Implementation and Usability Evaluation of an LLM-Based Résumé Builder and Interview Preparation Platform for Diverse Career Groups

Mogboluwaga Tayo Otegbayo
Doyenify OÜ, Estonia
Wire IT, Portugal
tayootegbayo@gmail.com

Olumide Oluyide
Doyenify OÜ, Estonia
gbengaoluyide@gmail.com

Ayomide Victoria Adeagbo
Doyenify OÜ, Estonia
vayomide2@gmail.com

Solagbade Enitilo
Doyenify OÜ, Estonia
solagbadeenitilo@gmail.com

Abstract— **Diverse career groups such as new graduates, job seekers, career switchers, and academics submit their résumés to screening systems that give preference to résumés that match a job description, but many new grads don't have experience writing résumés or experience in interviewing. In this paper, we present the design, implementation, and pilot usability testing of a web platform that generates résumés from templates and employs a large language model (LLM) to propose a summary and up to 20 interview questions specific to a job description. Requirements were derived from a questionnaire completed by job seekers and respondents on the recruiting side. It features a React and TypeScript client, a Node.js server, Firebase authentication and storage, and client-side PDF export. Ten participants finished the evaluation tasks and the System Usability Scale (SUS). No per-participant records were kept in analyzable form, and intervals were calculated as sensitivity analyses. Five of the nine people who responded to the ten questionnaire rows preferred a hybrid format (55.6%, 95% CI 26.7 to 81.1). 7/9 traced requirements were implemented. The mean SUS score was 77.0, and the reference mean of 68 was not included in the 95% interval only if the score standard deviation was less than 12.58 points. Task completion was 77.5%, which was below the 85% goal, depending on the unreported number of task attempts (exact p = 0.184 to 0.029). To achieve a 5% error rate and 22.5% task failure rate, there had to be at least 4.5 error opportunities per attempt. There is acceptable pilot usability evidence. Does not make any claims regarding output quality, applicant tracking system performance or hiring benefit.**

Keywords—résumé generation; interview preparation; large language models; usability evaluation; software design; fresh graduates.

I. Introduction

The transition from University to work is a screening issue. According to a 2024 mixed-methods study at the University of Lagos, the employability problem for graduates in Nigeria stems from a mismatch between the skills universities provide and those required by the economy, and approximately 1.7 million graduates enter the labor market annually [1]. The graduates in this position face competition from those with limited professional writing experience and little practice in structured interviews.

More and more screening is automated. In the United States, the United Kingdom, and Germany, over 90 percent of employers said they use recruiting software to filter or rank candidates, and 88 percent said that qualified candidates are filtered out when they don't meet the exact criteria of the job description [2]. A résumé needs to be readable and relevant to the posting. The quality of writing is important in and of itself. In a field experiment involving almost a million job seekers, non-generative writing support on resumes increased the likelihood of being hired by 8 percent [3].

Large language models (LLMs) offer a low-cost route to this assistance. Research prototypes already exist that customize a résumé to a job posting [4] and simulate or perform an interview [5], [6]. The same models also generate text [7], [8], perpetuate demographic bias in recruitment processes [9], [10], and make software engineering more difficult because prompts change over time without systematic documentation or testing [11]. A résumé is a document signed by the applicant, meaning any errors made by the résumé generator are the applicant's.

Student-created career platforms usually include a list of features and a usability rating [12], [13]. They do not typically map features to elicited requirements, report uncertainty in small-sample results, or report which components depend on the LLM. This paper fills that void for one particular system: a web-based system created at Obafemi Awolowo University to create résumés and prepare users for the interview process. It fully describes its requirements, design, and implementation, and reanalyses the evaluation data gathered for the original project with explicit intervals and sensitivity analyses.

The study is guided by three research questions.

RQ1. What résumé formats and sections did the respondents to the questionnaire value?

RQ2. How well are the requirements elicited in the platform implemented?

RQ3. How usable were the tasks and how well did the pilots perform them, and how uncertain are these results?

The paper has three contributions. First, it describes a requirements-driven design and implementation of an LLM-based résumé and interview platform, from architecture and

behavioral models to the entire prototype. Secondly, it presents the pilot findings in intervals and indicates the extent to which each conclusion relies on quantities not recorded in the original records. Thirdly, it offers error analysis and a logging protocol that enable the next evaluation to be reproduced. The study does not compare the quality of LLM output, the impact of the platform on the performance of applicant tracking systems, or hiring results. The related work is reviewed in Section II, the method is described in Section III, the design and implementation are reported in Section IV, the results, discussion, and threats are reported in Sections V, VI, and VII, respectively, and the conclusion is presented in Section VIII.

II. Background and Related Work

A. Résumé builders before large language models

Initial systems for resumes automated layout and form filling. Xie et al. presented a natural language processing-based builder that structures the user input into a structured document [14]. Tyagi et al. reported an Android application with one-time password login that exports PDF files [15], and Mhatre et al. integrated a builder with a portfolio feature and knowledge-base-assisted ranking [16]. Kungwani et al. proposed a form-based web application for creating a résumé [17]. Kanyal et al. introduced speech recognition for input [18], Yi and Fang introduced augmented reality markers [19] and Malhotra et al. described a website with video presentation features [20]. Kaur and Singh used machine learning and natural language processing to tailor résumés [21], Tejaswini et al. developed a machine learning résumé ranking algorithm [22] and Mgarbi et al. suggested offering jobs from a candidate's résumé [23]. The usability of a website has been found to influence candidates' perceptions of the organisations and a qualitative study of interactive résumés has revealed the pros and cons of web-based formats.

These systems set the groundwork for what template selection, structured input and PDF output will be. There is no cited paper that pairs builder with job-description-conditioned interview preparation, and the surveyed work predates LLM-based generation.

B. Large language models for résumé tasks

Recent studies explore how to use LLMs for three resume-related tasks. To extract, Zhu et al. propose a layout aware parser, an efficient LLM extractor, and a two-stage automated evaluation, as a result of the diversity of résumé formats and the lack of annotated data [24]. To generate, ResumeFlow receives a resume and a job post and uses GPT-4 or Gemini to generate a customized resume without fine tuning. Its authors also propose task specific metrics for alignment and hallucination [4]. Audits demonstrate that models used for résumé screening disadvantage some groups for assessment. In nine occupations, Wilson and Caliskan collected more than 500 résumés and more than 500 job descriptions and discovered that résumés containing black male names were disadvantaged in as many as 100 per cent of the comparisons [10]. Armstrong et al. expanded the audit to résumé generation and found that GPT-3.5 generated fewer experienced occupations for women's résumés and immigrant markers for Asian and Hispanic names [9].

There is a taxonomy of generation errors. The aforementioned 2026 preprint about résumé rewriting outlines anachronistic technology injection, cross-domain terminology contamination, structural mutation, and content fabrication. Across 25 synthetic résumés, undefended baselines yielded 2.48-5.36 hallucinations per résumé, while layered defenses reduced the rate [25]. This is a synthetic study and a preprint, but it demonstrates that fabrication is measurable and not hypothetical.

C. Interview preparation systems

LLMs are not the first to use research on interview practice. MACH developed a virtual coach that acted on the user's actions [26] and, more recently, recorded interviews were analyzed to predict performance, and verbal answers were analyzed on a mock interview platform. Mock interview generators are based on the resume and job description. EZInterviewer was evaluated by three human raters for the meaningfulness, usefulness, relevance, and coherence of the generated questions [27], and Qin et al. conducted research on skill-oriented question generation and recommendation [28].

LLM based systems include adaptive dialogue. Conversate uses GPT-3.5 Turbo to ask follow up questions and GPT-4 to give dialogic feedback. In an experiment with 19 users, the follow-up questions were appreciated, and the authors noted that the model was more likely to agree with users who disagreed with its feedback [5]. PolyInterview bases questions on the job description and CV, has a digital interviewer for spoken interviews, and has expert ratings of 4.62/5 for question plans, covering 1,564 sessions [6]. Both papers emphasize that the evaluation of the LLM output is an open work, and this is the reason for the cautious scope of the present paper.

Evidence on practice is more dated but more consistent. Situational interviews have a correlation with coaching, and structured interviews are a well established selection method [29]. Applicants who interact with an AI interviewer report having some difficulty communicating with the AI which is another reason for a low stakes practice setting. In addition, interview platforms that accept free text job descriptions also provide an input channel for prompt injection, which multi agent designs like CoMAI deal with by having a dedicated security agent [30].

D. Reliability, bias, privacy and engineering of LLM applications

The main reliability risk is hallucination [31]. It has been surveyed by Ji et al. for natural language generation [8] and by Huang et al. for distinguishing factuality errors from faithfulness errors of LLM [7]. The generated text can also taint downstream information and models can perpetuate gender stereotypes. A survey of LLM security and privacy catalogues leakage and misuse threats that apply to systems

storing personal data [32], [33]. The opportunities and risks of educational use of LLMs are similar.

Engineering practice lags behind these concerns. A study of 1262 changes across 243 repositories revealed that only 21.9% included documentation in commit messages [11]. There is also a role for regulation. The European Union Artificial Intelligence Act classifies systems that are used for recruiting or selecting individuals as high-risk systems [34]. The platform examined here is candidate-facing, and its legal status would have to be evaluated separately. The Data Protection Act 2023 in Nigeria imposes duties on processing personal data [35].

E. Measuring usability

The SUS is a ten-item questionnaire that gives a score between 0 and 100 [36]. It is the most widely used standardised usability questionnaire [37], and there are large data sets to interpret it. The average of the reference data is 68, and anchors a curved grading scale [38]. Adjective anchors are located at the "good" and "excellent" extremes, at scores around 71 and 85, respectively [39]. SUS provides more stable results with small sample sizes but the width of the interval is still dependent on the dispersion of scores, which cannot be determined exactly in small studies [38].

F. Synthesis and positioning

The comparison of the closest works of the LLM era with the platform studied here is given in Table I. Tailoring of resumes and interview simulation/hallucination control have been examined, but individually and primarily as prototypes, preprints or demonstrations. The present platform includes a résumé building and interview preparation in one product, job descriptions for generating questions and a usability evaluation. It lacks the ability to provide answer feedback or answer fabrication control and its LLM output is never evaluated. This paper's contribution is therefore a documented, traceable design with a clear and honest statement of what it can and cannot support in an evaluation, rather than a new-generation method.

TABLE I. COMPARING THE MOST SIMILAR SYSTEMS IN THE LLM ERA TO THE SYSTEM STUDIED HERE.

System	Focus	LLM	Interview practice	Job description input	Reported evaluation
ResumeFlow [4]	Job tailored résumé generation	Yes	No	Yes	Demonstration; proposed alignment and hallucination metrics
Grounded optimization [25]	Hallucination defences in résumé rewriting	Yes	No	Yes	Ablation on 25 synthetic résumés (preprint)
Layout aware parsing [24]	Résumé information extraction	Yes	No	No	Benchmarks and automated evaluation (preprint)
EZInterviewer [27]	Mock interview generation	Neural generator	Yes	Yes	Automatic metrics; three annotators
Conversate [5]	Interview simulation with dialogic feedback	Yes	Yes	Job title only	User study, 19 participants
PolyInterview [6]	Multimodal mock interviews	Yes	Yes	Yes, with CV	Ten expert ratings; usage snapshot (preprint)
This platform	Résumé builder with LLM summary suggestions and interview questions	Yes, model not reported	Yes, text questions	Yes	SUS pilot, n = 10

The comparison is based on the published abstracts and descriptions of each work. Entries marked preprint are not peer reviewed.

III. RESEARCH METHOD

A. Study design

The study adopts a design and development method that is supported by design science, which is a method that builds an artefact to solve a problem and then evaluates the artefact [40]. The development was done using Agile Scrum method and short sprints and user feedback. Fig. 1 shows the sequence of activities and the research question each addresses. The original project report supplies all data [41]. The present paper reanalyzes the questionnaire and pilot data explicitly with uncertainty, and fully describes the design.

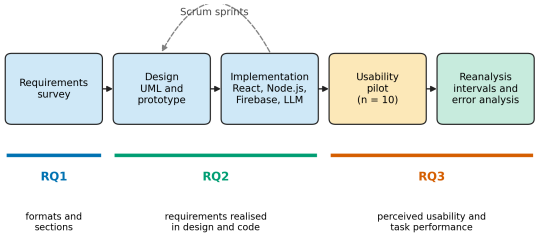

Fig. 1. Study design. The project activities are in blue boxes, orange box is the pilot, green box is the reanalysis reported here. The activities that provide evidence for RQ1-RQ3 are mapped below to the respective questions.

B. Requirements elicitation

A Google Forms survey was created to solicit feedback on their role, industry, type of résumé they were looking for, the sections they felt were necessary, and their preferred format. Hiring managers, recruiters and job seekers were respondents. No count of the number of respondents was provided in the report available. A few free text cells have been shortened in the source, so section counts are lower bounds. The ten rows reproduced in the report are transcribed in Appendix A.

C. Pilot usability evaluation

The four task groups were conducted with ten participants from the target group of fresh graduates and job seekers, who completed all four task groups. They then filled out the ten item SUS and gave their satisfaction rating on a scale of 1 to 10. No record of the number of tasks, setting, device and moderation procedure. The evaluation protocol was based on five measures. They are listed in Table II with the reference values used in this paper.

TABLE II. USABILITY MEASURES, DEFINITIONS AND REFERENCE VALUES.

Measure	Definition	Reference value
System Usability Scale (SUS)	Ten items, 5 point scale, scored 0 to 100 (Equation 1)	Mean of 68 in reference data [38]
Task completion rate (TCR)	Share of attempted tasks completed successfully (Equation 2)	Target above 85%, set in the evaluation protocol
Error rate (ER)	Errors divided by opportunities for error (Equation 3)	Target below 10%, set in the evaluation protocol
Time on task (ToT)	Time to complete the core tasks, in minutes	Target below 10 minutes, set in the evaluation protocol
Satisfaction score (USS)	Single rating from 1 to 10	Target above 7, set in the evaluation protocol

Other targets, other than the SUS reference, are not independently validated and are pre specified by the original protocol.

$$SUS = 2.5 \times [\ \Sigma(s_i - 1) \text{ over odd items} + \Sigma(5 - s_i) \text{ over even items}\ ] \quad (1)$$

$$TCR = 100 \times N_success / N_attempted \quad (2)$$

$$ER = 100 \times N_error / N_opportunity \quad (3)$$

In equation 2, the denominator is the number of attempted tasks. Time on task is reported as the arithmetic mean (as in the original analysis), but a geometric mean is preferred for skewed times [38].

D. Statistical analysis

Scores for each participant were not saved in an analyzable format and therefore a standard deviation could not be calculated. Intervals are then sensitivity analyses of the reported means. The t-based 95% intervals were calculated for standard deviations of 8, 10, 12.5, 15 and 17.5 points, with 12.5 being very close to the dispersion of SUS data [38]. Standard deviations for time on task and satisfaction were 1.5-3.0 minutes and 1.0-2.0 points, respectively. The 77.5% completion rate requires 40 attempts, and for 31 of 40, 62 of 80 and 93 of 120 opportunities, Wilson score intervals [42] were calculated; for 2 of 40, 4 of 80 and 6 of 120 opportunities, error rates were calculated. Completion was compared to the 85% target using exact binomial tests. Cohen's d and the power of the one sample t test were defined as usual. Fisher's exact test was used to compare groups in an exploratory manner.

No multiplicity correction is applied as there is no confirmatory claim. The p values indicate how conclusions will vary if the quantities are not reported and do not indicate significance. Calculations used Python 3.12.3 with SciPy 1.17.1 and statsmodels 0.15.0. As the source data are an extract, no interval is reported for the requirement coverage in RQ2, which is a census of nine listed requirements.

IV. SYSTEM DESIGN AND IMPLEMENTATION

A. Requirements

The design was guided by nine requirements [43]. They are listed in the project report in Table III, along with their source. Requirements R1 - R3 are derived from the questionnaire, R4 - R8 are derived from the aim, scope and design, and R9 is derived from the aim of the project.

TABLE III. PROJECT REQUIREMENTS AND SOURCES IN THE PROJECT REPORT.

ID	Requirement	Source
R1	Provide the sections respondents rated essential: work experience, skills, certifications and education	Questionnaire; Section 3.3
R2	Offer several templates with preview and confirmation	Questionnaire; Section 3.3
R3	Support a hybrid résumé format	Questionnaire; Section 3.3
R4	Show a live preview and progress while the user types	Design objective; Section 3.4
R5	Export the résumé as PDF	Problem statement; Section 1.2
R6	Suggest professional summaries with an LLM	Aim; Section 1.3
R7	Generate up to twenty job description specific interview questions	Aim and scope; Sections 1.3 and 1.5
R8	Provide accounts with email verification and Google sign in	Design and prototype; Section 3.5
R9	Import profile data from social or Google accounts	Aim; Section 1.3

B. Architecture

Fig. 2 shows the architecture. The six step wizard, the live preview and the interview screen are rendered by a React and TypeScript client. A Node.js server is used to expose a REST interface and to process requests to the LLM service. Firebase provides authentication and the database for users, résumés, templates and interview responses. The PDF is created by the client using html2pdf.js. The report doesn't specify if the database is accessed through the server or not, so both paths are shown.

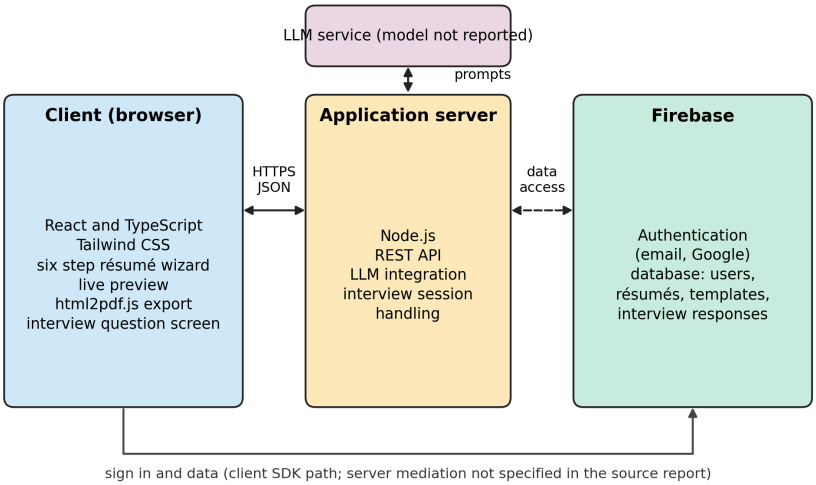

Fig. 2. Platform architecture. Documented interfaces are represented by solid arrows. The dashed arrow represents a data access path that is not defined in the project report. The provider and the model of the LLM are not disclosed.

C. Behavioural and structural design

The user flow is shown in Fig. 3. Users sign up, confirm their email, log in via email or Google, and are taken to a dashboard from which they can choose to build a résumé or to move on to the interview module. The Use Case Model is presented in Fig. 4. The Job Seeker conducts 8 use cases, and the LLM service assists with providing a summary suggestion and question generation. Administrator use cases were defined in the design, but not in the prototype evidence.

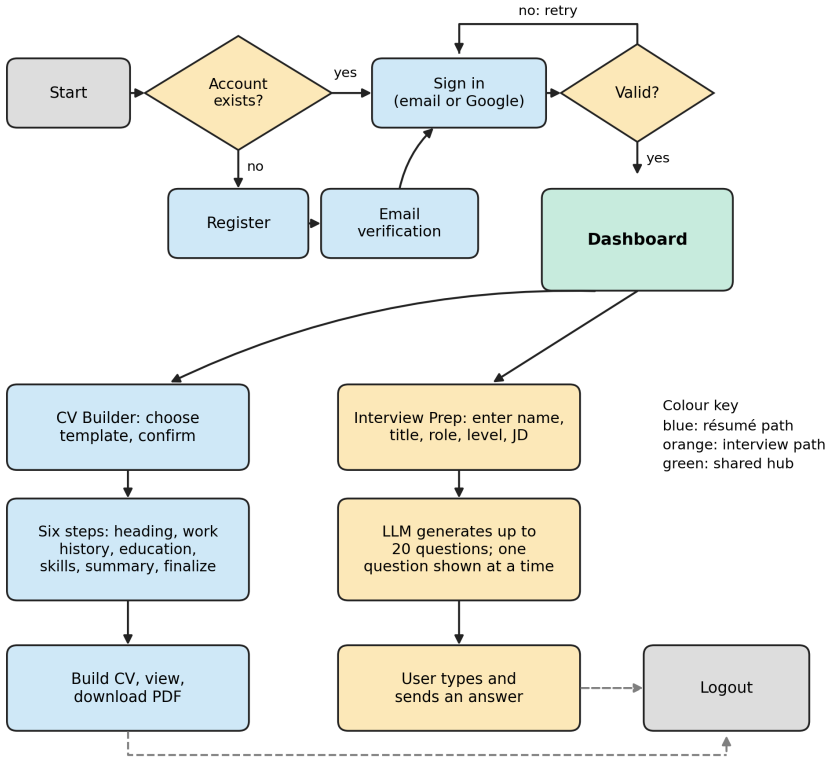

Fig. 3. User flow. Blue is for resumé path, orange for interview path, and green for shared dashboard. Dashed grey lines lead to logout.

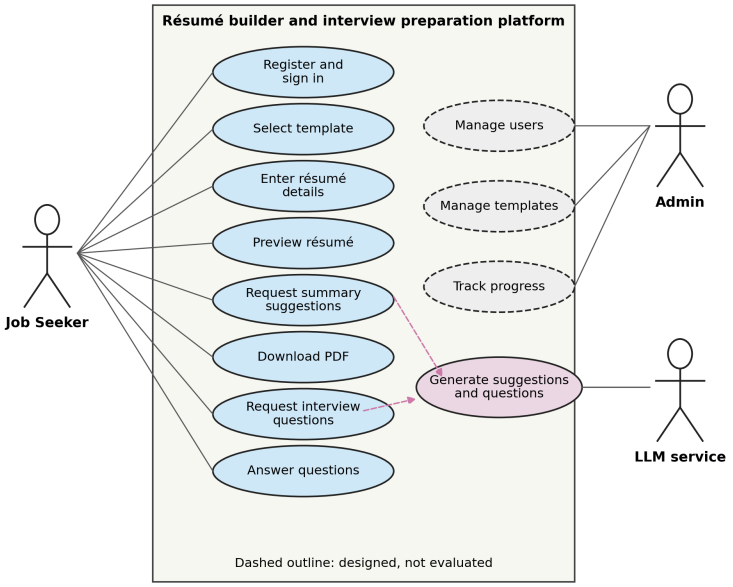

Fig. 4. Use case model. Dashed ellipses indicate administrator functions that were designed but not used in the evaluation. Dashed purple arrows indicate the location of the LLM service calls.

The class model is depicted in Fig. 5. The design specification has four classes: User, Resume, Admin and Template. The prototype needs two additional components: an interview session with title, role, level, job description, questions and responses, and an LLM service. These were deduced from the prototype and from the statement that interview answers are saved. Generation methods are not part of the Admin class but rather the LLM service.

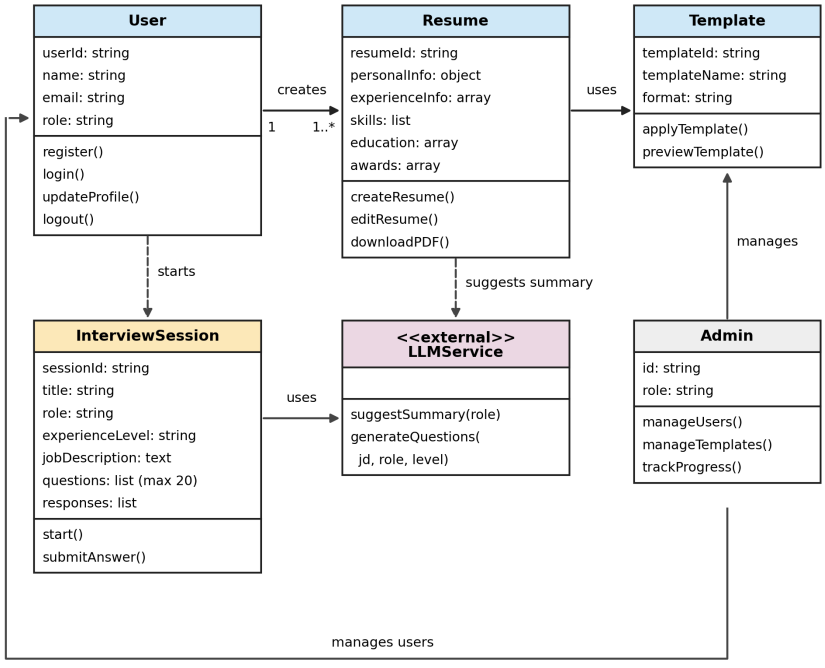

Fig. 5. Class model. The blue classes are designed based on the design specification, the orange class and the purple external service are deduced from the prototype, and the grey class represents the role of administrator.

The message flow for sign in, résumé building, summary suggestion, PDF export and interview preparation is depicted in Fig. 6. The sequence is drawn from the design specification and prototype behaviour. Prompt contents are not reported, and the messages sent to the LLM service name only contain its inputs.

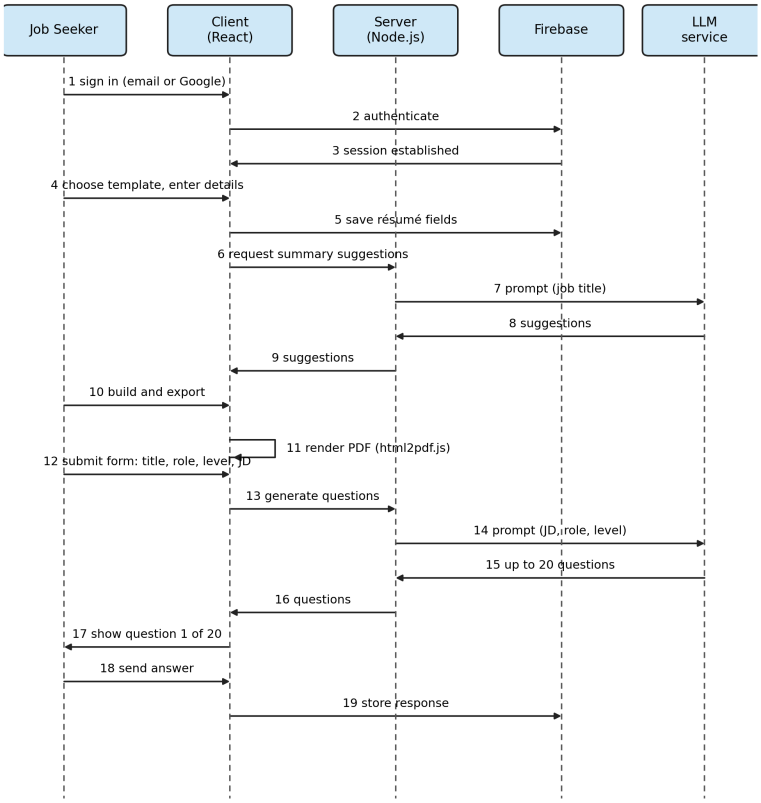

Fig. 6. Flow of interactions between the job seeker, client, server, Firebase and LLM service. The message order is indicated by numbers.

D. Implementation

The project report is summarised in Table IV with regard to implementation and what it does not include. The omissions concern the reproducibility and are included to be completed. The interview module can take a first name, last name, title, level of experience, role and a pasted job description, and returns a set of at most twenty questions, one at a time.

TABLE IV. ITEMS REPORTED IN THE PROJECT REPORT AND NOT REPORTED.

Component	**Reported specification**
Client	React, TypeScript and Tailwind CSS; responsive layout
Interface prototyping	Figma
Server	Node.js with a RESTful API
Authentication	Firebase Authentication with email verification; Google sign in visible in the prototype
Data store	Firebase; templates, résumés, interview responses and preferences
PDF export	html2pdf.js, client side
Process	Agile Scrum; sprint log not reported
LLM provider, model, version, parameters	Not reported
Prompt templates and output format	Not reported
Library versions, hosting and deployment address	Not reported

E. Prototype

The authentication screens and the dashboard are shown in Fig. 7. Sign up disables the Create Account button until the terms are accepted, displays validation messages for blank fields, and sends a verification email. Sign in is for Google or email and a toast is displayed if the credentials are invalid. The dashboard shows the user's first name, a progress ring and the template gallery.

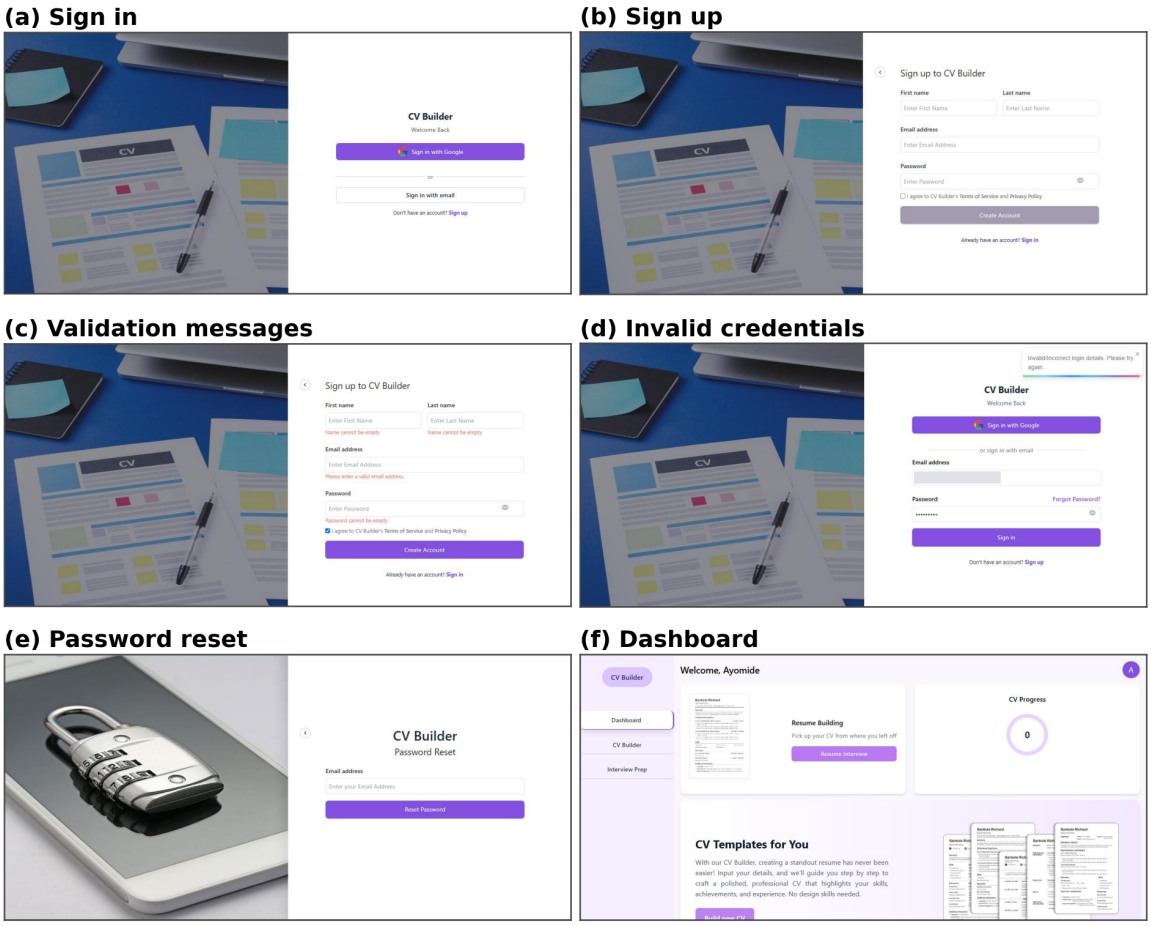

Fig. 7. Authentication and dashboard screens: (a) sign in, (b) sign up with disabled button, (c) validation messages, (d) invalid credentials, (e) password reset and (f) dashboard. Contact details of the test account are masked.

The résumé builder is displayed in Fig. 8 and Fig. 9. The user selects one of the three templates and clicks OK. The following 6 step wizard is followed: heading, work history, education, skills, summary, and finalize. A completeness bar and a real-time preview feature as fields are completed. Step 1 auto-fills the name and email based on the registration, and the work history step recommends role descriptions based on the typed job title. The summary step provides a list of suggested summaries that are provided by the report as LLM generated. The finalize step is to add optional certifications, awards, a website and languages. It will open in a preview window with download, finish and change template actions.

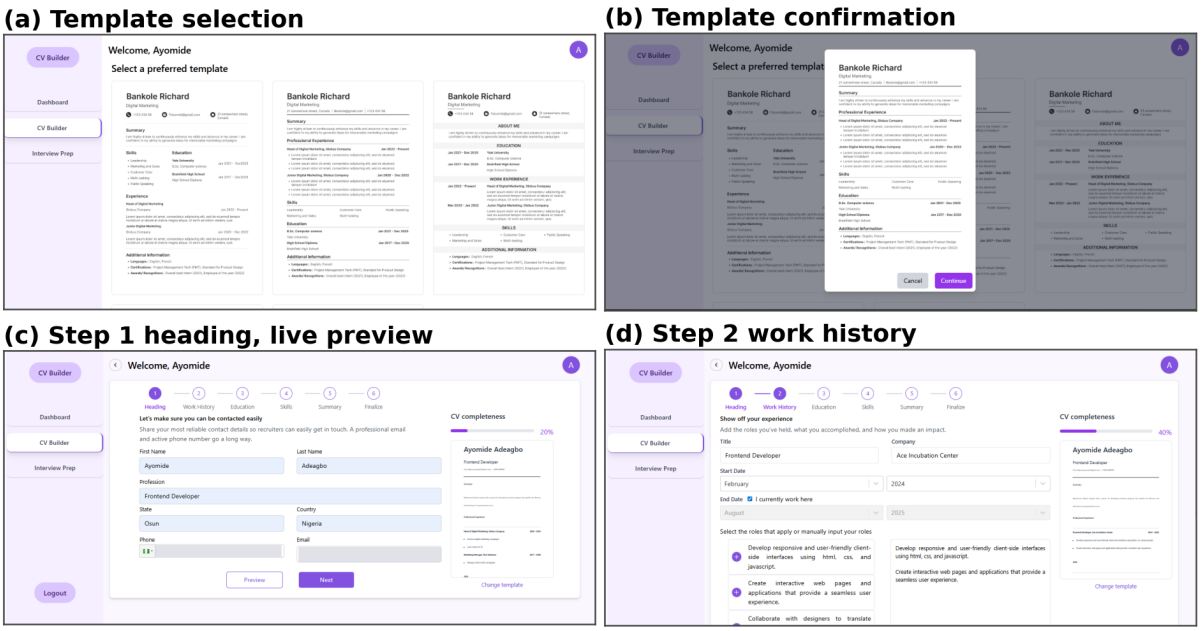

Fig. 8. Resume builder, (a) choose a template, (b) confirm, (c) step 1 with preview and completeness bar, (d) step 2 work history.

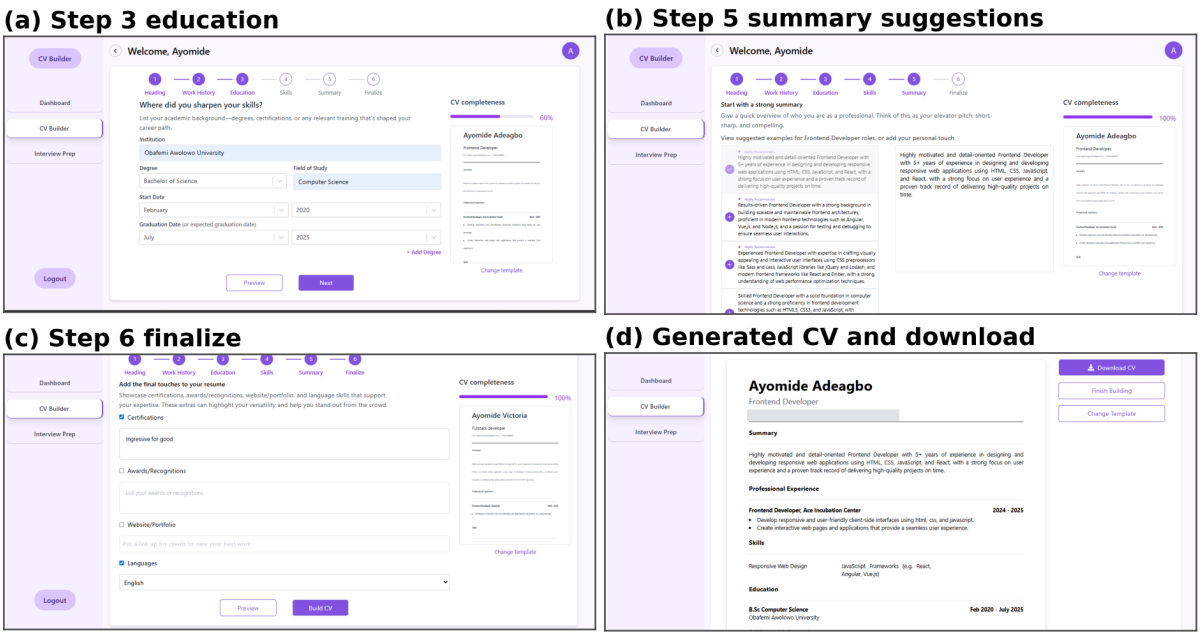

Fig. 9. Résumé builder, second part: (a) step 3 education, (b) step 5 summary suggestions, (c) step 6 finalize and (d) generated résumé with download action.

The interview module is presented in Fig. 10. On the entry page, there are personalized questions to prepare for an interview. The form has three sections: role, experience level and a job description, and the question page displays one of twenty questions with a text box and a Send button. The report is about the storage of responses and not about automated feedback on responses.

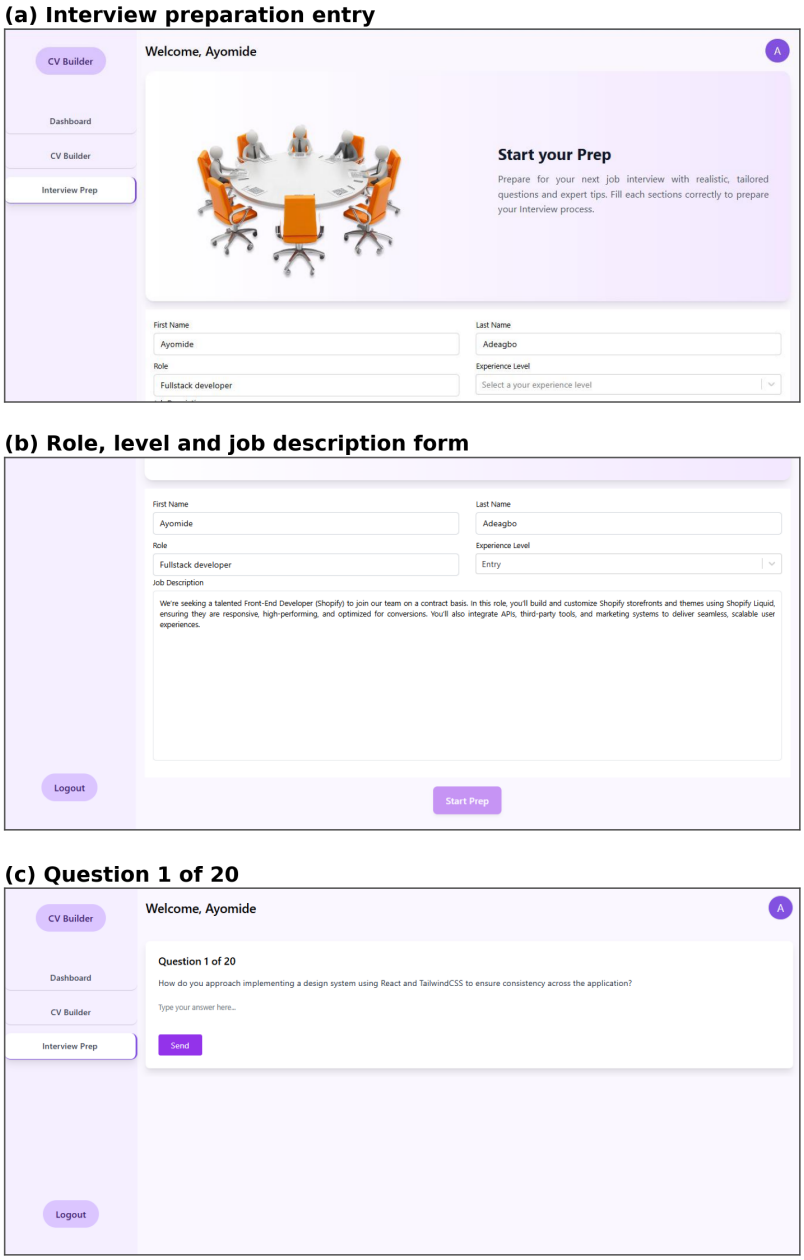

Fig. 10. Interview preparation: (a) entry page, (b) form with role, level and pasted job description, and (c) the first of 20 generated questions.

V. Results

A. Requirements questionnaire (RQ1)

Table V and Fig. 11 summarise the ten questionnaire rows. There were 5 job seekers, 2 recruiters, 1 hiring manager, 1 HR professional and 1 designer and student. Industries were information technology (four rows including one shared with healthcare), finance (two) and healthcare, education, hospitality and design (one each). Eight people required an industry résumé and two an academic résumé. Nine of the respondents indicated a format: five hybrid, three chronological, and one functional. A majority preference was not obtained with a hybrid share of 55.6% (wide interval).

At least eight rows were mentioned for work experience and at least seven for skills. There were at least six certifications, four with personal information, three with education information and one with technical projects. These lower bounds are obtained from the truncated cells. Four of five job seeker side respondents said they prefer hybrid, compared to one of four on the recruiting side. This difference was not significant (Fisher exact $p = 0.206$).

TABLE V. THE FOLLOWING IS AN EXTRACT FROM THE QUESTIONNAIRE INCLUDING THE RESPONDENT CHARACTERISTICS AND STATED PREFERENCES (TEN ROWS).

Characteristic	Count	Share (%)	95% Wilson interval (%)
Role: job seeker; recruiter; hiring manager; HR professional; designer and student	5; 2; 1; 1; 1	50; 20; 10; 10; 10	Not applicable
Résumé type: industry	8 of 10	80.0	49.0 to 94.3
Résumé type: academic	2 of 10	20.0	5.7 to 51.0
Format: hybrid (n = 9)	5	55.6	26.7 to 81.1
Format: chronological (n = 9)	3	33.3	12.1 to 64.6
Format: functional (n = 9)	1	11.1	2.0 to 43.5
Section named: work experience (lower bound)	8 of 10	80.0	49.0 to 94.3
Section named: skills (lower bound)	7 of 10	70.0	39.7 to 89.2
Section named: certifications (lower bound)	6 of 10	60.0	31.3 to 83.2
Section named: personal details (lower bound)	4 of 10	40.0	16.8 to 68.7
Section named: education (lower bound)	3 of 10	30.0	10.8 to 60.3
Section named: technical projects (lower bound)	1 of 10	10.0	1.8 to 40.4

One respondent did not provide a format. Only visible text is counted for sections.

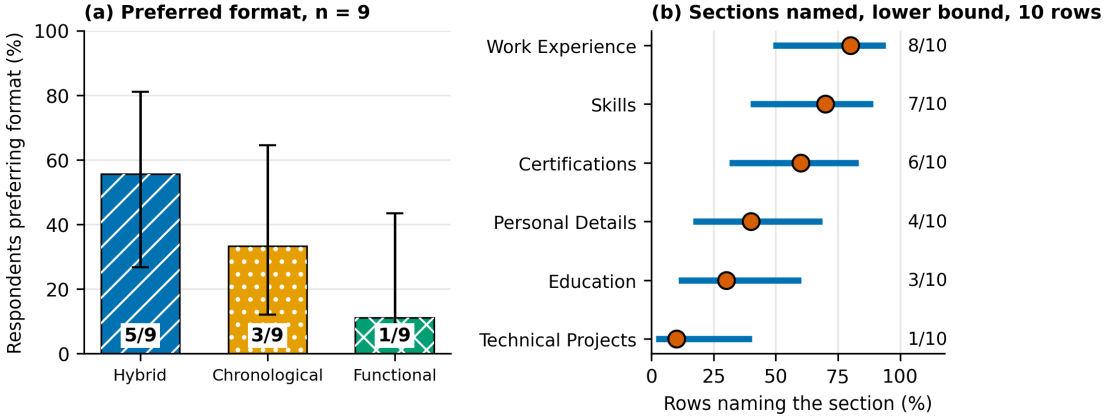

Fig. 11. Questionnaire extract. (a) The responses of nine people, showing bar patterns with 95% Wilson intervals, repeating the colours. The 95% Wilson intervals for the lower bounds of sections are shown as (b) Sections named in the ten visible rows.

Answer to RQ1. Respondents most commonly identified work experience and skills as important sections of the worksheet (at least 8 and 7 of 10 rows) and hybrid as the most frequent format (5 of 9, 55.6%, 95% CI 26.7 to 81.1). The size of the sample is not large enough to indicate a majority or difference between the responses of job seekers and recruiting side respondents.

B. Requirement realisation (RQ2)

Table VI gives the status of each requirement in Table III. Seven were implemented and shown in the prototype, one was not verifiable, and one was not implemented. The hybrid format is the unconfirmed one as there are three templates in the report which are not categorized for their structure. The social profile import has been removed. Google sign in provides only identity and there are no screens that show employment or education data being imported. Requirements are a census and therefore no interval is reported.

TABLE VI. WHAT EACH REQUIREMENT IS AND HOW IT IS SUPPORTED.

ID	Requirement	Status	Evidence
R1	Essential sections	Implemented	Wizard steps for work history, education and skills; finalize step for certifications (Figures 8 and 9)
R2	Template choice with preview	Implemented	Three templates and confirmation dialog (Fig. 8)
R3	Hybrid format	Not verifiable	Template structure not classified in the report
R4	Live preview and progress	Implemented	Preview and completeness bar (Fig. 8)
R5	PDF export	Implemented	Download action (Fig. 9); parsing by applicant tracking systems untested
R6	LLM summary suggestions	Implemented, output not evaluated	Suggested summaries (Fig. 9)
R7	Job description questions, up to 20	Implemented, output not evaluated	Form and question page (Fig. 10)
R8	Email and Google accounts	Implemented	Sign in, sign up, verification (Fig. 7)
R9	Social or Google data import	Not implemented	Google sign in gives identity only (Fig. 7)

Answer to RQ2. Seven out of the nine requirements were put in place and demonstrated. One couldn't be found in the available material and one, profile data import, wasn't implemented. The LLM supported requirements R6 and R7 are features, but not assessed for their quality.

C. *Usability and task performance (RQ3)*

Table VII presents the pilot results and sensitivity intervals, and Fig. 12 and Fig. 13 present these results graphically. The mean SUS score was 77.0. With a standard deviation of 12.5, the 95% interval was 68.1 to 85.9, with the lower bound of the interval being greater than 68 only when the standard deviation was less than 12.58 points (Fig. 13). The power of the test against 68 was 0.72, 0.53 and 0.40 at standard deviations of 10, 12.5 and 15. To detect the observed difference with 80% power, 12, 18, and 24 participants were required for the three sets of eyes, respectively.

Task completion was 77.5%. The exact p-values were 0.184 at 40 trials, 0.082 at 80, and 0.029 at 120 trials, indicating the result changed as the denominator increased—but the denominator was not reported. This error rate was 5.0%, with the upper interval bound exceeding the 10% target across all denominators (16.5%, 12.2%, and 10.5%). The mean time on task was 6.9 minutes, and the UTT did not exceed 10 minutes for all assumed dispersions. Satisfaction was 7.1, with the interval ranging from 5.7 to 8.5.

TABLE VII. USABILITY RESULTS FOR PILOTS (N = 10) WITH SENSITIVITY INTERVALS.

Measure	Reported	Reference	95% interval: favourable to conservative	Test against reference	Effect size
SUS	77.0	68	69.8 to 84.2 (SD 10); 66.3 to 87.7 (SD 15)	t = 2.85, p = 0.019 (SD 10); t = 2.28, p = 0.049 (SD 12.5); t = 1.90, p = 0.090 (SD 15)	d = 0.90 to 0.60
TCR (%)	77.5	85	69.2 to 84.1 (120 attempts); 62.5 to 87.7 (40)	Exact binomial p = 0.029 (120); 0.082 (80); 0.184 (40)	Difference of 7.5 points
ER (%)	5.0	10	2.3 to 10.5 (120 opportunities); 1.4 to 16.5 (40)	Not tested; upper bound at or above 10	Not applicable
ToT (min)	6.9	10	5.8 to 8.0 (SD 1.5); 4.8 to 9.0 (SD 3.0)	p = 0.0001 (SD 1.5); 0.0097 (SD 3.0)	d = 2.07 to 1.03
USS	7.1	7	6.4 to 7.8 (SD 1.0); 5.7 to 8.5 (SD 2.0)	p = 0.76 (SD 1.0); 0.88 (SD 2.0)	d = 0.10 to 0.05

Intervals are from sensitivity analysis, as no standard deviations or attempt counts were recorded. The assumed standard deviation of the measure is SD. For the complete SUS grid, see Appendix C.

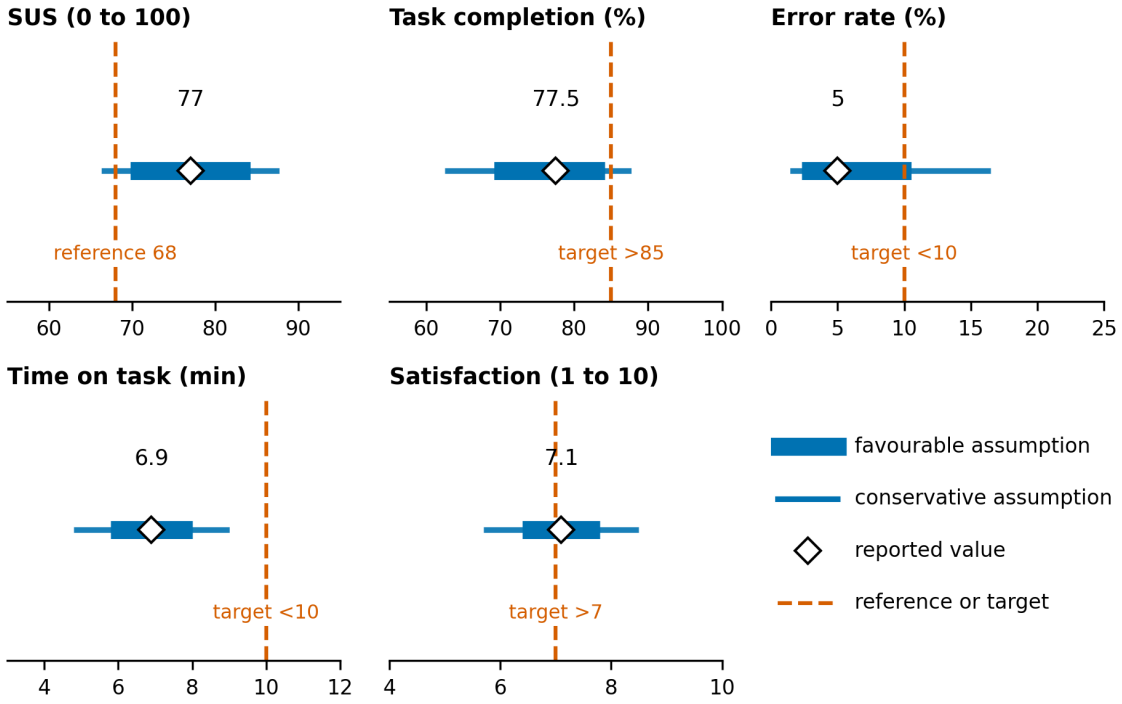

Fig. 12. Reported values to within a sensitivity range. The favourable assumption is used for thick bars, the conservative one for thin bars and the dashed line for the reference or target. The reported value is the diamond.

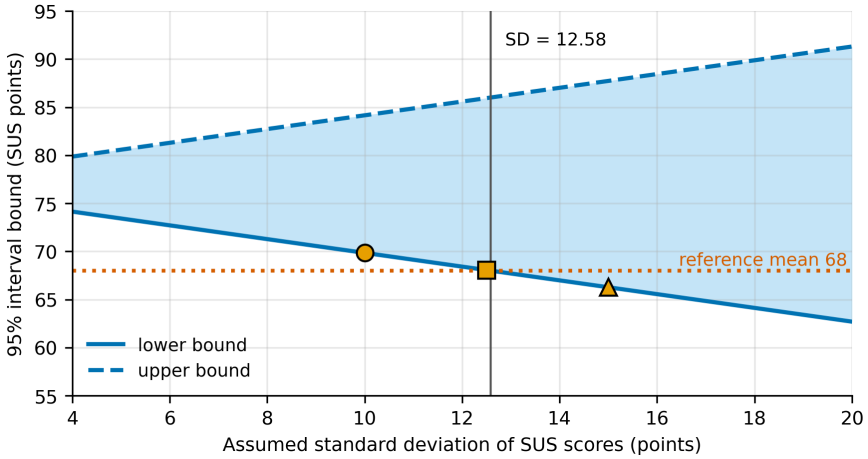

Fig. 13. The 95% interval for the mean SUS score (n = 10, mean = 77.0) for different assumed standard deviations. The value (12.58) at which the lower bound is equal to 68 is shown by the vertical line. The standard deviations are indicated by markers of 10, 12.5 and 15.

D. Error analysis

The type of errors supported by the material is listed in Table VIII. The first is called arithmetic. 22.5% of the attempts failed, which means the completion rate was 77.5%. An error rate of 5% will require at least 0.225 / 0.05 = 4.5 error opportunities per attempt (Fig. 14). The reported rates are consistent only if tasks had multiple opportunities for error or if some errors, e.g., abandonment, were not counted.

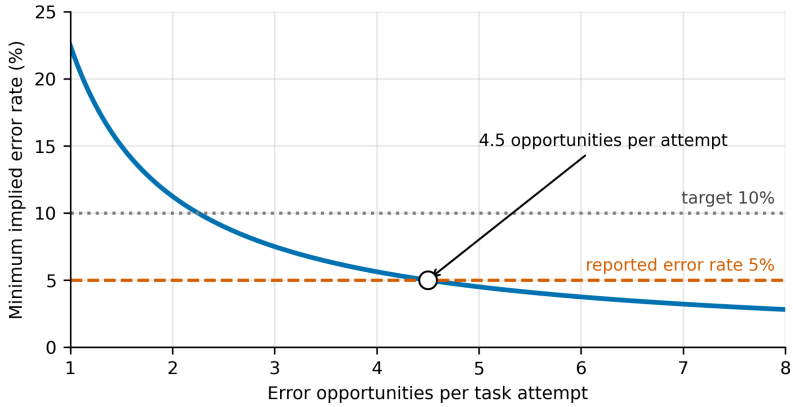

Fig. 14. The minimum error rate derived from a 22.5% rate of task failure, based on the number of error opportunities per attempt. The point at the open circle indicates the implied rate that is equal to the reported rate of 5%.

Two "observations" come from the prototype figures. For the selected summary, the work experience is over 5 years, as indicated in Fig. 9b, whereas the work experience in Fig. 8d and 9d starts in 2024. The pasted job description in Fig. 10 is for Shopify, and the visible question is about React and Tailwind design systems. The two screens can be from two different sessions, thus any inference about relevance is non-existent. Both observations illustrate the need for prompts and responses to be documented for each session.

TABLE VIII. THE ANALYSIS OF ERRORS OF THE PILOT EVALUATION.

Observation	**Evidence**	**Consequence**
Completion and error rates reconcile only under conditions	22.5% failure and 5% error rate imply at least 4.5 error opportunities per attempt (Fig. 14)	Report opportunities and failure causes per task
Completion verdict depends on attempts	Exact p = 0.184, 0.082 and 0.029 for 40, 80 and 120 attempts (Table VII)	Record the attempt count
Unsupported experience claim in a selected summary	More than five years stated; work history from 2024 (Figures 8d, 9b, 9d)	One instance only; LLM output must be audited against profile data
Question relevance not assessable	Job description and question may come from different sessions (Fig. 10)	Log prompt and output pairs
Failure causes not classifiable	No per task log retained	Adopt the schema in Appendix D
Target value inconsistent in the source	85% in the protocol; 93% in a summary table (Appendix E)	Protocol value used throughout

Answer to RQ3. The mean SUS was 77.0, and the dispersion of the scores is moderate (lower interval bound above 68 only below a standard deviation of 12.58), so the pilot can be considered acceptable for usability. The targets for task completion were not clearly met, with 77.5% and an error rate whose upper bound could not have been lower than 10.5%, and the verdict of task completion was based on an unreported number of attempts. In LLM suggested text, the error analysis identified one unsupported claim, which was an illustration, rather than an estimate of frequency.

VI. Discussion

Most of the questions and goals of the platform are achieved and the usability of the platform is promising in the pilot. The 77.0 falls within the "good" to "excellent" range of the SUS, closer to the "good" end. That position and the interval analysis, lend to a cautious reading. The claim of superiority over the reference requires a larger sample size consistent with the guidance that small SUS studies give wide intervals [38]. The sample size (12 to 24 subjects) is small but feasible.

The completion result should be given more weight than the SUS result. The error rate was such that a fifth of the task attempts failed, and it could only be confirmed with the fifth of the task attempts that succeeded, not the ones that failed. Perceived Usability and Task Success were not in line, as revealed by usability questionnaires. If the cause of each failure were recorded, it would be possible to determine if it was a result of authentication, template selection, form entry, or the LLM modules.

The more significant difference is in the LLM parts. Previous research and inspections reveal that generated text can present false information or bias [7], [9], [10], [25]. The one instance recorded here, of a risk that contradicts the work history, is consistent with that risk but is only one of many examples. A résumé tool that allows users to add such text without verification will be a burden on users. There are two design solutions based on the literature [44]: highlight the profile fields used for each generated sentence, and validate the generated text against the profile prior to display. The current platform does not support either.

The interview module is similar to the question generators (EZInterviewer) and the mock interview systems (Conversate, PolyInterview) in that it involves job context [5], [6], [27]. It's different because it does not provide feedback on answers, which is central to those systems. A free-text job description is also an injection surface [30], and the module stores answers with personal information, thus involving the obligations mentioned in Section II-D [32], [35]. These points prioritize the next version over any usability findings.

There are three lessons for practitioners. Requirements should be mapped to features so they can be seen when features are not implemented, e.g., profile import. Prompts and model identifiers should be versioned as artefacts [11]. Attempts, errors, and opportunities per task are to be recorded in usability logs, not summary rates.

VII. Threats to validity

Threats are grouped according to the taxonomy of Wohlin et al. [45] in Table IX. The main drawback is that only the usability assessment is performed. It did not assess the quality of the output, parsing by applicant tracking systems, or hiring results.

Table IX. Validity threats and how they are addressed.

Type	Threat	Handling or residual risk
Internal	Tasks designed by the developers; recruitment and moderation unreported; novelty effect; no comparison condition; LLM version unknown	Not mitigated; protocol and logging schema proposed in Appendix D
External	Ten participants from one institution; technology oriented scope; questionnaire extract of ten rows with mixed industries	Claims limited to a pilot with fresh graduates; no generalisation made
Construct	SUS measures perceived usability, not résumé quality or interview readiness; error and attempt definitions unrecorded; ATS suitability untested	Metrics defined in Table II; construct limits stated
Conclusion	Standard deviations and attempt counts unknown; targets set without independent validation; small sample	Sensitivity analysis over assumed values; no confirmatory claims
Data	Survey sample size unknown; truncated cells; screenshots of data tables; possible 10 versus 11 row discrepancy in the SUS screenshot	Counts reported as lower bounds; raw export requested from authors

VIII. Conclusion

This paper presents the design, implementation, and pilot testing of an LLM-based résumé builder and interview preparation website for fresh graduates. Seven of 9 traced requirements were implemented, while the profile import was not. The pilot usability was acceptable, with a mean SUS score of 77.0, which did not fall below the reference under moderate dispersion. The completion of the task was below target, based on the unreported number of attempts. The study does not demonstrate accuracy, impartiality, or utility of the LLM output to employers.

There are three steps to be followed in future work. First, the model and prompts will be fixed, and the generated summaries and questions will be compared against profile data and expert ratings on a documented corpus. Second, conduct the usability study again using 18-24 users and per-task logs. Third, provide feedback on the answers given during an interview, then test the exported PDF using open-source text-extraction software before claiming compatibility with applicant-tracking systems.

DECLARATIONS

Data availability. The ten questionnaire rows and the total of the pilot results are repeated in Appendix A and Appendix C. The full questionnaire export and per-participant SUS responses will be made available in a public repository with a persistent identifier prior to submission [repository DOI to be added] and retained by the authors. The figure and statistics scripts used in this paper are available upon request from the authors and will be sent out with the deposit.

Ethics. There is no record of ethics committee approval or exemption for the questionnaire and the pilot. The platform was used without requiring acceptance of terms at registration, and there was no record of a consent procedure for the study.

Use of AI assistance. This manuscript was prepared with the help of a large language model, which helped in drafting, editing, and figure scripts. The authors take full responsibility for the contents of this paper.

REFERENCES

[1] A. Adewolu Ogwo, "Higher education, skills development and students' preparedness for employability: A case study of the University of Lagos, Nigeria," thesis, Univ. College London, London, U.K., 2024.

[2] J. Fuller, M. Raman, E. Sage-Gavin, and K. Hines, "Hidden workers: Untapped talent," Harvard Bus. School Project Managing Future Work and Accenture, Boston, MA, USA, Tech. Rep., Sep. 2021.

[3] E. Wiles, Z. T. Munyikwa, and J. J. Horton, "Algorithmic writing assistance on jobseekers' resumes increases hires," *Manage. Sci.*, vol. 71, no. 12, pp. 10144–10164, 2025, doi: 10.1287/mnsc.2024.04528.

[4] S. B. Zinjad, A. Bhattacharjee, A. Bhilegaonkar, and H. Liu, "ResumeFlow: An LLM-facilitated pipeline for personalized resume generation and refinement," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (Demonstration Track)*, 2024, arXiv:2402.06221.

[5] T. Daryanto, X. Ding, L. T. Wilhelm, S. Stil, K. McInnis Knutsen, and E. H. Rho, "Conversate: Supporting reflective learning in interview practice through interactive simulation and dialogic feedback," *Proc. ACM Hum.-Comput. Interact.*, vol. 9, no. GROUP, Art. no. 9, 2025, doi: 10.1145/3701188.

[6] Z. Wen *et al.*, "PolyInterview: An LLM-based platform for immersive mock interview practice with comprehensive multimodal assessment," arXiv:2607.10310, 2026.

[7] L. Huang *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, 2025, doi: 10.1145/3703155.

[8] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, Art. no. 248, 2023, doi: 10.1145/3571730.

[9] L. Armstrong, A. Liu, S. MacNeil, and D. Metaxa, "The silicon ceiling: Auditing GPT's race and gender biases in hiring," arXiv:2405.04412, 2024.

[10] K. Wilson and A. Caliskan, "Gender, race, and intersectional bias in resume screening via language model retrieval," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, vol. 7, no. 1, 2024, doi: 10.1609/aies.v7i1.31748.

[11] M. Tafreshipour *et al.*, "Prompting in the wild: An empirical study of prompt evolution in software repositories," in *Proc. Int. Conf. Mining Softw. Repositories (MSR)*, 2025, arXiv:2412.17298.

[12] C. Agbonkhese, O. Gambo, T. Akande, I. Gambo, and A. Adewuyi, "Design and implementation of an AI-driven study assistant mobile application for personalized learning and academic management," in *ICT for Intelligent Systems (ICTIS 2025)* (Smart Innovation, Systems and Technologies, vol. 126), J. Choudrie, E. Tuba, T. Perumal, and A. Joshi, Eds. Springer, 2025, doi: 10.1007/978-981-95-1361-1_48.

[13] I. P. Gambo, S. Aina, C. Agbonkhese, O. O. Gambo, and B. J. Ojo, "Software analysis and design of a vocational career system: Development perspective," *Ife J. Inf. Commun. Technol.*, vol. 3, no. 1, pp. 40–50, 2018.

[14] Y. Xie, B. Zhu, and H. Huang, "Research on the design and implementation of a resumé building system based on natural language processing," in *Proc. 3rd Int. Conf. Inf. Manage., Big Data Artif. Intell.*, IEEE, 2019, pp. 51–55, doi: 10.1109/IMBDABI.2019.00019.

[15] R. Tyagi, N. Singh, A. Baghel, and A. Singh, "Resumé builder application," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 8, 2020.

[16] C. V. Mhatre, V. R. Mishra, M. M. J. Momin, and V. Nemade, "Real resumé: A web-app for building resumé and portfolio," *Int. Res. J. Eng. Technol.*, vol. 7, no. 4, 2020.

[17] B. Kungwani, A. Manglani, N. Dembal, H. Hirani, and L. Sawlani, "Analytical resumé builder: A web application for creating a resumé which gives the best impact in this competitive world," *Ann. Rom. Soc. Cell Biol.*, vol. 24, no. 2, pp. 235–238, 2020.

[18] H. S. Kanyal, M. Goel, R. C. Tripathi, A. Kumar, B. K. Saraswat, and S. Singh, "Resumé designing using speech recognition system," in *Proc. 10th Int. Conf. Syst. Modeling Adv. Res. Trends (SMART)*, IEEE, 2021, pp. 245–248, doi: 10.1109/SMART52563.2021.9676297.

[19] M. C. J. Yi and O. H. Fang, "A resumé generator with augmented reality features," in *Proc. 22nd Int. Conf. Inf. Integr. Web-Based Appl. Services (iiWAS)*, ACM, 2020, pp. 357–361.

[20] A. Malhotra, A. Sharma, D. Sharma, and F. Ahmed, *Résumé Builder Over Rejuvenated AI Features in Website Development*, 1st ed. Boca Raton, FL, USA: CRC Press, doi: 10.1201/9781003167488.

[21] J. Kaur and K. Singh, "Development of an intelligent resumé builder system," in *Proc. 3rd Int. Conf. Comput. Methodology*, 2021.

[22] K. Tejaswini, V. Umadevi, S. M. Kadiwal, and S. Revanna, "Design and development of machine learning based resumé ranking system," *Global Transitions Proc.*, vol. 3, no. 2, pp. 371–375, 2022, doi: 10.1016/j.gltp.2022.10.067.

[23] H. Mgarbi, M. Y. Chkouri, and A. Tahiri, "Towards a new job offers recommendation system based on the candidate resumé," *Int. J. Comput. Digit. Syst.*, vol. 14, no. 1, 2023.

[24] F. Zhu *et al.*, "Layout-aware parsing meets efficient LLMs: A unified, scalable framework for resume information extraction and evaluation," arXiv:2510.09722, 2025.

[25] S. Indukuri and A. Agrawal, "Grounded optimization: A layered engineering framework for reducing LLM hallucination in automated personal document rewriting," arXiv:2607.01457, 2026.

[26] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "MACH: My automated conversation coach," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. (UbiComp)*, 2013, pp. 697–706, doi: 10.1145/2493432.2493502.

[27] M. Li *et al.*, "EZInterviewer: To improve job interview performance with mock interview generator," arXiv:2301.00972, 2023.

[28] C. Qin *et al.*, "Automatic skill-oriented question generation and recommendation for intelligent job interviews," *ACM Trans. Inf. Syst.*, vol. 42, no. 1, Art. no. 27, 2023, doi: 10.1145/3604552.

[29] T. Maurer, J. Solamon, and D. Troxtel, "Relationship of coaching with performance in situational employment interviews," *J. Appl. Psychol.*, vol. 83, no. 1, pp. 128–136, 1998, doi: 10.1037/0021-9010.83.1.128.

[30] G. Sun *et al.*, "CoMAI: A collaborative multi-agent framework for robust and equitable interview evaluation," arXiv:2603.16215, 2026.

[31] R. Massenon, I. Gambo, J. A. Khan, C. Agbonkhese, and A. Alwadain, "'My AI is lying to me': User-reported LLM hallucinations in AI mobile apps reviews," *Sci. Rep.*, vol. 15, no. 1, Art. no. 30397, 2025.

[32] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *High-Confidence Comput.*, Art. no. 100211, 2024, doi: 10.1016/j.hcc.2024.100211.

[33] R. Massenon, I. Gambo, and J. A. Khan, "Unveiling hidden permissions: An LLM framework for detecting privacy and security concerns in AI mobile apps reviews," *Autom. Softw. Eng.*, vol. 33, Art. no. 42, 2026, doi: 10.1007/s10515-025-00567-9.

[34] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Off. J. Eur. Union*, 2024.

[35] Federal Republic of Nigeria, "Nigeria Data Protection Act, 2023," 2023.

[36] J. Brooke, "SUS: A 'quick and dirty' usability scale," in *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland, Eds. London, U.K.: Taylor & Francis, 1996, pp. 189–194.

[37] J. R. Lewis, "The System Usability Scale: Past, present, and future," *Int. J. Hum.-Comput. Interact.*, vol. 34, no. 7, pp. 577–590, 2018, doi: 10.1080/10447318.2018.1455307.

[38] J. Sauro and J. R. Lewis, *Quantifying the User Experience: Practical Statistics for User Research*, 2nd ed. Cambridge, MA, USA: Morgan Kaufmann, 2016.

[39] A. Bangor, P. Kortum, and J. Miller, "Determining what individual SUS scores mean: Adding an adjective rating scale," *J. Usability Stud.*, vol. 4, no. 3, pp. 114–123, 2009.

[40] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Q.*, vol. 28, no. 1, pp. 75–105, 2004.

[41] A. A. Adeagbo, "Development of an AI-powered resumé builder and interview preparation platform," B.Sc. project report, Dept. Comput. Sci. Eng., Obafemi Awolowo Univ., Ile-Ife, Nigeria, Sep. 2025.

[42] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *J. Amer. Statist. Assoc.*, vol. 22, no. 158, pp. 209–212, 1927.

[43] I. P. Gambo, H. O. Odukoya, A. A. Oke, and E. R. Adagunodo, "Analysis and classification of requirements specification for web application development: A case study approach," *J. Comput. Sci. Appl.*, vol. 27, no. 1, pp. 144–160, 2020.

[44] I. Gambo, R. Massenon, C.-C. Lin, R. O. Ogundokun, S. Agarwal, and W. Pak, "Enhancing user trust and interpretability in AI-driven feature request detection for mobile app reviews: An explainable approach," *IEEE Access*, vol. 12, pp. 114023–114045, 2024.

[45] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*. Berlin, Germany: Springer, 2012.

APPENDIX A. QUESTIONNAIRE EXTRACT

Table A1 records the ten rows reproduced in the project report. Text cut off in the source is marked with an ellipsis. The résumé type column reads "Industry Resume (For corporate, business, or tec…" or "Academic Resume (For research, teaching, or gra…".

TABLE A1. TRANSCRIPTION OF THE QUESTIONNAIRE EXTRACT.

Row	Role	Industry	Résumé type	Sections named (visible text)	Format
1	Hiring manager	IT	Industry	Work Experience	Functional
2	Recruiter	Healthcare	Academic	Personal Details, Teaching Experience (For Acade…	Chronological
3	Job seeker	Education	Academic	Certifications, Education	Chronological
4	HR professional	Finance	Industry	Personal Details, Work Experience, Skills, Certifica…	Hybrid
5	Job seeker	IT	Industry	Work Experience, Skills, Certifications, Education, …	Hybrid
6	Job seeker	IT	Industry	Work Experience, Skills, Technical Projects (For In…	No answer
7	Recruiter	Hospitality management	Industry	Personal Details, Work Experience, Skills, Certifica…	Chronological
8	Designer and student	Design	Industry	Work Experience, Skills, Certifications, Education	Hybrid
9	Job seeker	IT, healthcare	Industry	Work Experience, Skills	Hybrid
10	Job seeker	Finance	Industry	Personal Details, Work Experience, Skills, Certifica…	Hybrid

APPENDIX B. SYSTEM USABILITY SCALE ITEMS

Table A2 lists the standard SUS items [36] scored by Equation 1. Odd items are positively worded and even items negatively worded.

TABLE A2. SUS ITEMS.

Item	Statement
1	I think that I would like to use this system frequently.
2	I found the system unnecessarily complex.
3	I thought the system was easy to use.
4	I think that I would need the support of a technical person to be able to use this system.
5	I found the various functions in this system were well integrated.
6	I thought there was too much inconsistency in this system.
7	I would imagine that most people would learn to use this system very quickly.

8	I found the system very cumbersome to use.
9	I felt very confident using the system.
10	I needed to learn a lot of things before I could get going with this system.

APPENDIX C. SENSITIVITY GRID FOR THE SUS RESULT

Table A3 gives the 95% interval, the test against 68, the effect size, the power and the sample size needed for 80% power across assumed standard deviations (n = 10, mean 77.0).

TABLE A3. SUS SENSITIVITY GRID.

SD	SE	95% interval	t (df 9), p	d	Power	n for 80%
8.0	2.53	71.3 to 82.7	3.56, 0.006	1.12	0.88	9
10.0	3.16	69.8 to 84.2	2.85, 0.019	0.9	0.72	12
12.5	3.95	68.1 to 85.9	2.28, 0.049	0.72	0.53	18
15.0	4.74	66.3 to 87.7	1.9, 0.09	0.6	0.4	24
17.5	5.53	64.5 to 89.5	1.63, 0.138	0.51	0.31	32

The lower bound equals 68 at a standard deviation of 12.58 points.

APPENDIX D. LOGGING PROTOCOL AND ITEMS TO BE SUPPLIED

Table A4 defines the records that a replication should keep, and lists items that the authors must supply.

TABLE A4. LOGGING SCHEMA AND OUTSTANDING ITEMS.

Record or item	Content
Task attempt log	Participant identifier, task identifier, start and end time, outcome (success, failure, abandonment), error count, error opportunities, error type, free text note
LLM call log	Timestamp, model identifier and version, parameters, prompt template identifier, input fields, output text, user action (accepted, edited, rejected)
SUS responses	Ten item scores per participant; computed SUS per participant
Outstanding: ethics	Approval or exemption reference; consent procedure
Outstanding: data	Full questionnaire export with respondent count; per participant SUS items 1 to 10; per task outcomes and times
Outstanding: artefact	Model identifier, prompt templates, library versions, repository and persistent identifier
Outstanding: administrative	Author details, funding and competing interests

APPENDIX E. SOURCE REPORT DISCREPANCIES AND THEIR TREATMENT

Table A5 lists inconsistencies in the project report and how this paper treats them.

TABLE A5. DISCREPANCIES IN THE PROJECT REPORT.

Discrepancy	Treatment in this paper
The task completion target appears as above 85% in the protocol, below 93% in the summary table and 85 in the chart	The protocol value of 85% is used
SUS 77.0 is labelled excellent in the summary table	Interpreted against adjective anchors; closer to good [39]
Equation for completion rate divides by tasks completed	Denominator corrected to attempted tasks (Equation 2)
The SUS data screenshot appears to show eleven rows, and the text states ten participants	n = 10 used; raw export requested
One journal article is listed twice in the reference list under two author orders and years	Cited once
Social profile and Google data integration appear in the aim and contributions	Google sign in documented; data import treated as not implemented