JobResQA: Semi-Automatic Multilingual Benchmark Creation for LLM Machine Reading Comprehension on Résumés and Job Descriptions

Casimiro Pio Carrino[1,2], **Paula Estrella**[2], **Rabih Zbib**[2], **Carlos Escolano**[1], **José A. R. Fonollosa**[1]
Universitat Politècnica de Catalunya[1], Avature Machine Learning[2]
{casimiro.pio.carrino, carlos.escolano, jose.fonollosa}@upc.edu
{casimiro.carrino, paula.estrella, rabih.zbib}@avature.net

Abstract

We present a methodology for building privacy-preserving multilingual QA benchmarks in low-resource and sensitive domains, demonstrated through JobResQA, a multilingual MRC benchmark over synthetic HR documents. The dataset comprises 581 QA pairs across 105 synthetic résumé-job description pairs in five languages (English, Spanish, Italian, German, and Chinese), with questions spanning four types based on document source (intra vs. cross-document) and reasoning complexity (single-hop vs. multi-hop). We propose an anonymization synthetic data pipeline, with controlled attributes (via placeholders) to enable future fairness studies. Our cost-effective, human-in-the-loop translation pipeline based on TEaR methodology incorporates MQM error annotations and selective post-editing. Baseline evaluations across multiple open-weight LLM families using LLM-as-judge reveal higher performance on English and Spanish but substantial degradation for other languages, highlighting critical cross-lingual MRC gaps. Our pipeline, where LLMs act as synthesizers, translators, and evaluators under human oversight, constitutes a reusable methodology for resource creation and a case study in evaluation-integrity challenges of LLM-era benchmark construction.

Keywords: machine reading comprehension, HR, multilingual QA

1. Introduction

Sensitive domains such as Human Resources (HR), medicine, and law face a shared bottleneck in resource creation: real data cannot be shared due to privacy constraints, yet annotation requires expensive domain expertise. Synthetic data generation addresses the first barrier while human-in-the-loop pipelines address the second. HR is a particularly high-impact instantiation of this challenge, as LLMs are increasingly applied to résumé parsing, candidate-job matching, interview evaluation, and conversational support, already outperforming traditional keyword-based systems in candidate matching (Bevara et al., 2025), while HR-focused dialogue datasets demonstrate the potential of conversational HR agents (Xu et al., 2024).

However, this rapid adoption raises concerns about accuracy, reproducibility, and fairness, as controlled experiments show that current models often perpetuate demographic and cultural biases (Nghiem et al., 2024; Rao et al., 2025), with implications under emerging AI regulatory frameworks such as the *EU AI Act*[1].

Addressing these risks requires reproducible and publicly available benchmarks for LLM performance assessment, especially in multilingual contexts (Otani et al., 2025). Recent works have begun providing such resources, including annotated datasets for skills and job matching (Gasco et al., 2025; Zhang et al., 2022) and LLM-generated synthetic résumés and job descriptions (JDs) that reduce privacy exposure while enabling controlled fairness studies (Skondras et al., 2023; Saldivar et al., 2025).

One important use case of LLMs in HR is the analysis of résumés for matching with JDs. This task involves asking questions about the skills, experience, and background of a candidate in relation to a JD. Framing this process as a Machine Reading Comprehension (MRC) task enables knowledge-intensive Question Answering (QA) approaches that can better assess LLM's reasoning about candidate-job suitability. While a few works have introduced HR-related QA datasets (Xu et al., 2024; Luo et al., 2023; van Toledo et al., 2022), existing resources either focus on extractive, single-document CV questions or lack realistic, multilingual, and bias-controllable résumé-JD QA pairs.

Motivated by these challenges, we introduce JobResQA, a synthetic multilingual QA benchmark designed to approximate realistic HR scenarios with recruiter-style questions over résumé-JD pairs. The dataset is derived from real-world data through a de-identification and synthesis pipeline, resulting in anonymized yet realistic résumés and JDs. Jo-

[1] https://artificialintelligenceact.eu/the-act/

Proceedings of The Fourth Workshop on the Role of Resources in the Age of Large Language Models (RESOURCEFUL 2026), pages 161–176 May 11, 2026.

Q. Type	Definition	Example (EN)	Example (ES)
Intra-Doc Single-hop	Answerable from a single document (résumé *or* JD) using one piece of information.	*What is the highest degree the candidate has earned?*	*¿Cuál es el título más alto que ha obtenido el/la candidato/a?*
Intra-Doc Multi-hop	Requires combining multiple pieces of information within a single document (résumé *or* JD).	*What is the candidate's most specialized area of competence?*	*¿Cuál es el área de competencia más especializada del/de la candidato/a?*
Cross-Doc Single-hop	Requires one piece of information from each document (résumé *and* JD).	*Does the candidate meet the basic technical requirements for MS Office proficiency?*	*¿Cumple el/la candidato/a con los requisitos técnicos básicos de competencia en MS Office?*
Cross-Doc Multi-hop	Requires combining multiple pieces of information from both documents.	*Does the candidate's educational background exceed the preferred qualifications for this position?*	*¿La formación académica del/de la candidato/a supera las cualificaciones preferidas para este puesto?*

Table 1: Question types with parallel examples in English and Spanish.

bResQA spans question types from basic factual extraction to complex, cross-document reasoning, and includes controlled demographic attributes that may support future bias analysis. It is annotated in English and extended to Spanish, Italian, German, and Chinese using a human-in-the-loop LLM translation pipeline.

Notably, this paper exemplifies the full LLM-as-resource-creator loop: the same model families evaluated here also generated the documents, translated them, and judge the answers, making human oversight the key mechanism for evaluation integrity (Arnardóttir et al., 2025). This scenario is increasingly common in resource creation for sensitive, data-scarce domains, and motivates the design choices we document below.

Our contributions are as follows:

- We present a reusable pipeline for building privacy-preserving multilingual QA benchmarks in sensitive, data-scarce domains, instantiated as JobResQA: 105 synthetic résumé-JD pairs, 581 QA items, five languages, and four question types spanning document source (intra vs. cross-document) and reasoning complexity (single-hop vs. multi-hop).
- We present a cost-effective, human-in-the-loop LLM translation pipeline using MQM error annotations and selective post-editing, producing quality-controlled parallel data in Spanish, Italian, German, and Chinese.
- We establish an initial cross-lingual evaluation baseline for LLM machine reading comprehension on résumés and JDs across several open-weight model families.

2. Related Works

We group related research into three main areas. QA and MRC tasks in HR have been explored by Xu et al. (2024) with HR-MultiWOZ, the first HR-focused dialogue dataset, and Luo et al. (2023) who modeled résumé understanding as multilingual MRC by generating QA pairs from English and Dutch résumés.

Synthetic data generation has proven effective for addressing data scarcity, with Skondras et al. (2023) showing that ChatGPT-generated résumés improve job classification, while Lorincz et al. (2022) and Yu et al. (2025) advanced vacancy generation and résumé matching through transfer learning and hypothetical embeddings.

Bias and fairness research has identified critical issues, as Saldivar et al. (2025) introduced demographic attributes in synthetic CVs for bias evaluation, Nghiem et al. (2024) revealed name-based and gender biases in LLM employment recommendations, and Rao et al. (2025) exposed cultural biases in interview evaluations.

Resource construction methodology and evaluation integrity form a fourth relevant strand. Wang et al. (2024) document positional and systemic biases in LLM-as-judge evaluation, motivating the human oversight we incorporate. Magar and Schwartz (2022) show that benchmark data encountered during LLM pre-training inflates evaluation scores, a risk our multi-step synthetic transformation pipeline is designed to mitigate. Arnardóttir et al. (2025) present a parallel case of LLM-assisted benchmark construction with automated evaluation in a different domain, showing the broader applicability of such pipelines. These four directions collectively inform both JobResQA's design and its methodological framing.

3. The JobResQA Dataset

JobResQA is a QA benchmark that instantiates our proposed methodology for privacy-preserving multilingual resource creation in sensitive domains, using HR as the application domain. The dataset contains 581 question-answer (QA) pairs anno-

tated over a set of 105 unique pairs of résumé and JD. The résumés and JDs are derived from real-world data through a data synthesis pipeline that produces synthetic, anonymized, yet realistic versions (see Section 4). The QA pairs are annotated manually following detailed guidelines to ensure quality and diversity (see Section 5), spanning four question types defined by document source and reasoning complexity defined in Table 1. The entire dataset is multi-way parallel across five languages: English (en), Spanish (es), Italian (it), German (de), and Chinese. The translations are produced by an LLM-based pipeline with human-in-the-loop corrections (see Section 6). The set enables cross-lingual QA evaluation, where an English instruction prompt is used regardless of the question and document language.

3.1. Main Characteristics

We designed the JobResQA benchmark to be realistic and representative of practical HR applications, capturing the complexity of real persons' career-related information and job requirements. We preserve the data's privacy and anonymity, while at the same time, we ensured certain properties to enable controlled studies in multilingual and fairness settings, as detailed below. The synthetic résumés and JDs are gender-inclusive and anonymized through a set of controlled attributes spanning multiple bias dimensions (demographic, socioeconomic, educational, etc.), enabling future systematic investigation of fairness in HR applications (see Appendix A.1).

3.2. Statistics and Data Fields

We report the main statistics of the JobResQA benchmark in Table 2 and describe briefly the main dataset's textual fields[2] as below:

- `resume`: text of synthetic candidate's résumé.
- `jd`: synthetic description of a role.
- `question`: recruiter-style question on the résumé in relation to the JD.
- `short_answer`: concise answer to the question, as a span, phrase, number, or yes/no.
- `explanation`: longer answer with explanatory rationale and evidence supporting the short answer.
- `question_type`: four-way categorization across two dimensions, document source (intra vs. cross-document) and reasoning complexity (single-hop vs. multi-hop) (see Table 1).
- `industry`: industry sector of the JD.
- `language`: language of all text fields.

Statistic	Value
QAs (#)	581
Unique résumés (#)	105
Unique JDs (#)	101
Unique résumé-JD pairs (#)	105
Industries (#)	24
Question types (%)	
- Cross-document Multi-hop	79.7%
- Intra-document Single-hop	9.0%
- Intra-document Multi-hop	8.4%
- Cross-document Single-hop	2.9%
Languages supported	en, es, de, it, zh

Table 2: JobResQA dataset statistics.

3.3. Accessibility and Reproducibility

We release JobResQA under the Creative Commons BY-SA 2.0 license[3]. We provide both data and code to support reproducibility at the GitHub repository (https://github.com/Avature/jobresqa-benchmark).

It includes the complete multilingual dataset, MQM error annotations from human evaluation, placeholders for all target languages, prompts for data synthesis, translation, and LLM-as-judge evaluation, as well as runnable scripts for experimentation.

4. Résumés and Job Synthesis

We detail the generation of realistic, anonymized synthetic résumés and JDs, along with the QA annotation to create recruiter-style questions and answers as illustrated in Figure 1.

4.1. Data Collection, Job Matching and Industry Classification

We start by collecting real-world résumés and JDs from a large pool of public job boards that are randomly sampled from diverse locations and industries to target a wide array of roles and domains. Then, we align candidates with suitable roles by performing semantic matching using the job titles of the résumés and JDs. We use the multilingual job title encoder in Deniz et al. (2024) to encode the job title of résumés and JDs into a shared embedding space, and compute cosine similarity to identify the most similar pairs. In particular, given a résumé job title, we obtain the top-10 JD titles from the ranking, and then we manually review and select the best

[2] For brevity, we omit fields containing numerical identifiers

[3] https://creativecommons.org/licenses/by-sa/2.0/deed.en

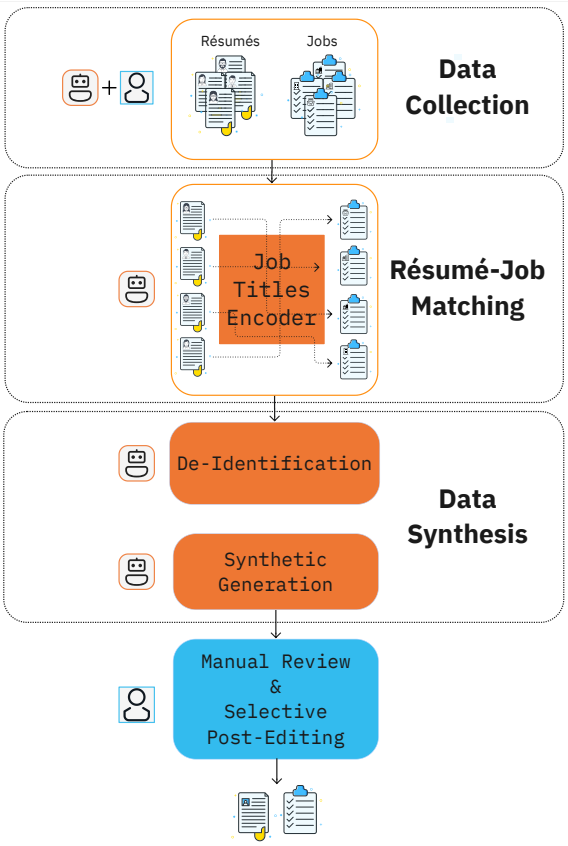

Figure 1: Data synthesis pipeline: collection and matching, de-identification, LLM-based synthesis, manual review.

match based on title similarity and industry alignment, improving over automatic threshold-based selection.

The final result is a selection of 105 matched résumé-JD pairs covering a total of 24 industries. We manually annotated the industry for each JD of the 105 résumé-JD pairs in the dataset, based on the job titles of the JD, following our internal taxonomy of 24 industries that groups similar sectors together. The distribution in Figure 2 shows a diverse range of industries, with the more common ones being Healthcare, Accounting/Finance, and Computer/Internet, while containing also less common ones such as Construction/Facilities, Government/Military, and Real Estate. This diversity ensures that the benchmark covers a wide variety of professional contexts and job requirements.

4.2. De-identification

We then pass the records through a de-identification stage to preserve the privacy of the data. For résumés we use the model in Retyk et al. (2023) to extract relevant entities such as contact information, work experience, education, and languages, and we replace all but

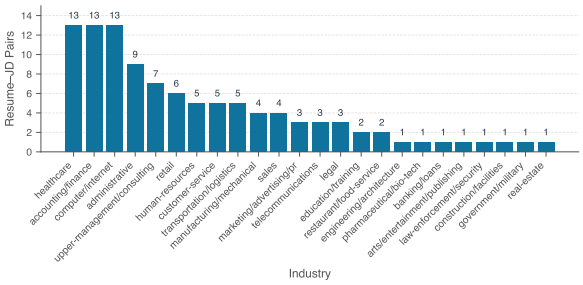

Figure 2: JD's industry distribution across the 105 résumé-JD pairs.

the job titles and skills with placeholders (`[NAME]`, `[PHONE]`, `etc.`).

For JDs, we implement a rule-based de-identification approach by creating a list of companies, branches, products and company-related identifiable entities and then extracting and replacing those with placeholders (e.g., `[COMPANY]`, `[PRODUCT]`, etc.) to remove traceability to the original company.

4.3. Synthetic Generation

We generated synthetic versions from de-identified résumés and JDs using carefully crafted prompts with OpenAI's GPT-4.1 (temperature 0.7, *top_p* 1).

For résumé generation, we apply three key transformations:

1. *anonymizing personal information* by replacing all PII with a standard set of placeholders (e.g., `[NAME]`, `[EMAIL]`, `[COMPANY]`, `[SCHOOL]`)
2. *modifying career-related content* to prevent traceability, job titles are replaced with different but career-progression-consistent alternatives, skills are substituted with pertinent equivalents, responsibilities and achievements are rephrased, language proficiencies are replaced with plausible alternatives, and dates are shifted forward while preserving chronological consistency
3. *normalizing structure* by mapping all content to a fixed set of predefined section names, with layout deliberately varied from the source to reduce traceability.

Crucially, while individual identifiers are replaced, the overall career narrative is preserved from the real-world source, including career gaps, non-linear trajectories, and authentic professional histories, ensuring that the synthetic résumés reflect realistic career patterns.

Similarly, JD generation involves:

1. *anonymizing company information* by replacing all company-related identifiable details with placeholders;

2. *rephrasing job content* to remove distinctive wording while preserving role-specific aspects including job title, skills, responsibilities, and requirements;

3. *preserving format and style* by maintaining comparable length, professional tone, and realistic formatting.

Collectively, these multi-step transformations can also play as a contamination-mitigation measure (Magar and Schwartz, 2022), since the resulting documents diverge substantially from any web-crawled source text, preserving evaluation integrity even when assessed models were trained on public corpora.

4.4. Manual Review and Selective Post-Editing

Finally, we manually reviewed and selectively post-edited the synthetic résumés and JDs to ensure high quality. We corrected minor issues (e.g., typos and formatting inconsistencies) and conducted a manual privacy audit over all 105 résumé-JD pairs to verify that no personally identifiable information (PII) remained after the de-identification and synthesis steps, removing any residual identifiers found. Then, we detected both sex-related terms (e.g., *female*, *male*) and gender-related terms[4] (e.g., *woman*, *man*) and replaced them with person-centered alternatives using *person* to ensure inclusivity. We also normalized all placeholders to ensure consistency across the dataset (see Table 5 in Appendix A.1) and translated them into each target language to ensure cross-lingual parallelism. Finally, we compared the synthetic documents against their original de-identified versions to verify that the overall career narrative and job requirements were preserved. This combined process of de-identification, synthesis, manual review, and post-editing yields 105 unique synthetic résumé-JD pairs (105 résumés and 101 JDs), preserving realistic professional content while ensuring anonymity.

5. Question-Answering Annotations

We consulted with HR experts and Talent Acquisition professionals to develop a curated question bank of recruiter-relevant questions suitable for real-world HR screening applications. Following prior work on QA resource development (**??**Lan et al., 2023), we conducted a pilot study on a small set of résumé-JD pairs, which informed the design of comprehensive annotation guidelines for non-expert annotators. The subsequent QA annotation was performed by linguists following these guidelines, with each annotator independently creating QA pairs for half of the résumé-JD pairs[5]. Each annotator created triplets of *(question, short answer, explanation)* focusing on specific candidate aspects (e.g., work experience). Each triplet was assigned to one of four question types across two dimensions: *document source* (intra-document: answerable from a single résumé or JD; cross-document: requires both documents) and *reasoning complexity* (single-hop: direct lookup from a single fact; multi-hop: synthesis across multiple facts), yielding four categories as shown in Table 1.

Annotators provided both short answers and explanations detailing their reasoning process. Importantly, they avoided targeting placeholders or gender-specific information in résumés and JDs, focusing instead on generalizable skills, experiences, and qualifications. Given the dataset's broad industry coverage, annotators consulted the question bank from HR experts, the ESCO dictionary (esc, 2020) and the O*NET database (National Center for O*NET Development) to clarify unfamiliar job titles, skills, or domain-specific terminology. Finally, a third non-expert annotator was instructed to review the entire dataset to assess the relevance of each question to practical HR screening tasks, ensuring the questions' applicability to real-world recruitment scenarios.

6. Human-in-the-Loop Machine Translation Pipeline

To evaluate LLMs' capabilities in HR-specific MRC tasks across multiple languages, we extended JobResQA to four additional languages: Spanish, Italian, German, and Chinese. Building on recent studies showing that LLMs can produce translations in controlled settings (Feng et al., 2025; Zhu et al., 2024; Cui et al., 2025; Koshkin et al., 2024), we developed an LLM-based machine translation pipeline with selective human review and feedback.

Our multi-stage translation process, designed specifically for résumés and JDs, includes machine translation, human error annotation, selective post-editing, and post-processing (Figure 3) . We combined automatic translation using *Claude Sonnet 4*[6] (temperature = 0) with review and error annotation by native speakers, thus balancing translation quality with efficiency. We run inference using AWS Bedrock service[7]

[4]For simplicity, we treat gender as binary (woman/-man) in this current version of the dataset.

[5]Due to the division of labor rather than overlap, inter-annotator agreement metrics are not available.

[6]anthropic.claude-sonnet-4-20250514-v1:0

[7]https://aws.amazon.com/bedrock/

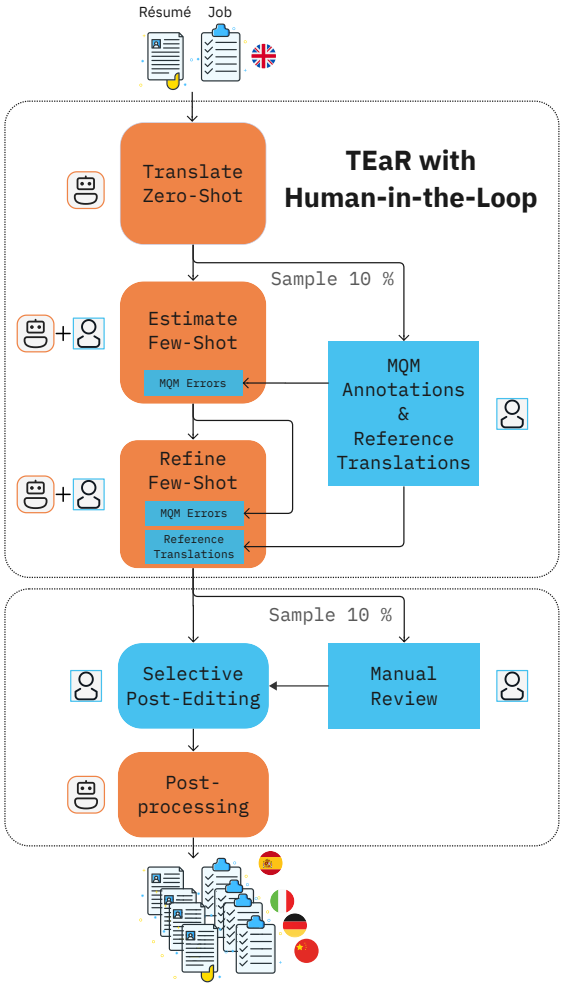

Figure 3: Human-in-the-loop TEaR translation pipeline for JobResQA: zero-shot translation, MQM error annotation & corrections, few-shot estimation, few-shot refinement and selective post-editing.

6.1. TEaR with Human-in-the-Loop

We implemented a human-in-the-loop variant of Translate-Estimate-Refine (TEaR) (Feng et al., 2025) guided by Multidimensional Quality Metrics (MQM) (Lommel et al., 2013) error annotations. MQM is an established framework for human evaluation of translation quality that structures annotator feedback into a hierarchical taxonomy of error types (e.g., accuracy, terminology, fluency) each assigned a severity level (critical, major, minor, or neutral). This structured approach transforms subjective translator judgements into actionable, fine-grained error signals, making it well suited to drive iterative LLM refinement in a human-in-the-loop pipeline. Full MQM category and severity definitions used in our annotations are provided in Appendix A.3.

Zero-Shot Translation. We started with an initial translation following the zero-shot translation prompt strategy in Feng et al. (2025), with an additional instruction that preserve placeholders and formatting. We translated résumés and JDs at the paragraph-level, while other fields were processed entirely. At this stage, the idea is to produce translations without any human guidance, which are later improved with human-in-the-loop feedbacks.

Human Feedback: MQM Errors Annotations and Corrected Translations. We sample approximately 10% of the résumé-JD pairs for manual review. Annotators identify translation errors using our custom MQM categories designed for HR documents: Terminology, Accuracy, Linguistic, Style, Locale, Design, and Custom.

Notably, we introduced *Hallucination* under the *Custom* category to capture AI-generated content errors, and *Gender-Inclusive* under the Style category to address concerns about inclusiveness in the translations. The gender-inclusive error category targets gender-specific translations and enforces a corrections that uses slashed forms (e.g., “des/der Kandidaten/-in”, “el/la candidato/a”, “del/la candidato/a”). All errors were rated across four severity levels: Critical, Major, Minor, and Neutral. This feedback guided the LLM in subsequent *Estimate and Refine* steps, driving it towards higher-quality translations and better aligned with human preferences.

Few-Shot Estimation. We utilized the errors from human MQM annotations to apply the few-shot estimation prompting strategy from Feng et al. (2025). This allowed us to automatically scale error estimation to the entire dataset following the MQM error categories we defined. These errors provide feedback to improve subsequent translations.

Few-Shot Refinement. Finally, we fed both the estimated MQM errors and corrected translations (used as references) to apply the few-shot refinement prompting strategy in Feng et al. (2025). Similar to the estimation step, this allows us to scale the refinement to the entire dataset. The corrected translations provide references that guide the LLM to refine the initial translations based on human feedback and preferences.

6.2. Manual Review, Selective Post-Editing, and Post-Processing.

To ensure high-quality translations, we sampled 10% of translated résumé-JD pairs for manual review by native speakers to identify main issues. We then addressed these issues through further selective post-editing on the full dataset, either manually or automatically. Below we describe the main issues we detected and how we addressed them:

Job Titles Consistency. We detected remaining untranslated English job titles in the 10% review sample and subsequently reviewed and replaced them with target-language equivalents across the full dataset, ensuring consistency across all dataset fields.

Automatic Verb Tense and Pronoun Consistency. We detected mixed verb tenses (present and past) and inconsistent pronoun perspectives (first- and third-person) across résumé sections. To correct this, we applied an LLM-based post-editing step using *Claude Sonnet 4*[8] (temperature = 0), followed by a final manual review. The LLM was instructed to use present-tense verbs or nominalized forms for the candidate's current or most recent position, nominalized forms for all past positions, and to remove first-person pronouns throughout to match standard résumé conventions across all languages.

Gender-Inclusive Forms. For Spanish, Italian and German, we detected and fixed gender-inclusive form issues. We automatically extracted all words containing the gender-inclusive slash "/" using a rule-based approach (e.g., "des/der Kandidaten/-in", "el/la candidato/a", "del/la candidato/a"), then manually fixed each occurrence to ensure consistency with MQM annotated errors. While this ensures formal correctness and inclusiveness, it may produce structures less common in authentic résumés from some locales.

Placeholders Translations. We manually translated all placeholders for non-English languages, validating semantic equivalence after translation and performing typological consistency checks, to maintain full parallelism across languages.

6.3. Translation Quality Evaluation

To provide an automated estimate of translation quality, we employed COMETKiwi (Rei et al., 2022, 2023), a reference-free metric specifically designed for quality estimation of machine translations without reference translations, serving as a proxy in the absence of human evaluation.

Table 3 presents average COMETKiwi scores for final translations and improvements (delta) over zero-shot baselines. Scores range from 83.07 to 85.51, with positive deltas (0.05 to 0.45) indicating that human-in-the-loop feedback and selective post-editing improved translation quality. Despite formal statistical significance testing was not conducted, the consistent improvements and the overall high scores across all languages suggest high-quality translations.

[8] anthropic.claude-sonnet-4-20250514-v1:0

Lang	COMETKiwi (2022)	COMETKiwi (2023, XL)
de	83.36 (**+0.09**)	74.65 (**+0.16**)
es	85.25 (**+0.33**)	78.08 (**+0.43**)
it	85.51 (**+0.24**)	78.95 (**+0.45**)
zh	83.07 (**+0.05**)	74.86 (**+0.19**)

Table 3: Translation quality scores and delta (Δ) over zero-shot baseline.

7. Evaluation Experiments

The goal of this section is to establish a first baseline evaluation on the JobResQA benchmark to assess LLMs machine-reading comprehension through cross-lingual question answering on résumé and JDs, with an English instruction prompt across all five languages. In the following, we describe the experimental setup, including the models, prompting strategy and evaluation metrics, and then we discuss the results. We run inference using AWS Bedrock service[9], which provides access to a variety of foundation models through API endpoints. We note that our use of GPT-4 for data synthesis and Claude Sonnet 4 for both translation and evaluation may introduce evaluation biases, as these models may share similar reasoning patterns.

7.1. Experimental Setup

Performing QA with LLMs. We designed a zero-shot prompt that instructs the model to act as an expert hiring assistant professional, answering questions about a candidate using only the JD and the provided résumé. Following the QA annotation guidelines in Section 5, the model is prompted to produce a concise short answer and a detailed explanation strictly grounded in the résume and/or JD. Responses should be factual, objective, and in the same language as the question, with explanations referencing specific details, quotes as evidence, and with justification for any information inferred from résumés and JDs. Since the instruction prompt is written in English regardless of the question and document language, this constitutes a cross-lingual evaluation setting.

For the QA task, we experimented with several open-weight, multilingual LLM models from various families and sizes, from medium to large. The selected models are Llama 3.1 Instruct (8B, 70B), Llama 3.2 Instruct (1B, 3B), and Llama 3.3 70B Instruct (Grattafiori et al., 2024), Mistral Small 2402 and Mistral Large 2402 (Jiang et al., 2023), and Gemma 3 Instruct (1B, 4B) (Team, 2025). We consider models between 1B and 8B parameters as medium-sized, and those above 8B as large. For generation, we set the temperature to 0 and max-

[9] https://aws.amazon.com/bedrock/

Model	en	es	de	it	zh
Mistral Large (2402)	0.69 ± 0.26	0.67 ± 0.25	**0.65** ± 0.25	0.66 ± 0.24	0.61 ± 0.29
Mistral Small (2402)	0.65 ± 0.28	0.61 ± 0.28	0.59 ± 0.29	0.60 ± 0.29	0.40 ± 0.28
Llama 3.3 70B Instruct	**0.73** ± 0.26	**0.69** ± 0.25	0.47 ± 0.38	**0.70** ± 0.25	0.48 ± 0.39
Llama 3.1 70B Instruct	0.72 ± 0.26	0.68 ± 0.25	0.64 ± 0.27	0.43 ± 0.38	**0.66** ± 0.27
Llama 3.1 8B Instruct	0.62 ± 0.30	0.57 ± 0.29	0.52 ± 0.28	0.56 ± 0.29	0.56 ± 0.30
Gemma 3 4B Instruct	0.64 ± 0.29	0.39 ± 0.21	0.48 ± 0.28	0.40 ± 0.21	0.41 ± 0.22
Llama 3.2 3B Instruct	0.55 ± 0.27	0.49 ± 0.26	0.45 ± 0.25	0.47 ± 0.26	0.47 ± 0.28
Llama 3.2 1B Instruct	0.35 ± 0.25	0.27 ± 0.24	0.25 ± 0.20	0.24 ± 0.20	0.26 ± 0.21
Gemma 3 1B Instruct	0.29 ± 0.17	0.15 ± 0.11	0.15 ± 0.11	0.16 ± 0.11	0.15 ± 0.10

Table 4: Average G-Eval scores (mean ± std) by model and language. The score ranges are based on evaluation rubrics: factually incorrect (0.0-0.3), mostly correct (0.3-0.6), correct but missing minor details (0.6-0.9). Higher scores indicate better alignment with human reference answers.

imum response length to 512 tokens to produce deterministic outputs aligned with the short answer and explanation format.

LLM-as-a-Judge Evaluation Despite the issues associated with fully automated evaluation Bavaresco et al. (2025), due to the scale of our evaluation (nine models across five languages and 581 QA items), human evaluation was not feasible. Instead, we employed an automated LLM-as-a-judge framework using the G-EVAL metric (Liu et al., 2023), which has been shown to correlate well with human judgments in various evaluation settings. This approach allows us to inject human expertise into the evaluation process through the design of the evaluation steps and rubrics, while leveraging the scalability of automated evaluation. Concretely, we provided a list of evaluation steps that guide the judge to compare the short answer and explanations from both the model and human responses. The judge checks that the short answer is concise and it uses minimal wording, and that the explanation provides detailed justification with specific references to the JD or résumé. The judge determines whether both answers communicate the same main factual conclusion based only on the provided documents, ensuring objectivity and factual accuracy. Reasoning and evidence in the actual output must be semantically equivalent to the human output, while ignoring stylistic differences. The judge also verifies that the model's answer is in the same language as the question and provides a brief justification for any major omissions, additions, mismatches, or failures to reference source documents. We also instructed the model to produce calibrated scores based on rubrics ranging from 0.0 "Factually incorrect" (0.0-0.3), "Mostly incorrect" (0.3-0.6), "Correct but missing minor details." (0.6-0.9), "100% correct" (0.9-1.0), using the G-Eval implementation from the open-source DeepEval library[10].

[10] https://github.com/confident-ai/deepeval

For the QA evaluation, we used *Claude Sonnet 4*[11] with temperature set to 0.7 and *top_p* to 0.9.

7.2. Results and Discussions

Table 4 reports QA performance (G-Eval mean ± std) for each model and language. Score bands follow the rubrics in Section 7.1. Higher scores indicate closer agreement with human reference answers.

The results reveal consistent patterns along both the model and language dimensions, interpreted through the G-Eval score bands defined in Section 7.1.

Correct but missing minor details (0.6 to 0.9). This band is reached almost exclusively in English and Spanish by the strongest models. Mistral Large is the only model to sustain it across all five languages, making it the most consistently multilingual performer. Llama 3.1 70B also attains it for German and Chinese, though it drops sharply for Italian. Models in the 4B to 8B range reach this band only in English.

Mostly correct (0.3 to 0.6). The majority of model and language combinations fall here, covering most non-English results for stronger models and nearly all results for smaller ones. A clear cross-lingual gap is evident in larger models, which achieve high scores in English but drop substantially for German and Chinese, with elevated standard deviations reflecting considerable variability across question types. Smaller models remain uniformly in this band across all five languages, showing a flatter cross-lingual profile driven by a lower performance ceiling rather than genuine multilingual robustness.

Factually incorrect (0.0 to 0.3). The two 1B models fall into this band for most or all languages, with minimal cross-lingual variation and low standard

[11] anthropic.claude-sonnet-4-20250514-v1:0

deviations, reflecting uniformly poor and undifferentiated performance regardless of language.

Notably, the 100% correct band (0.9 to 1.0) is not reached by any model in any language, indicating that substantial room for improvement remains across the board. Moreover, the substantial standard deviations observed in reflect variability driven by question-type complexity, as detailed in Appendix A.4.

Overall, the results confirm a consistent cross-lingual gap across all model families and sizes. Larger models benefit English and Spanish the most, while performance degrades markedly for lower-resource languages, suggesting that current multilingual pretraining is not sufficient for HR-specific cross-lingual MRC settings.

8. Conclusion and Future Works

We presented a methodology for building privacy-preserving multilingual QA benchmarks in sensitive, data-scarce domains, demonstrated through JobResQA in the HR domain. The pipeline, with de-identification, LLM synthesis, human-in-the-loop translation with MQM feedback, and LLM-as-judge evaluation, is designed to be reusable across other privacy-sensitive resource creation efforts. By incorporating controlled demographic and professional attributes (via placeholders) and gender-inclusive design, JobResQA provides infrastructure that may support future systematic fairness evaluation, though such studies remain to be conducted. Our human-in-the-loop translation pipeline with MQM annotations demonstrates a quality-cost tradeoff in producing multilingual datasets. Baseline evaluations reveal substantial performance gaps across languages and model sizes, highlighting the need for improved cross-lingual capabilities in HR contexts. Future work will extend the benchmarks with more questions and languages, and perform bias studies using the controlled attributes.

9. Limitations

We acknowledge the main limitations of our work:

Localization and Translation Quality Despite human-in-the-loop review, our translations may not fully capture native writing styles, and gender-inclusive rewriting may reduce perceived naturalness. Some inconsistencies in narrative voice can arise from paragraph-level translation and synthetic generation. These factors may limit task performance on authentic linguistic patterns.

QA Annotations Annotator subjectivity, particularly for complex cross-document questions, may reduce alignment with LLM outputs. The absence of inter-annotator agreement metrics limits verification of annotation consistency and reliability.

Synthetic Data and Privacy Our benchmark relies entirely on LLM-generated documents preserving career narratives but not authentic formatting inconsistencies. Predefined section structures and uniform layout reduce ecological validity, limiting generalizability to diverse real-world résumé formats. Additionally, whilst a manual privacy audit was conducted over all documents, the privacy-preserving effectiveness of the de-identification and synthesis process has not been formally quantified.

Dataset Scale At 105 résumé-JD pairs and 581 QA items, the dataset is relatively small and may limit generalizability across diverse HR scenarios and employment patterns.

Evaluation Methodology LLM-as-judge evaluation is susceptible to systemic biases and may not fully capture answer quality compared to human assessment. Using the same model for both translation and evaluation may introduce circularity, and sharing model families across synthesis, translation, and evaluation stages risks inflated scores. Calibrating judge scores against human judgments remains a priority for future work.

10. Ethical Statement

Potential Applications. The dataset can facilitate the responsible development and evaluation of HR-oriented language technologies, such as chatbots or virtual assistants for candidate screening, résumé parsing, and job-candidate matching. These high-risk applications should always be evaluated rigorously before being deployed in real-world settings, ensuring fairness, transparency, and accountability in decision-making. Importantly, the automation of hiring decisions raises specific

ethical concerns, including the risk of opaque or unaccountable filtering of candidates and the potential for historical biases encoded in training data to be perpetuated at scale. Therefore, any deployment should be subject to human oversight and regular auditing.

Human-in-the-Loop Annotations. Professional annotators and native speakers were involved throughout all the stages of the dataset creation process. We ensured fair compensation and clear guidelines to support ethical labor practices and document generation and annotation processes for transparency and reproducibility. We believe and emphasize the role of human expertise to ensure high-quality data and produce ethical outcomes.

Implications for Bias Research in LLMs. Our dataset is fully synthetic and anonymized, containing placeholder entities and gender-inclusive language. These design choices enable controlled investigation of potential bias attributes, such as demographic, gender, racial, and educational ones in LLMs applied to HR-related tasks. By systematically varying bias-related variables while removing personally identifiable information, our approach supports the study of model behavior based on content rather than identity cues.

11. Bibliographical References

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Philipp Martins, Andre andj Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Ravi Varma Kumar Bevara, Nishith Reddy Mannuru, Sai Pranathi Karedla, Brady Lund, Ting Xiao, Harshitha Pasem, Sri Chandra Dronavalli, and Siddhanth Rupeshkumar. 2025. Resume2vec: Transforming applicant tracking systems with intelligent resume embeddings for precise candidate matching. *Electronics*, 14(4).

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.

Daniel Deniz, Federico Retyk, Laura García-Sardiña, Hermenegildo Fabregat, Luis Gasco, and Rabih Zbib. 2024. Combined unsupervised and contrastive learning for multilingual job recommendation.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. TEaR: Improving LLM-based machine translation with systematic self-refinement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3922–3938, Albuquerque, New Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika

Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant

Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024. LLMs are zero-shot context-aware simultaneous translators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1207, Miami, Florida, USA. Association for Computational Linguistics.

Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Complex knowledge base question answering: A survey. *IEEE Trans. on Knowl. and Data Eng.*, 35(11):11196–11215.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*.

Anna Lorincz, David Graus, Dor Lavi, and Joao Lebre Magalhaes Pereira. 2022. Transfer learning for multilingual vacancy text generation. In *Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 207–222, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

Huy Nghiem, John Prindle, Jieyu Zhao, and Hal Daumé Iii. 2024. "you gotta be a doctor, lin" : An investigation of name-based bias of large language models in employment recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7268–7287, Miami, Florida, USA. Association for Computational Linguistics.

Naoki Otani, Nikita Bhutani, and Estevam Hruschka. 2025. Natural language processing for human resources: A survey. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 583–597, Albuquerque, New Mexico. Association for Computational Linguistics.

Pooja S. B. Rao, Laxminarayen Nagarajan Venkatesan, Mauro Cherubini, and Dinesh Babu Jayagopi. 2025. Invisible filters: Cultural bias in hiring evaluations using large language models.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In

Proceedings of the Eighth Conference on Machine Translation, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Federico Retyk, Hermenegildo Fabregat, Juan Aizpuru, Mariana Taglio, and Rabih Zbib. 2023. Résumé parsing as hierarchical sequence labeling: An empirical study. In *Proceedings of the 3rd Workshop on Recommender Systems for Human Resources (RecSys in HR 2023) co-located with the 17th ACM Conference on Recommender Systems (RecSys 2023), Singapore, Singapore, 18th-22nd September 2023*, volume 3490 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Gemma Team. 2025. Gemma 3 technical report.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Xiao Yu, Ruize Xu, Chengyuan Xue, Jinzhong Zhang, Xu Ma, and Zhou Yu. 2025. ConFit v2: Improving resume-job matching using hypothetical resume embedding and runner-up hard-negative mining. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12775–12790, Vienna, Austria. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

12. Language Resource References

2020. *European skills, competences, qualifications and occupations – ESCO annual report 2019*. Publications Office.

Þórunn Arnardóttir, Elías Bjartur Einarsson, Garðar Ingvarsson Juto, Þorvaldur Páll Helgason, and Hafsteinn Einarsson. 2025. WikiQA-IS: Assisted benchmark generation and automated evaluation of Icelandic cultural knowledge in LLMs. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 64–73, Tallinn, Estonia. University of Tartu Library, Estonia.

Gasco, Luis and Fabregat, Hermenegildo and García-Sardi na, Laura and Estrella, Paula and Deniz, Daniel and Rodrigo, Alvaro and Zbib, Rabih. 2025. *Overview of the TalentCLEF 2025: Skill and Job Title Intelligence for Human Capital Management*.

Yuxin Luo and Feng Lu and Vaishali Pal and David Graus. 2023. *Enhancing Resume Content Extraction in Question Answering Systems through T5 Model Variants*. CEUR-WS.org.

National Center for O*NET Development. *O*NET OnLine*. Retrieved October 14, 2025.

Jorge Saldivar and Anna Gatzioura and Carlos Castillo. 2025. *Synthetic CVs To Build and Test Fairness-Aware Hiring Tools*.

Skondras, Panagiotis and Zervas, Panagiotis and Tzimas, Giannis. 2023. *Generating Synthetic Resume Data with Large Language Models for Enhanced Job Description Classification*.

van Toledo, Chaïm and Schraagen, Marijn and van Dijk, Friso and Brinkhuis, Matthieu and Spruit, Marco. 2022. *Exploring the Utility of Dutch Question Answering Datasets for Human Resource Contact Centres*.

Xu, Weijie and Huang, Zicheng and Hu, Wenxiang and Fang, Xi and Cherukuri, Rajesh and Nayyar, Naumaan and Malandri, Lorenzo and Sengamedu, Srinivasan. 2024. *HR-MultiWOZ: A Task Oriented Dialogue (TOD) Dataset for HR LLM Agent*. Association for Computational Linguistics.

Zhang, Mike and Jensen, Kristian and Sonniks, Sif and Plank, Barbara. 2022. *SkillSpan: Hard and Soft Skill Extraction from English Job Postings*. Association for Computational Linguistics.

A. Appendices

A.1. Controlled Attribute Categories and Placeholders

Table 5 provides the attribute categories with the most frequent placeholders and their associated potential bias dimensions, extracted from the English data, along with bias-related dimensions they might impact. For the other languages, we translate them to ensure parallelism, as described in Section 4.4.

Attribute & Bias	Top Placeholders
Attribute Category: PII Bias Dimensions: Demographic, geographic, socioeconomic and privacy	[EMAIL], [PHONE], [CITY], [STATE], [COUNTRY], [NAME], ...
Attribute Category: Affiliation Bias Dimensions: Prestige, socio-educational, and domain	[COMPANY], [SCHOOL], [ORGANIZATION], [UNIVERSITY], ...
Attribute Category: Professional Context Bias Dimensions: Core job qualifications, prestige and domain	[PLATFORM], [POSITION], [SUPERVISOR], [TEAM], [CERTIFICATION], [LICENSE], [AWARD], [PRODUCT], ...

Table 5: Controlled attribute categories with most frequent placeholders (ordered by frequency) and potential associated bias dimensions for the English dataset.

Table 6 also shows the frequency distribution of the most frequent top 10 placeholders across the English dataset, reflecting the different types of information typically found in résumés versus job descriptions, with contact and organizational details being common to both, while personal identifiers and educational background are specific to résumés.

A.2. Translation Post-Editing Edit Counts

Table 7 reports total edit counts after the selective post-editing step, split by language and dataset field. Résumés required the most editing (1,821-2,776 edits), with Spanish and Italian needing the most corrections. Job descriptions required fewer edits (368-612), while QA fields needed minimal corrections, particularly explanations (23-142 edits). These counts reflect the varying complexity of each field, with longer, more complex fields like résumés and JDs naturally requiring more post-editing to ensure quality and localization accuracy across languages.

Rank	Placeholder	Total	Résumé	JD
1	[EMAIL]	200	103	97
2	[COMPANY]	188	91	97
3	[PHONE]	180	99	81
4	[CITY]	136	102	34
5	[STATE]	120	93	27
6	[COUNTRY]	116	63	53
7	[NAME]	105	105	0
8	[SCHOOL]	90	90	0
9	[ZIPCODE]	74	74	0
10	[ADDRESS]	54	54	0

Table 6: Top 10 most frequent placeholders in the English dataset across résumés and job descriptions.

Field	de	es	it	zh
resume	2494	2776	2754	1821
jd	612	500	543	368
short_answer	267	137	278	315
explanation	142	62	110	23

Table 7: Total edit counts from manual post-editing and automated post-processing per language and dataset field.

A.3. MQM Categories Definition

We employed Multidimensional Quality Metrics (MQM) (Lommel et al., 2013) for human annotation of translation errors, adapting the taxonomy to the specific context of résumé and JD translation. The main categories and error types are summarized in Table 8, with detailed descriptions provided in the table. Each identified error is further classified according to its severity level:

- **Critical**: errors that render the content unfit for purpose or pose a risk for serious harm.
- **Major**: errors that seriously affect the understandability or usability of the content due to significant meaning changes.
- **Minor**: errors that do not seriously impede the usability or understandability but impact accuracy, consistency, or fluency.
- **Neutral**: cases where the evaluator would prefer a different translation but the current translation.

Category	Error Type	Description
Terminology	Inconsistent terminology	The target contains multiple terms used for the same concept.
	Wrong term	Use of term that is not what a domain expert would use or creates a conceptual mismatch.
Accuracy	Mistranslation	The target content does not accurately represent the source content.
	Addition	The target includes content not present in the source, it was translated when it should not have been, or is overly specified.
	Omission	The target is missing content present in the source, is not translated when it should have been, or is oversimplified.
Linguistic	Grammar	A text string in the translation violates the grammatical rules of the target language.
	Punctuation	Punctuation incorrect according to target language conventions.
	Spelling	Error occurring when a word is misspelled.
Style	Inconsistent style	Style that varies inconsistently throughout the text.
	Gender inclusive	Output should include both feminine and masculine forms using appropriate target language conventions (e.g., “des/der Kandidaten/-in”, “el/la candidato/a”, “del/la candidato/a”).
Locale	Entity format	Format for entities such as numbers, date, time, currency, address, etc. is wrongly rendered.
Design	Layout	Errors related to the physical design or presentation of the translation.
Custom	Hallucination	Parts of the target content are completely decoupled from the input sentence.

Table 8: MQM (Multidimensional Quality Metrics) error categories and error types used in translation evaluation.

A.4. Impact of Question Type on QA Performance

Figure 4 shows G-Eval score distributions stratified by question type across all nine models and five languages. Question types follow the two-dimensional taxonomy in Table 1, resulting in four categories: Cross-Document Single-Hop (CD-SH), Cross-Document Multi-Hop (CD-MH), In-Document Single-Hop (ID-SH), and In-Document Multi-Hop (ID-MH). The variability in scores across question types aligns with the large standard deviations observed in the main results (Table 4).

The figure reveals patterns in question-type difficulty that vary by model capability and language. For large models (> 8B parameters), results indicate that single-hop questions (CD-SH and ID-SH) tend to have tighter IQRs and higher medians in English and Spanish, suggesting more predictable performance on question requiring direct retrieval in high-resource languages. Multi-hop questions (CD-MH and ID-MH) generally exhibit wider distributions, though this effect appears more pronounced in some language-model combinations than others (notably DE and ZH). For medium-sized models (4B–8B parameters), the stratification between single-hop and multi-hop performance becomes less clear. While some models show a visible gap in distribution shapes, the IQR widths become more comparable across question types. For small models (< 4B parameters), results suggest that question-type distinctions have minimal bearing on overall performance, with all four question types exhibiting uniformly wide distributions and low medians across all languages, consistent with the factually incorrect band in Table 4.

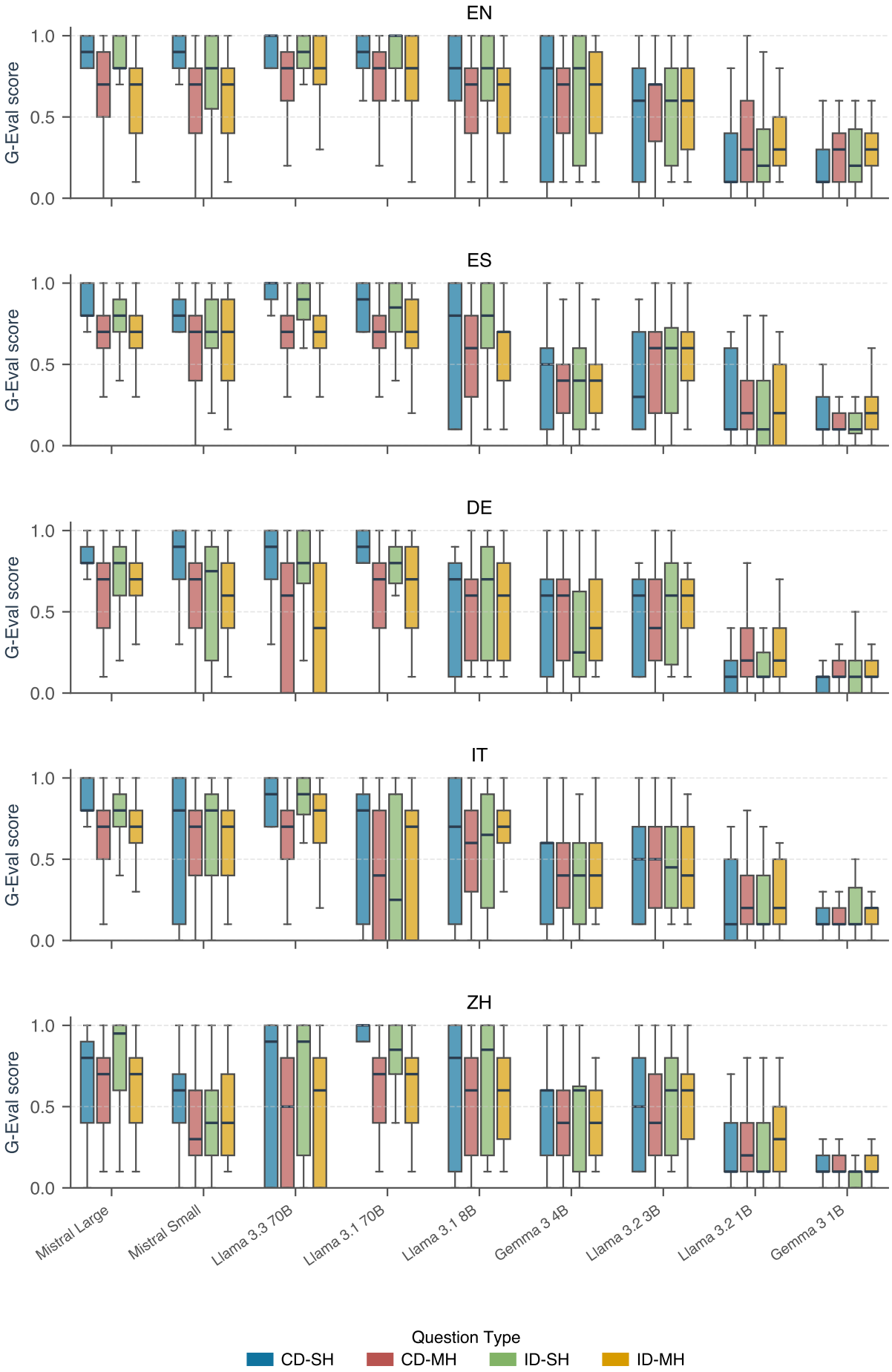

Figure 4: G-Eval score distributions by question type (CD-SH, CD-MH, ID-SH, ID-MH), model, and language. Boxes show the interquartile range (IQR), horizontal lines indicate the median and whiskers extend to $1.5\times$IQR.