-weight LLMs exhibit significantly higher reasoning capabilities and adherence to complex prompt constraints when operating in English, due to the predominant language of their training corpora. Therefore, the core matching engine is prompted exclusively in English.

The LLM is provided with both the CV JSON and the Job JSON and instructed to output a rigid evaluation schema. The algorithm executes the matching based on several critical constraints:

- **Evidence-Based Scoring:** The model cannot simply mark a requirement as "met." For every mandatory requirement fulfilled, the LLM must extract and cite explicit evidence from the CV, ensuring algorithmic transparency and mitigating LLM hallucination.
- **Contextual Skill Evaluation:** A skill is only counted as a match if it is directly tied to a specific work experience or project. Aspirational skills or skills briefly mentioned outside a professional context are excluded.
- **Relevance over Duration:** Experience is evaluated based on relevance, not mere duration. For example, five years of unrelated industry experience scores lower than two years in the precise target role. Soft skills are matched based on semantic meaning (e.g., "communication" vs. "team coordination") rather than exact lexical matches.

To accommodate entry-level candidates (students), the algorithm calculates an overall score (0–100) using re-calibrated weights: mandatory prerequisites (30%), technical/soft competencies (25%), formal education (20%), practical exposure (15%), and logistical alignment (10%). This prioritizes foundational knowledge over long-term corporate tenure. Crucially, a "Cap Rule" ensures that if essential prerequisites are missing, the score is mathematically capped at 40, preventing inflated ratings for unqualified applicants. The engine then outputs this holistic evaluation - including specific academic strengths, skill gaps, and logistical risks - in a strict JSON format.

4.2. Stage 2: Constrained Linguistic Localization

While the reasoning is performed in English to maximize algorithmic fidelity, the final output must be accessible to local HR professionals and academic advisors in the target language (Macedonian). To achieve this without compromising the structured JSON format, the architecture utilizes a secondary, specialized translation LLM.

This localization model is governed by strict linguistic constraints. It is explicitly instructed to translate all string values (including metadata labels like "fitLabel" and arrays containing strengths and risks) from English to pure Macedonian. It is constrained to avoid regional suffixes or vocabulary from neighboring languages (e.g., Bulgarian or Serbian) that frequently contaminate general-purpose translation models. Furthermore, the model enforces exceptions for specific technical nomenclature, ensuring that software tools and frameworks (e.g., SQL, SAP, C#) remain in their native English forms. By splitting the reasoning and translation tasks into two distinct, constrained processes, the Edge-AI architecture guarantees both deep semantic accuracy and localized usability without sacrificing the strict JSON ontology.

5. EXPERIMENTAL EVALUATION AND SYSTEM PERFORMANCE

5.1. Experimental Setup: High-Performance Edge Computing Architecture

To conduct a CV to Job matching without leaking user data to third parties and AI APIs, also put in risk GDPR standards, the methodology of this study relies on a custom-engineered, high-performance local computing architecture. The system is designed to execute heavy natural language processing pipelines, vector embeddings, and large language model inference entirely on-premises, serving as a scalable blueprint for secure edge computing.

The hardware environment deployed for this research consists of the following primary components given in Table 1. The architecture is inherently scalable, permitting the integration of distributed scheduling and optimization algorithms to parallelize workflows across multiple nodes for enhanced performance (Kakasevski & Mishev, 2017).

Table 1. Hardware Specifications of the Local Edge Computing Architecture

Component Category	Hardware Specification	Function in NLP Architecture
Graphics Processing Unit (GPU)	NVIDIA GeForce RTX 5090 (32 GB VRAM, Driver: 591.86, CUDA 13.1)	Critical inference engine. Houses the quantized LLM natively in VRAM for rapid, parallelized tensor operations with zero-latency.
Central Processing Unit (CPU)	AMD Ryzen 9 9950X (16 Physical Cores, 32 Logical Threads)	Handles the heavy computational load of local data indexing, preprocessing, and orchestration of the AI models.
System Memory (RAM)	64 GB DDR5 (Speed: 6000 MHz)	Manages high-throughput data ingestion, holding large, unoptimized corpora text in fast memory.
Motherboard	Gigabyte X870E AORUS ELITE WIFI7 ICE	Provides advanced PCIe Gen 5 lanes to ensure maximum bandwidth between the CPU, system memory, and the GPU.

Source: Authors' own elaboration.

The 32 GB VRAM threshold of the RTX 5090 is uniquely suited for edge AI; it provides the necessary memory overhead to load state-of-the-art models directly into the graphics memory. This prevents the severe latency penalties associated with offloading inference layers to the system RAM. Simultaneously, the 32-thread AMD Ryzen processor ensures that the GPU is not bottlenecked during data ingestion, leaving the GPU entirely dedicated to neural network inference.

5.2. Multi-Stage Edge LLM Selection and Evaluation

To optimize the architecture for the distinct computational requirements of each phase (digitization, structuring, and localization), a rigorous evaluation of modern open-weight LLMs was conducted. The system was tested across a diverse suite of eleven models, evaluated both with and without chain-of-thought ("thinking") reasoning parameters enabled. The tested models included: the Gemma 3 family (4B, 27B), the Gemma 4 family (26B, 31B), GLM-4.7-Flash, GPT-OSS:20B, Llama 3.1:8B, the Qwen 3.5 family (9B, 27B, 35B), and TranslateGemma:27B.

The selection criteria prioritized three operational metrics critical to an automated educational tool: zero-shot precision in semantic reasoning, inference latency on the local RTX 5090 hardware, and - most importantly - absolute adherence to deterministic JSON outputs without schema breakage or syntax corruption.

Based on these experimental trials, the architecture was segmented to utilize highly specialized models for each distinct computational task:

- **Vision and OCR Extraction:** The Qwen 3.5:35B model was selected for the initial digitization phase. Experimental results demonstrated that it possessed superior spatial awareness, effectively interpreting multi-column, highly designed corporate job offer images where traditional OCR engines produced fragmented noise.
- **Semantic Structuring and Matching:** For the core task of parsing unstructured CVs and calculating the multi-weighted semantic matches, the Gemma 3:27B model was deployed. While newer architectures, including the Gemma 4 family (26B and 31B), were rigorously evaluated for potential inference speed advantages, empirical testing revealed critical generative instabilities within their outputs. Specifically, the Gemma 4 models exhibited high failure rates in constrained generation, suffering from frequent JSON schema corruption, severe lexical repetition, and out-of-distribution token hallucinations (e.g., injecting unexpected Chinese characters into English/Macedonian outputs).
- **Constrained Linguistic Localization:** To translate the final output into the target language (Macedonian) without corrupting the English JSON keys, TranslateGemma:27B was chosen. General-purpose models frequently failed by either translating the programmatic keys or hallucinating unsupported regional dialects. TranslateGemma:27B provided the highest precision in maintaining strict structural integrity while executing pure language translation.

5.3. System Performance and Deterministic Output Rates

Running massive 31-billion and 35-billion parameter models on edge hardware requires careful resource orchestration. By deploying heavily quantized versions of these models, the architecture successfully loaded the required tensors entirely within the 32 GB VRAM threshold of the RTX 5090, avoiding the severe latency penalties of CPU offloading.

The experimental performance of the selected models was measured against an evaluation dataset of 100 heterogeneous student CVs and 100 complex corporate job descriptions. The results, detailed in Table 2, validate the feasibility of edge-based processing for high-volume academic environments.

Table 2. Experimental Performance Metrics of the Multi-Stage LLM Architecture

Pipeline Stage	Selected Edge Model	Avg. Inference Time (sec)	JSON Schema Validity Rate	Task Success Criteria
OCR & Spatial Parsing	Qwen 3.5:35B	14.2 s	N/A (Text Output)	>92% accurate text reconstruction from complex images.
CV Structuring	Gemma 3:27B	23.8 s	99.4%	Complete extraction of formal education and skills.
Job Structuring	Gemma 3:27B	8.1 s	99.1%	Accurate isolation of "must-have" vs. "soft" skills.
Semantic Matching Engine	Gemma 3:27B	13.4 s	100%	Flawless multi-weighted scoring with extracted evidence.
Localization (Macedonian)	TranslateGemma:27B	15.5 s	100%	Zero corruption of JSON keys; high linguistic fidelity.

Source: Authors' original research.

As demonstrated, Gemma 3:27B for the core reasoning engine resulted in near-perfect JSON validation rates (>99%). This is a critical threshold for software integration, as broken JSON outputs cause catastrophic pipeline failures in automated systems. Furthermore, the total end-to-end processing time for a complete candidate-to-job

match averaged under 75 seconds, although the first phase of CV analyzing is done only once while matching phase last under 30 seconds for each next job offer.

These results definitively prove that educational institutions do not need to rely on external, privacy-compromising cloud APIs. A properly orchestrated local edge-computing architecture can achieve enterprise-grade parsing, matching, and translation while keeping all sensitive student data completely air-gapped and strictly compliant with modern data sovereignty standards.

6. CONCLUSION AND FUTURE WORK

Bridging the gap between academic preparation and dynamic labor market demands requires advanced, data-driven alignment. However, integrating state-of-the-art AI into the educational sector has historically forced institutions to compromise student data privacy by relying on external, cloud-based commercial APIs. This paper proposed and validated a decentralized Edge-AI architecture that fully resolves this tension, enabling highly accurate, automated student-to-career matching entirely on-premises.

By engineering a multi-stage pipeline of specialized open-weight Large Language Models, the system successfully managed the entire extraction and matching lifecycle. The deployment of Qwen 3.5:35B effectively digitized visually complex job offers, while the Gemma 3:27B model demonstrated exceptional zero-shot reasoning and schema adherence (>99% valid JSON) in semantically structuring resumes and matching candidate profiles. Finally, TranslateGemma:27B ensured high-fidelity localization into Macedonian without corrupting the underlying programmatic architecture.

While the proposed edge architecture establishes a secure and highly functional foundation for automated career matching, the intersection of educational technology and local AI presents several high-value avenues for future research. Future iterations of this system could aggregate the "skill gaps" identified across hundreds of individual student evaluations. By analyzing these macroscopic, the LLM can generate automated curriculum recommendations. Research should explore the development of secure, lightweight middleware to integrate this local AI pipeline directly into existing university Learning Management Systems (LMS) or student portals. As the labor market increasingly values demonstrative work over written resumes, future architectures should integrate advanced Vision-Language Models (VLMs) capable of natively analyzing multimodal student portfolios.

REFERENCES

Bogen, M., & Rieke, A. (2018). Help wanted: An examination of hiring algorithms, equity, and bias. Upturn.

Fazel-Zarandi, M., & Fox, M. S. (2009). Semantic matchmaking for job recruitment: An ontology-based hybrid approach. CEUR Workshop Proceedings, 525.

Feretzakis, G., Vagena, E., Kalodanis, K., Peristera, P., Kalles, D., & Anastasiou, A. (2025). GDPR and Large Language Models: Technical and Legal Obstacles. Future Internet, 17(4), 151.

Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine, 15(3/4).

Kakasevski, G., & Mishev, A. (2017). Optimization and scheduling algorithm for data intensive workflows in distributed data mining architecture. Eurocon, Ohrid, North Macedonia.

Khaouja, I., Kassou, I., & Ghogho, M. (2021). A Survey on Skill Identification From Online Job Ads. IEEE Access, 1-1. 10.1109/ACCESS.2021.3106120.

Ling, Z., Zhang, H., Cui, J., Wu, Z., Sun, X., Li, G., & He, X. (2025). Beyond Human Labels: A Multi-Linguistic Auto-Generated Benchmark for Evaluating Large Language Models on Resume Parsing. 31907-31933. 10.18653/v1/2025.emnlp-main.1626.

Manish, V., Manchala, Y., Vijayalata, Y., Chopra, S.B., & Reddy, K.Y. (2024). Optimizing Resume Parsing Processes by Leveraging Large Language Models. 2024 IEEE Region 10 Symposium (TENSYMP), 1-5.

Montuschi, P., Gatteschi, V., Lamberti, F., Sanna, A., & Demartini, C. (2014). Job Recruitment and Job Seeking Processes: How Technology Can Help. IT Professional 16(5):41-49.

Spada I, Fabbroni V, Chiarello F, Fantoni G (2024) Standardising job descriptions in the humanitarian supply chain: A text mining approach for recruitment process. PLoS ONE 19(7): e0305961.

Suleman, F. (2018). The employability skills of higher education graduates: Insights into conceptual frameworks and methodological options. Higher Education, 76:263-278.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.