technologies

MDPI

Article

Comparative Analysis of Local Large Language Models for Ranking Higher Education Programmes Based on Applicant Digital Profiles

Artem Sveshnikov, Yury Nikitnikov, Maxim Shiltsyn, Denis Dedov and Artem Obukhov *

Laboratory of VR Simulators, Tambov State Technical University, Tambov 392000, Russia; artlinden99@gmail.com (A.S.); yury.nikitnikov@yandex.ru (Y.N.); maks.shilcin@mail.ru (M.S.); hammer68@mail.ru (D.D.)

* Correspondence: obuhov.art@gmail.com

Abstract

Choosing among closely related higher-education programmes requires interpretation of heterogeneous applicant data while preserving data confidentiality. This study compares ten locally executable large language model (LLM) configurations as semantic rankers, contextualises their performance against a hybrid term frequency–inverse document frequency (TF–IDF) cosine baseline, and evaluates robustness when relevant competencies are expressed indirectly. The main benchmark comprised 50 anonymised applicant profiles and 10 degree programmes; an additional processed-profile robustness set comprising 10 profiles was used to reduce direct lexical overlap with the programme catalogue. Rankings were evaluated using Accuracy@1, Accuracy@3, normalised discounted cumulative gain at 5 (NDCG@5) and mean reciprocal rank at 5 (MRR@5), while structured-output validity and inference time were assessed for the LLMs. Among the local LLMs, gpt-oss-20b-MXFP4 achieved the highest Accuracy@1 (0.76), whereas gemma-4-E4B-it-Q8_0 achieved the highest Accuracy@3 (0.92), and Ministral-3-14B-Reasoning-2512-Q4_K_M provided a favourable quality–latency balance. On the main benchmark, TF–IDF achieved Accuracy@1 = 0.88 and NDCG@5 = 0.9557, exceeding all LLM configurations and demonstrating a strong lexical signal. On the processed-profile robustness set, the Accuracy@1 of the TF–IDF baseline decreased to 0.60, whereas gemma achieved 0.90. The results provide initial evidence that several local LLM configurations can reproduce observed programme-selection patterns in a limited pilot benchmark. However, the findings should be interpreted as task-specific model-comparison results rather than as full validation of an autonomous career guidance system.

Keywords: large language models; educational trajectories; applicant digital profile; decision support system; educational recommender systems

Check for updates

Academic Editor: Hongying Meng

Received: 15 June 2026
Revised: 15 July 2026
Accepted: 16 July 2026
Published: 17 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Selecting a field of study when applying to a university is a decision-making task under conditions of incomplete and unstructured information [1]. Applicants need to relate their own interests, entrance examination results, achievements, and perceptions of future professional activity to the content of degree programmes, the disciplines studied, and the competencies to be developed. This task is also important for an educational organisation, since an erroneous or poorly justified programme choice may lead to reduced student engagement, a change in educational trajectory after admission, or withdrawal from study.

Therefore, preliminary career guidance support tools should not replace the decision of the applicant or the expert, but should instead provide a justified ranked set of relevant fields of study.

An applicant's digital profile may include various academic indicators, information on Olympiads and competitions, descriptions of completed projects, technological interests, and non-formalised textual characteristics of achievements. A degree programme, in turn, is described not only by its code and title, but also by a list of disciplines, professional tasks, technologies used, and expected learning outcomes, including competencies. Consequently, selecting an appropriate field of study requires the comparison of heterogeneous features: numerical scores, categorical characteristics, and textual descriptions, between which not only direct but also semantic relationships may exist. For example, an applicant's participation in the development of a computer vision software system may be relevant not to a single field of study, but to several programmes that differ in their emphasis on software engineering, artificial intelligence, robotics, or information systems. Manually performing such large-scale data processing during an admissions campaign is highly labour-intensive and is difficult to implement under limited human, time, and computational resources, which highlights the need for automation in this subject area [2].

Decision support systems in education have traditionally used expert rules, competency models, multi-criteria procedures, and machine learning methods. These approaches make it possible to formalise individual parameters of correspondence between a learner and a degree programme; however, their application becomes more difficult when processing extended textual descriptions of achievements and when it is necessary to account for contextual semantic relationships. In recent years, large language models (LLMs) have formed a new area of intelligent educational data processing owing to their ability to analyse natural language, extract features from semi-structured descriptions, and generate explainable textual recommendations. Studies in education have demonstrated the applicability of LLMs to automated assessment of written work, analysis of educational experience, interpretation of application materials, and the development of advisory chatbots for applicants [3,4].

Despite the growing number of studies related to LLMs, their application has mainly focused on the assessment of textual data, the analysis of learner characteristics, and the development of chatbots and reference systems [4,5], including various types of retrieval-augmented generation (RAG) systems [6]. Such systems improve access to information, but they do not evaluate the quality of ranking degree programmes on the basis of a combined set of academic indicators, interests, and portfolio data.

The analysis of existing studies confirms the feasibility of using LLMs to process unstructured data; however, a research gap remains in the task of ranking degree programmes according to a multi-component applicant digital profile [4,5,7]. This task differs from answering a reference question or assessing a single text: the model must compare several alternatives, determine the degree of their correspondence to the user profile, and correctly order the programmes, with an error in the first position having greater practical importance than an error at the end of the recommendation list. Therefore, studying this task requires the use of ranking metrics, including first-choice accuracy, the inclusion of the target field of study among the top-ranked positions, and position-sensitive indicators of recommendation quality.

A separate factor underlying the relevance of this study is the use of cloud-based models, including models provided by Google, OpenAI, and other developers, which at the preliminary stage demonstrated the feasibility of the task. However, applicants' personal data constitute sensitive information, and sending such data to external LLM services may conflict with regulatory and legal requirements for the processing of this type of data.

Therefore, an important task is to implement the decision support system (DSS) within the infrastructure of the educational organisation, ensuring full control over the processes of data collection, transfer, and storage [8].

At the same time, the specific characteristics of local LLMs must be taken into account: they differ in architecture, number of parameters, degree of quantisation, support for reasoning modes, inference speed, and stability of output in the required format. A larger model does not necessarily provide more accurate ranking, while a model with acceptable semantic capability may prove unsuitable for integration into a DSS if it does not reliably return structured output or requires an unacceptable computation time. In this context, the selection of an LLM for an educational DSS is considered a multi-criteria task of experimental model comparison. For practical use, it is not sufficient to establish which model most frequently places the reference degree programme in the first position. It is also necessary to assess whether the model preserves the target field of actual admission among the top alternatives; how correctly it orders the recommendations; whether it can reliably generate a machine-processable response; and whether its inference time meets the requirements of an interactive advisory system.

The research gap addressed in this study emerges at the intersection of several fields. Educational recommender systems provide methods for ranking learning objects, but they are usually applied to courses, resources, or learning paths after enrolment. Career guidance requires the interpretation of heterogeneous applicant data before enrolment, when no interaction history is available. Large language models can process such semi-structured textual information, but their use in educational decision support is often based on cloud services, which raises privacy and reproducibility concerns. Therefore, it remains unclear whether locally deployed LLMs can serve as reliable semantic rankers of degree programmes under a fixed institutional catalogue and a controlled evaluation protocol.

The study is interdisciplinary because it connects theories of educational and career choice with educational recommender systems, information retrieval, natural-language processing, and applied decision support. The scientific hypothesis is that sparse lexical retrieval and local generative rankers exploit different information signals: lexical methods should be strongest when applicant and catalogue terminology overlap, whereas selected LLMs may retain more ranking quality when evidence of competence is expressed indirectly. This hypothesis is tested under a common programme catalogue, a unified prompt protocol, and identical ranking metrics.

The aim of this study is to compare locally executable LLMs as semantic rankers of degree programmes, contextualise their performance against a traditional TF–IDF content-based baseline, and conduct an exploratory evaluation of ranking robustness.

To achieve this aim, the study addresses the following research questions (RQs):

1. How substantially does the quality of ranking degree programmes differ between local LLMs under a unified task formulation and a common set of digital profiles?
2. Which models are the most reliable in a scenario requiring the identification of a single priority field of study, and which models are more appropriate for forming an extended list of alternatives?
3. What is the relationship between ranking quality, stability of structured output, and computational latency in local models?
4. Is it possible to select a local LLM that, in terms of its combined characteristics, is suitable for use as an intelligent component of a career guidance DSS?
5. How does the performance of local LLMs compare with a traditional lexical retrieval baseline, and how does this relationship change when relevant applicant competencies are expressed indirectly?

The main contribution of this study is as follows. First, the task of matching an applicant digital profile to a register of fields of study is formulated as a problem of ranking alternatives, which makes it possible to apply metrics from information retrieval and recommender systems. Second, the necessary pipeline is developed, including a mathematical and algorithmic description of the required steps for solving the task, and an experimental comparison of ten locally executable LLMs is performed under a unified setting that controls for differences in input data and task formulation. Third, alongside Accuracy@1 and Accuracy@3, position-sensitive metrics, namely NDCG and MRR, are used to distinguish between models. Fourth, structured-output validity and inference time are evaluated, since in an applied DSS, format errors and excessive latency are independent factors that may render a model unsuitable for use. Fifth, a traditional TF–IDF cosine baseline and an additional processed-profile robustness set are introduced to distinguish lexical matching capability from contextual semantic ranking. Finally, the results provide an empirical basis for selecting a local LLM that can be considered a candidate for subsequent implementation in a university career guidance service, while preserving mandatory expert control over the final recommendation.

The experimental pipeline was implemented in Python 3.11, with local model inference performed using llama.cpp. On the main benchmark, the TF–IDF baseline achieved the highest ranking metrics, whereas selected local LLMs preserved more ranking quality on the processed-profile robustness set with reduced direct lexical overlap.

The remainder of the article is organised as follows. Section 2 presents an analysis of studies on educational recommender systems, the use of LLMs in education, and the local deployment of generative models. Section 3 formalises the ranking task and describes the data, models, TF–IDF baseline, robustness protocol, and evaluation procedure. Section 4 presents the results of the comparative experiment and the statistical testing of differences between models. Section 5 discusses the practical implications of the results obtained, the limitations of the study, and the requirements for using local LLMs in an educational DSS. The conclusions summarise the main findings and outline directions for further research. RQ1–RQ4 are addressed through the comparative and statistical analyses in Sections 3.3–3.6 and Sections 4.1–4.4 and interpreted in Sections 5.1–5.3, whereas RQ5 is addressed through the TF–IDF baseline and robustness evaluation in Sections 3.4, 3.6, 4.5 and 5.4.

2. Related Works

This section positions the proposed study at the intersection of educational recommender systems, career guidance, and local large language model deployment. The review first considers the educational task of selecting a field of study, then discusses recommender-system approaches in education, the use of LLMs for processing educational and admission-related data, and finally the specific research gap addressed by the present benchmark.

2.1. Analysis of the Task of Selecting a Field of Study

Personalisation in education covers a broad range of tasks, including the formation of individual curricula, the selection of disciplines and learning resources, the adaptation of educational content, the prediction of academic risks, and support for professional self-determination. The present study considers one of the earliest stages, which has a substantial influence on the entire educational process: the selection of a field of study by an applicant when entering a university. At this stage, the educational trajectory is represented as a set of available bachelor's and specialist degree programmes, one or several of which should be recommended to the user on the basis of their digital profile.

This task differs from traditional admissions counselling. A user may formulate a question about a preferred profession or discipline of interest; however, their actual

correspondence to a degree programme is determined by a more complex set of factors: academic results, profile-specific subjects, achievements, project experience, and interests in technologies and professional domains. At the same time, degree programmes have a predefined formalised description, including lists of disciplines, competencies to be developed, technologies, professional tasks, and areas of future professional activity. Consequently, direct lexical matching may provide a strong baseline when applicant and programme descriptions use explicit and consistent terminology, but it may be insufficient when relevant competencies are expressed indirectly or through heterogeneous contextual evidence. The task is further complicated by the fact that the format in which degree programmes are presented is often difficult for applicants to interpret.

In decision support systems, this scenario can be represented as a task of ranking alternatives. Let there be an applicant profile containing numerical features of academic preparation and a textual description of interests and achievements, as well as a set of degree programmes, each characterised by the content of its curriculum and its competency profile. The task is then to order the degree programmes by their degree of correspondence to the applicant, while preserving the human right to make the final choice.

Studies on artificial intelligence in higher education show that profiling, prediction, personalisation, and intelligent support are stable areas of development in educational technologies. In the systematic review by Zawacki-Richter et al., four major groups of AI applications in higher education are identified: profiling and prediction, assessment, adaptive systems and personalisation, and intelligent tutoring systems [3]. For the task of selecting a field of study, the first two groups are of particular importance: it is necessary to construct a user model on the basis of heterogeneous data and compare it with a set of degree-programme alternatives. At the same time, the authors of the review note that many AI-based developments have an insufficient connection with pedagogical foundations, and that issues of privacy, transparency, and the role of humans in decision-making are underrepresented.

A classical approach to the selection and management of a professional educational trajectory is the use of a competency model and mathematical modelling of adaptive educational systems. In such work, the state and goals of the learner are related to the requirements of the educational system [9]. This line of research forms an important theoretical basis for the present study: an educational recommendation should not be generated as an arbitrary textual response, but as the result of matching an individual profile with the competency content of a degree programme. The new element is the use of an LLM as a semantic ranker capable of processing extended descriptions of achievements and semantic relationships between an applicant's interests and the disciplines of a degree programme.

2.2. Educational Recommender Systems

Educational recommender systems (ERSs) have developed as a specific class of recommender systems aimed at supporting learning. An educational recommendation is associated not only with the user's current preferences, but also with competency development, the achievement of learning outcomes, and compliance with the sequence in which knowledge is acquired. The literature considers recommendations of learning materials, online courses, assignments, educational pathways, interaction groups, and additional resources [10,11].

From a methodological perspective, ERSs use three main groups of approaches. Collaborative filtering relies on user similarity and the history of interactions with resources. Content-based methods compare the features of the user with the properties of the recommended object. Knowledge-based and hybrid methods use rules, competency ontologies, knowledge requirements, and combinations of several recommendation mechanisms. For

selecting a field of study at the admission stage, content-based and knowledge-based mechanisms are of the greatest importance, since a new applicant, as a rule, has no history of interaction with the university's educational environment.

The substantive complexity of the task under consideration further limits the applicability of simple similarity models. Two fields of study may use similar terms while differing in meaning and professional orientation. For example, the concepts of programming, data analysis, and modelling may occur in descriptions of programmes in information systems, applied informatics, and systems analysis; however, the relative significance of these disciplines for a particular applicant is determined by the context of their projects and achievements. Similarly, an interest in unmanned systems may be related to robotics, automation, radio engineering, or software engineering, depending on whether the applicant designed mechanical components, electronic circuits, a control system, or software.

Reviews of ERSs also identify the problem of evaluating the quality of educational recommendations. Many studies use metrics of accuracy, rating correspondence, or clicks, whereas the influence of recommendations on educational outcomes and trajectory quality is assessed much less frequently [12]. For the selection of a field of study, this problem is particularly important: a recommendation is a decision with a delayed effect. Therefore, at the initial stage, a realistic evaluation can be based on the correspondence between the ranking and an expert reference or an observed educational choice, such as the applicant's actual field of admission. In the longer term, however, such an evaluation should be supplemented by indicators of satisfaction with the choice, academic success, stability of interest in the programme, and expert assessment of explanations.

Recent studies further distinguish context-aware and knowledge-based educational recommender systems, which incorporate learner characteristics, competency structures, domain ontologies, and situational information into the recommendation process [13,14]. These approaches are particularly relevant at the admission stage, where interaction histories are unavailable and recommendations must instead be formed from explicit profile and programme content.

Thus, educational recommender systems provide the conceptual framework for the present study: a user, a set of educational objects, a relevance model, and a ranked result. What is needed is a tool capable of simultaneously taking into account academic features, portfolio evidence, professional interests, and the semantic content of disciplines. Large language models potentially satisfy these requirements.

2.3. Application of Large Language Models in Education

The emergence of LLMs has substantially expanded the range of tasks in which educational data can be analysed in natural-language form. Kasneci et al. show that LLMs can be used for personalised learning, the preparation of educational content, feedback, counselling, and decision support; however, they also require control of errors, biases, privacy risks, and mandatory human oversight [15]. For the present study, it is important that LLMs allow a transition from the formal matching of predefined tags to the interpretation of an extended textual context: a model can establish a relationship between the description of an applicant's project and the content of a degree programme even when different formulations are used in the two texts.

The most developed area of empirical evaluation of LLMs in education is automated text assessment. Pack, Barrett, and Escalante investigated the use of several LLMs for assessing written work by learners of English and showed that model quality should be analysed not only in terms of agreement with expert assessment, but also with respect to the validity and reliability of the result [16]. Li and Liu considered automated essay scoring in Japanese as a foreign language and found that LLMs can outperform a number

of traditional approaches when processing complex linguistic material [17]. These studies demonstrate an important capability of LLMs: the extraction of meaningful features from coherent text and their alignment with specified criteria.

However, the task of essay assessment is fundamentally different from the selection of an educational field. In automated assessment, the model usually compares one text with a rubric and returns a score or category. In the career guidance task, the model must compare a multi-component profile with several alternatives and form their order. In other words, it is necessary not only to recognise the presence of relevant characteristics, but also to determine which programme corresponds to them to the greatest extent and which alternatives should be placed lower in the list. This shifts the task from the domain of text assessment to the domain of semantic ranking.

A study closer to the topic of the present work is that by Lira et al., which focuses on the use of artificial intelligence to extract personal characteristics from applicants' materials in college admissions [18]. The authors demonstrated the possibility of automatically detecting features related to learning motivation, leadership, teamwork, and persistence. This work confirms that documents accompanying admission can be analysed by intelligent methods. At the same time, its subject is the assessment of candidate characteristics rather than the selection of a degree programme. A successful applicant and an educational trajectory suitable for that applicant are related but distinct objects of decision-making: the first task evaluates the candidate against admission criteria, whereas the second matches their interests and competencies with the content of several programmes.

Another important class of studies is associated with the extraction of educational experience from unstructured texts. Kondo and Nishigori examined the ability of a generative model to extract learning events from clinical practice logs; this type of task demonstrates the applicability of LLMs to structuring professionally significant experience described in free form [19]. For the applicant digital profile, this finding has important methodological significance: a portfolio may contain the names of competitions, descriptions of projects, technologies used, and participation outcomes, all of which need to be transformed into a representation comparable with the competencies of degree programmes.

A separate line of research concerns intelligent services for supporting applicants. Advisory chatbots for admissions campaigns are intended to answer questions about admission rules, deadlines, requirements, degree programmes, and organisational procedures. Nguyen et al. describe the use of LLMs for university admissions counselling, where the model provides a natural-language interface to information about the educational organisation [20]. In the work by Nguyen et al. on the URAG architecture, vector and lexical search are combined with a generative model to improve the accuracy of responses from a university chatbot [21].

Such systems are important for the digitalisation of admissions campaigns; however, their objective function differs from that considered in the present article. A chatbot answers queries such as "Which examinations are required for admission to this programme?" or "Which documents must be submitted?", that is, it retrieves and formulates relevant reference information. A system for selecting a field of study solves the inverse task: given a description of a particular applicant, it must determine which programmes are most relevant to that applicant's achievements and interests. This task requires a model of preference or correspondence, not merely reliable retrieval of information from a knowledge base.

In the broader field of recommender systems, studies are emerging in which LLMs are used as components of personalised recommendation. Models can enrich textual representations of objects, interpret user preferences, construct semantic chains of correspondence, and act as generative rankers. Ma et al. investigated the capabilities of LLMs for course recommendation in massive open online courses and showed that generative models can

be comparable with traditional recommender methods in terms of recommendation quality criteria [22]. This line of work is the closest to the present study, since it considers LLMs not only as a means of generating a response, but also as a mechanism for selecting educational objects.

Most early studies of generative AI applications in education used cloud-based models or closed services, since these were the first to demonstrate high-quality dialogue interaction and natural-language processing. This is an acceptable scenario for experimental hypothesis testing. However, when moving towards a career guidance DSS for an educational organisation, additional requirements arise that are not reducible to the quality of an individual response.

First, an applicant digital profile may include personal information, academic results, and portfolio data. The use of a remote application programming interface (API) in such a scenario requires legal regulation of data flows, whereas local deployment makes it possible to design the processing environment within the university infrastructure. Second, for scientific reproducibility, it is important to be able to fix a specific model, its parameters, and its inference configuration. Third, large-scale counselling of applicants requires an assessment of computational cost and recommendation generation time, since a model with high quality but unacceptable latency may be unsuitable for an interactive service.

The development of open-weight LLMs has created a technical possibility for investigating such systems locally. Touvron et al. presented open Llama 2 models of different sizes, demonstrating that language models suitable for dialogue scenarios can be distributed in a form that enables their independent deployment and study [23]. At the same time, quantisation methods can reduce memory and computational requirements. Frantar et al. proposed the GPTQ method, which enables large generative transformers to be quantised to 3–4 bits per weight with limited quality loss and a substantial reduction in inference requirements [24].

Nevertheless, the mere availability of local execution does not prove that a model is suitable for educational ranking. General LLM benchmarks measure linguistic competence, reasoning, factual knowledge, or instruction following, but they do not reflect the specific ability to match a Russian-language applicant portfolio with a catalogue of fields of study. Chang et al. emphasise that LLM evaluation should be linked to the applied task and should include not only general indicators, but also characteristics that are critical for the specific use scenario [25]. For a career guidance DSS, such characteristics include the quality of programme ordering, the preservation of a relevant field among the top alternatives, the ability to obtain a structured result, and acceptable profile processing time.

Consequently, the comparison of local models in the present study is not merely an engineering choice of computational configuration. It is an investigation of the applied effectiveness of different LLMs as an intelligent ranking component of an educational DSS. The central question is whether local models can provide sufficient semantic quality for recommending fields of study, and whether there is a model that demonstrates a rational balance between accuracy and operational applicability [26].

2.4. Synthesis of the Interdisciplinary Research Gap

The analysis conducted above makes it possible to identify four interrelated research gaps.

The first gap concerns the object of recommendation. Most educational recommender systems focus on learning resources, individual courses, or the sequence of learning for an already enrolled student. In the present study, the object of recommendation is the field of study at the admission stage, that is, a decision made before the user has accumulated a history of interaction with the university information system. This makes

the information contained in the digital profile primary: interests, academic results, and portfolio achievements.

The second gap concerns the type of data analysed. Studies on LLMs in education convincingly demonstrate the analysis of essays, feedback, and other individual texts; however, career guidance recommendation requires the simultaneous interpretation of several sources of information and their comparison with the content of several degree programmes. The model must extract competency indicators from the portfolio, recognise semantic proximity to disciplines, and form an order of alternatives. Therefore, the task is not classificatory but ranking-oriented.

The third gap relates to the type of models used. In the educational literature, a substantial proportion of studies on generative AI are based on cloud-based or closed LLMs. Meanwhile, the use scenario in a university DSS imposes requirements for profile confidentiality, model reproducibility, configuration control, and computational cost. Local models can potentially meet these requirements, but their comparative effectiveness in selecting a field of study must be established empirically rather than assumed on the basis of general benchmarks.

The fourth gap concerns evaluation criteria. For the use of LLMs in a DSS, it is not sufficient to show that a model can generate a plausible explanation or list related programmes. The order of recommendations is important: the first position has the greatest influence on the advisory decision, whereas the inclusion of the correct, or most suitable, programme in the top-k reflects the usefulness of the model for forming a set of alternatives. Therefore, the study must use ranking metrics, and model comparison must take into account not only the agreement of the ranking with the observed admission outcome, but also the suitability of the result for integration into an applied system.

The present study addresses these gaps as follows. The selection of a field of study is formulated as the task of ranking a fixed set of degree programmes according to a multi-component applicant digital profile. Locally executable large language models are compared as the intelligent component, which makes it possible to evaluate the applicability of LLMs in a career guidance DSS. The result produced by each model is analysed using Accuracy@1, Accuracy@3, MRR, and NDCG, which reflect different aspects of ranking quality. Thus, the study is aimed not at demonstrating the generative capabilities of LLMs as such, but at making a justified selection of a local model capable of supporting decision-making on the most appropriate field of study.

From the perspective of educational choice, the study is also related to Social Cognitive Career Theory [27], which links academic and career choices to self-efficacy, outcome expectations, goals, and contextual supports or barriers, and to expectancy—value theory [28], which emphasises expected success and the subjective value of an activity. The profile variables used here—examination results, achievements, and stated interests—represent observable proxies for only part of these constructs; they do not directly measure self-efficacy, perceived value, family influence, or contextual barriers. Accordingly, the proposed system should be viewed as a profile-to-programme decision-support layer that complements, rather than replaces, theory-informed counselling.

3. Materials and Methods

The methodological scheme of the study includes six interrelated components:

1. A formalised description of the applicant digital profile and the fixed set of available fields of study;
2. A unified protocol for generating prompts for the evaluated local LLMs;
3. The generation of a ranked list of fields of study and brief explanations;
4. The implementation of a non-generative TF–IDF cosine content-based baseline;

5. A comparative evaluation using ranking-quality, structured-output, and computational metrics;
6. An additional robustness evaluation using processed applicant profiles with reduced direct lexical overlap.

The main object of comparison is the agreement between the priorities generated by the model and the applicant's actual field of admission. The system should rank programmes according to their substantive correspondence to the digital profile and present several justified alternatives, while the final decision remains with the applicant and the career guidance specialist.

The experimental hypothesis is that locally executable LLMs are able to extract semantic relationships between applicants' academic results, interests, achievements, and the content of degree programmes; however, different models differ substantially in their ability to rank alternatives correctly. An additional hypothesis is that the best model for the considered scenario is not necessarily the largest in terms of the number of parameters: practical suitability is determined by a combination of the accuracy of the top recommendation, the quality of placing the actually selected programme in the upper part of the ranked list, result stability, and inference time. A further hypothesis is that the relative advantage of lexical retrieval decreases when relevant competencies are described indirectly, whereas selected LLMs preserve a larger proportion of their ranking quality through contextual interpretation.

The complete experimental workflow is summarised in Figure 1.

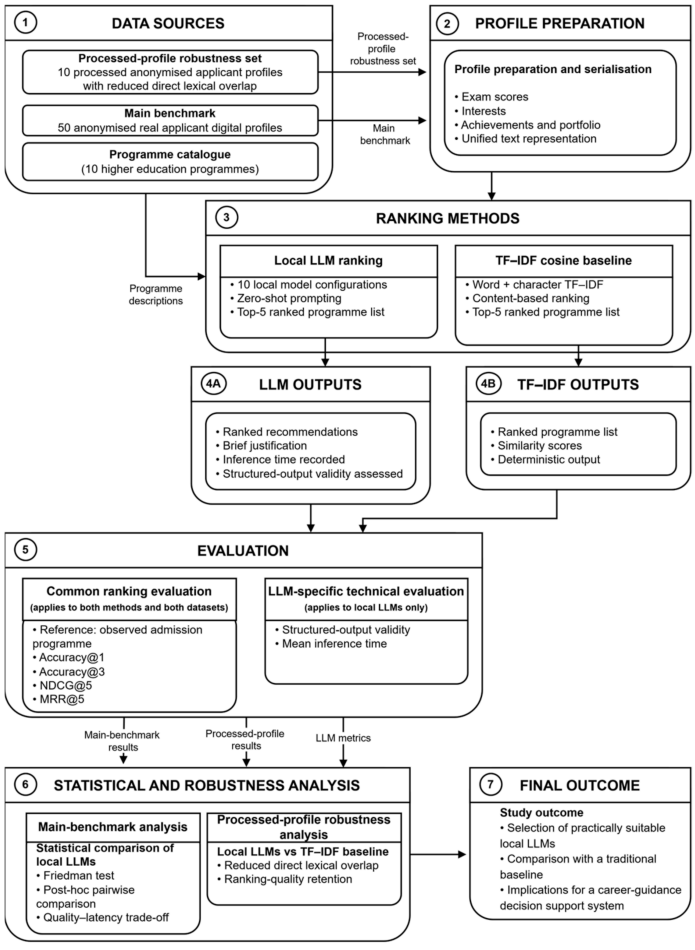


Figure 1. Experimental workflow for comparing local LLMs with the TF–IDF baseline and for conducting the additional robustness evaluation.

3.1. Formal Statement of the Task of Ranking Degree Programmes

In this study, each field-of-study code is represented by one degree programme in the fixed institutional catalogue. For terminological consistency, the recommended alternatives are hereafter referred to as degree programmes.

Let there be a set of applicant digital profiles

$$\mathcal{A} = \{a_1, a_2, \ldots, a_N\}, \tag{1}$$

where N is the number of analysed profiles. Each profile $a_i \in \mathcal{A}$ is represented as a tuple

$$a_i = (\mathbf{e}_i, \mathbf{h}_i, \mathcal{G}_i, t_i), \tag{2}$$

where $\mathbf{e}_i$ is a vector of the applicant's academic indicators, including entrance examination or Unified State Examination results; $\mathbf{h}_i$ is the set of explicitly specified interests and preferred topics; $\mathcal{G}_i = \{g_{i1}, g_{i2}, \ldots, g_{iq_i}\}$ is the set of achievements, projects, Olympiads, and other portfolio elements; and t_i is an aggregated textual description of the profile used for semantic analysis.

Let the set of fields of study be denoted as

$$\mathcal{P} = \{p_1, p_2, \ldots, p_M\}, \tag{3}$$

where M is the number of available degree programmes. Each field of study $p_j \in \mathcal{P}$ is represented as

$$p_j = (c_j, n_j, \mathcal{D}_j, \mathcal{C}_j, s_j), \tag{4}$$

where c_j is the code of the field of study; n_j is its title; $\mathcal{D}_j$ is the set of key disciplines; $\mathcal{C}_j$ is the set of competencies and professionally significant skills to be developed; and s_j is a textual description of the substantive orientation of the programme.

The set of admissible field-of-study codes is defined as

$$\mathcal{Y} = \{c_j \mid p_j \in \mathcal{P}\}. \tag{5}$$

For each profile a_i, it is necessary to obtain an ordered list of programmes

$$\pi_i^{(m)} = \left(p_{i,(1)}^{(m)}, p_{i,(2)}^{(m)}, \ldots, p_{i,(K)}^{(m)}\right), \tag{6}$$

where m denotes the ranking method under evaluation, including either a local LLM configuration or the non-generative TF–IDF baseline, K is the size of the recommendation list, and $p_{i,(r)}^{(m)}$ is the programme placed by model m at position r for profile a_i.

The task can be expressed through a relevance function

$$\rho : \mathcal{A} \times \mathcal{P} \to \mathbb{R}, \rho(a_i, p_j) = \rho_{ij}, \tag{7}$$

which determines the degree of correspondence between the applicant digital profile and the degree programme. A correct ranking then satisfies the condition

$$\rho\left(a_i, p_{i,(1)}^{(m)}\right) \geq \rho\left(a_i, p_{i,(2)}^{(m)}\right) \geq \ldots \geq \rho\left(a_i, p_{i,(K)}^{(m)}\right). \tag{8}$$

In this study, the relevance function is approximated by the ranking method under evaluation. For local LLMs, relevance is inferred from a unified prompt without additional fine-tuning. For the non-generative baseline, relevance is calculated from hybrid word-level and character-level TF–IDF cosine similarity.

For quantitative evaluation, each profile is associated with the actually selected degree programme p_i^* and its code y_i:

$$p_i^* \in \mathcal{P}, y_i = c(p_i^*) \in \mathcal{Y}. \quad (9)$$

Comparing the model output with the code of the actual field of admission y_i makes it possible to evaluate both the correctness of the first recommendation and the quality of placing this field within the recommendation list.

3.2. Formalisation of the Applicant Digital Profile

The applicant digital profile combines structured and semi-structured data. In contrast to models based only on examination results or questionnaire preferences, the proposed formulation provides for the simultaneous consideration of the following classes of features:

- Academic indicators characterising the applicant's basic preparation;
- Interests and thematic preferences reflecting the presumed direction of professional self-determination;
- Achievements and project activities serving as evidence of actually developed skills;
- Textual descriptions of works, technologies, and subject areas, which make it possible to establish semantic links with degree programmes.

Academic indicators are represented by a vector $\mathbf{e}_i = (e_{i1}, e_{i2}, \ldots, e_{id}) \in \mathbb{R}^d$, where the individual components correspond to results in profile-specific subjects. The applicant's interests are defined as a set of thematic descriptors $\mathbf{h}_i = \{h_{i1}, h_{i2}, \ldots, h_{iv_i}\}$. Each achievement $g_{il} \in \mathcal{G}_i$ can be described as

$$g_{il} = (u_{il}, r_{il}, \tau_{il}, d_{il}), \; l = 1, \ldots, q_i, \quad (10)$$

where u_{il} is the name of the event, competition, or project; r_{il} is the participation result; τ_{il} is a temporal or organisational characteristic; and d_{il} is a textual description of the completed work, the methods applied, or the technologies used.

In the ranking context, an achievement is considered a stronger confirmation of thematic orientation than a declaratively stated interest. For example, an indicated interest in programming is informative, whereas the description of a specific project or participation in an Olympiad is a more objective indicator of interest and ability in the field of IT. Similarly, information on the development of a robotic system or a control system may justify the formation of engineering orientations even in the absence of an exact terminological match in the title of the degree programme.

To pass the data to the model, the initial components of the profile are transformed into a consistent textual representation

$$T_{\mathcal{A}}(a_i) = \text{Serialize}_{\mathcal{A}}(\mathbf{e}_i, \mathbf{h}_i, \mathcal{G}_i, t_i), \quad (11)$$

where the function $\text{Serialize}_{\mathcal{A}}(\cdot)$ defines a unified format for representing data in the prompt. This transformation ensures a homogeneous task formulation for all models under study and allows the LLM to interpret both explicit features and semantic relationships within the portfolio.

In the DSS under study, a degree programme is considered a decision alternative characterised not only by a formal code but also by its content. For each programme, a textual-structured representation is used, including:

- The code and official title of the field of study;
- The list of key disciplines;
- Associated hard skills and subject-specific competencies;

- Thematic orientation and expected professional areas.

The textual representation of the programme is defined as

$$T_{\mathcal{P}}(p_j) \;=\; \text{Serialize}_{\mathcal{P}}\big(c_j, n_j, \mathcal{D}_j, \mathcal{C}_j, s_j\big). \tag{12}$$

In the implementation, $\text{Serialize}_{\mathcal{A}}(\cdot)$ and $\text{Serialize}_{\mathcal{P}}(\cdot)$ are deterministic template-based operations rather than learned encoders. They place each source field under a fixed textual label, preserve the original numerical and textual values, normalise whitespace, and concatenate the fields in a fixed order. They do not infer additional characteristics, vectorise the data, or include the reference admission code. Thus, serialisation ensures that every model receives semantically identical information in the same prompt structure.

The descriptions of degree programmes were formed on the basis of official information on the university's degree programmes, including the code and title of the field of study, the list of key disciplines, professionally significant skills, and the thematic orientation of training. The brief programme descriptions were prepared by the authors on the basis of official educational materials and were converted into a unified format to eliminate differences in the amount and structure of information between fields of study.

The experiment used a register of $M = 10$ fields of study (Table 1). It covers information technology, engineering, management, and economics-and-information educational areas that are sufficiently close in orientation, which complicates the task of selecting the optimal field and requires a careful analysis of the applicant digital profile. This set of alternatives makes it possible to evaluate not only the recognition of obvious correspondences, for example between programming projects and IT programmes, but also the model's ability to distinguish between related trajectories with overlapping sets of disciplines and competencies.

This representation is fundamental to the methodology of the study: the LLM must select a field not by matching the programme title with the applicant's interest, but by assessing the substantive correspondence between the applicant's achievements and the disciplines and competencies of the programme.

For each applicant profile and programme set, a unified prompt is formed, containing the system role, the task description, the full list of alternatives, the digital profile, and the constraint on the output structure. Let Γ denote the prompt generation function. Then the prompt for profile a_i is defined as

$$q_i \;=\; \Gamma\Big(I, T_{\mathcal{A}}(a_i), \{T_{\mathcal{P}}(p_j)\}_{j=1}^{M}, O\Big), \tag{13}$$

where I is the instruction that defines the role of the model as an expert in career guidance matching, and O is the specification of the required output format.

For each model m under study, the generative response is defined as

$$z_i^{(m)} \;=\; \text{LLM}_m(q_i; \boldsymbol{\lambda}), \tag{14}$$

where $\boldsymbol{\lambda}$ is the general set of inference parameters. From the response, an ordered list of programme codes is extracted:

$$\hat{\mathbf{y}}_i^{(m)} \;=\; \mathit{ExtractRanking}(z_i^{(m)}) \;=\; (y_{i,1}^{(m)}, y_{i,2}^{(m)}, \ldots, y_{i,K}^{(m)}) \in \mathcal{Y}^K. \tag{15}$$

The size of the recommendation list is set to $K = 5$. The first recommendation is then interpreted as the most appropriate field, while the remaining positions represent admissible alternatives for discussion with the applicant.

Table 1. Fields of study used in the experimental ranking task.

Group of Fields	Code and Title of the Field of Study	Brief Description of Disciplines
Information technology	09.03.01 Computer Science and Computer Engineering	Java, Python, C++; algorithms and optimisation; computer graphics and 3D modelling; databases; mobile and web technologies; AI and CAD.
Information technology	09.03.02 Information Systems and Technologies	Web programming; computer networks; integration of information systems and APIs; data analysis and statistics; system programming; machine learning.
Information technology	09.03.03 Applied Informatics	Applied AI and neural networks; big data; 3D and animation; UX/UI; VR/AR; Unity and game application development.
Information technology	10.05.03 Information Security of Automated Systems	Network administration; protection of operating systems and databases; information security management and risk analysis; secure automated systems; telecommunications and signals.
Radio engineering and infocommunications	11.03.01 Radio Engineering	Electromagnetic fields and electrodynamics; digital signal processing; mobile communications; modelling of radio engineering devices; television; microprocessors.
Radio engineering and infocommunications	11.03.02 Infocommunication Technologies and Communication Systems	Circuit theory and radio measurements; power supply; microwave engineering and antennas; transmitting and receiving devices; control and diagnostics of radio-electronic systems; television.
Energy engineering	13.03.01 Thermal Power Engineering and Heat Engineering	Heat engines; boiler installations and steam generators; energy saving; automation of thermal power engineering; water and gas supply; HVAC.
Control and automation	27.03.03 Systems Analysis and Control	Statistical and intelligent data analysis; systems engineering; modelling; control theory; organisational systems; metrology.
Control and automation	27.03.04 Control in Technical Systems	Microcontrollers; metrology; automation systems; optimal control; industrial networks; SCADA; databases for control systems.
Economics and information systems	38.03.05 Business Informatics	Economic analysis and business analytics; IT services and the life cycle of information systems; web and PR technologies; projects; fundamentals of AI and Python.

AI, artificial intelligence; CAD, computer-aided design; API, application programming interface; UX, user experience; UI, user interface; VR, virtual reality; AR, augmented reality; HVAC, heating, ventilation, and air conditioning; SCADA, supervisory control and data acquisition; PR, public relations.

The prompt structure includes the following semantic constraints:

1. The model must match the applicant's achievements with the disciplines and skills of the programmes, rather than relying exclusively on programme titles;
2. Specific achievements and project activities must be treated as stronger evidence than general declarations of interests;
3. The model must return an ordered list only from the available catalogue of degree programmes;
4. A brief explanation of correspondence to the applicant profile is required for each field;
5. The result must be presented in a structured format suitable for subsequent display in the DSS interface.

The generalised input and output structure is as follows:

```
{
  "input": {
    "programs": [
      {
        "code": "XX.XX.XX",
        "name": "Field title",
        "disciplines": ["..."],
        "competencies": ["..."]
      }
    ],
    "applicant": {
      "interests": ["..."],
      "exam_scores": {"...": 0},
      "achievements": ["..."]
    }
  },
  "output": {
    "ranking": [
      {
        "code": "XX.XX.XX",
        "reason": "Explanation of correspondence"
      }
    ]
  }
}
```

The use of a unified prompt structure for all models is necessary to ensure the comparability of the results.

3.3. Compared Local Language Models

The experiment considers ten locally executable models in GPT-Generated Unified Format (GGUF) used by llama.cpp (Table 2), which makes it possible to deploy the models within the computational environment of the educational organisation without transferring digital profiles to an external cloud provider. The comparison includes models of different size classes and quantisation variants, which makes it possible to assess the influence of these parameters on the final ranking quality.

Table 2. Local LLMs included in the comparative experiment.

No.	Model Configuration	Stated Parameter Size Class	Quantisation
1	DeepSeek-R1-Distill-Qwen-7B-Q4_K_M	7B	Q4_K_M
2	gemma-4-E4B-it-Q8_0	4B	Q8_0
3	gpt-oss-20b-MXFP4	20B	MXFP4
4	LFM2-24B-A2B-Q4_K_M	24B	Q4_K_M
5	Ministral-3-14B-Reasoning-2512-Q4_K_M	14B	Q4_K_M
6	NVIDIA-Nemotron-3-Nano-4B-Q8_0	4B	Q8_0
7	Qwen3.5-9B-Q4_K_M	9B	Q4_K_M
8	Qwen3.5-9B-Q8_0	9B	Q8_0
9	Qwen3.6-27B-UD-IQ3_XXS	27B	IQ3_XXS
10	Qwen3.6-35B-A3B-UD-Q4_K_S	35B	Q4_K_S

All models are used in inference-only mode: no additional training, fine-tuning, or adaptation to the corpus of degree programmes is performed in this experiment. This design makes it possible to evaluate the initial ability of local models to interpret a digital profile and rank fields of study under a unified zero-shot protocol.

The compared entries were treated as complete deployed configurations rather than as isolated base architectures. They differ simultaneously in model family, pretraining and instruction-tuning history, parameter size, quantisation, and implementation characteristics. Consequently, the experiment estimates configuration-level performance and does not permit causal attribution of an observed advantage to a single factor such as the training corpus, architecture, or quantisation method.

The experiments were performed using the following computational configuration: CPU, AMD Ryzen 9 5900X 12 cores (Advanced Micro Devices, Inc., Santa Clara, CA, USA); GPU, NVIDIA GeForce RTX 4080 (NVIDIA Corporation, Santa Clara, CA, USA); VRAM, 16 GB; RAM, 64 GB DDR4; operating system, Windows 10 64-bit. The models were run locally using llama.cpp version b8861.

3.4. TF–IDF Cosine Content-Based Baseline

To provide a traditional non-LLM reference point, a content-based ranking baseline based on term frequency–inverse document frequency vectorisation and cosine similarity was added to the experimental design [29]. No manually defined keyword-to-programme mappings, programme-specific lexicons, or classification rules were supplied to the baseline.

For each applicant, the textual representation included interests, Unified State Examination results, achievement titles, achievement outcomes, and achievement descriptions. Applicant names, cities, identifiers, reference programme codes, and other fields capable of revealing the target label were excluded. For each degree programme, a textual representation was formed from its code, title, description, disciplines, competencies, and other catalogue information available to all evaluated methods.

Two TF–IDF representations were calculated. The word-level representation used unigrams and bigrams, while the character-level representation used character n-grams of length three to five. Character n-grams were included to reduce sensitivity to Russian inflectional forms, spelling variants, and differences in word endings. The similarity between applicant a_i and programme p_j was calculated as

$$s_{hybrid}(a_i, p_j) = 0.45 s_{word}(a_i, p_j) + 0.55 s_{char}(a_i, p_j), \tag{16}$$

where

$$s(a_i, p_j) = \frac{x_i \cdot u_j}{\|x_i\| \|u_j\|} \tag{17}$$

is the cosine similarity between the corresponding TF–IDF vectors.

The weighting coefficients and n-gram ranges were fixed before evaluation and were not optimised using the reference labels.

The word-level and character-level TF–IDF vectorisers were fitted on the combined unlabeled corpus of applicant-profile texts and programme-description texts separately for each evaluated dataset.

All candidate programmes were ranked in descending order of s_{hybrid}. The method did not use LLM inference, prompting, external knowledge, supervised training, or access to the reference label during vectorisation and ranking. Reference programme codes were used only after ranking to calculate Accuracy@1, Accuracy@3, NDCG@5, and MRR@5. Because the TF–IDF baseline directly returned a deterministic ranked list rather than generating a natural-language response, structured-output validity was not applicable to this method.

The baseline was reported as an external deterministic reference point. It was not included in the LLM-specific post hoc statistical comparison, which was originally designed to compare the evaluated local generative-model configurations.

3.5. Experimental Sample and Comparison Procedure

The experimental dataset includes $N = 50$ applicant digital profiles and $M = 10$ degree programmes. The profiles were obtained from institutional admission and portfolio records for 2025 year admission cycle. Each profile contains interests, academic indicators, and a portfolio of achievements. For each profile, the code of the actual field of admission y_i is specified and used exclusively at the evaluation stage. The actually selected field of admission was retained as the observed reference label. All direct personal identifiers were removed before the data were used in the experiment.

In this study, the actual field of admission is treated as an observed reference label rather than as an objective ground truth of educational suitability. The applicant's final choice may be affected by admission quotas, strategic considerations, financial and geographic constraints, personal circumstances, and available opportunities. Therefore, the evaluation measures how well each model reproduces the observed programme-selection outcome under a fixed catalogue, not whether the model identifies the universally optimal educational trajectory. A more complete validation would require expert-graded relevance labels assigned by career guidance specialists or academic advisors.

The comparison includes $L = 10$ local models. Each model processes the same set of 50 profiles and receives the same register of degree programmes. Therefore, the total number of ranking procedures is $N \times L = 50 \times 10 = 500$.

For each experiment, the following sequence is performed:

- The digital profile a_i and the full list of programmes $\mathcal{P}$ are transformed into a unified prompt q_i;
- Model m generates a structured ranked list of $K = 5$ fields of study;
- The sequence of codes $\hat{\mathbf{y}}_i^{(m)}$ is extracted from the result;
- The ranking is compared with the code y_i;
- Quality metrics are calculated for the individual profile and aggregated metrics for the model;
- The recommendation generation time and the possibility of obtaining a structured result are recorded.

To ensure model comparability, a unified generation mode is used: $\boldsymbol{\lambda} = (C, T, P, R, K_g)$, where $C = 8192$ is the context window size; $T = 0.1$ is the generation temperature; $P = 0.9$ is the nucleus sampling parameter; $R = 1.1$ is the repetition penalty; and $K_g = 40$ is the top-k generation parameter. The maximum response length is limited to 8192 tokens. The low temperature value was chosen to reduce random variability in the response and improve the stability of the ranked output.

Because the system must return an ordered set of degree programmes rather than arbitrary text of fields of study, the evaluation is performed using ranking metrics. Let $r_i^{(m)}$ be the position of code y_i in the output of model m. If the field is absent from the K recommendations, then $r_i^{(m)} = \infty$.

To prevent target leakage from the input data, direct references to the actual field of admission, competition group number, selected degree programme, and other features that uniquely disclose the admission result were excluded. The models received only the applicant digital profile and the general catalogue of available programmes. Before use in the experiment, the digital profiles were anonymised: surnames, names, contact details, application identifiers, and other direct personal identifiers were removed. Where necessary, textual descriptions of achievements and projects were generalised in such a way

as to preserve the substantive features of the profile while reducing the risk of identifying a specific applicant.

The main metrics for evaluating the models are considered below.

Accuracy@1 reflects the proportion of profiles for which the model placed the actually selected field of study in the first position:

$$\text{Accuracy@1}(m) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[r_i^{(m)} = 1\right], \tag{18}$$

where $\mathbb{I}[\cdot]$ is the indicator function.

For the task under consideration, Accuracy@1 is the primary metric for the accuracy of the priority recommendation. A high value of this indicator means that the model is able to identify the field of study that should be presented to the applicant as the most relevant to their profile.

Accuracy@3 characterises the proportion of cases in which the actually selected field of study is present among the first three recommendations:

$$\text{Accuracy@3}(m) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[r_i^{(m)} \le 3\right]. \tag{19}$$

This indicator is interpreted as the model's ability to preserve the actual field of admission among the top three recommendations. Unlike Accuracy@1, Accuracy@3 allows for the final choice to be refined during consultation with an expert.

To account for the position of the actually selected field of study within the whole recommendation list, the Normalized Discounted Cumulative Gain metric is used. In the present study, binary relevance is applied [30]:

$$\text{rel}_{i,r}^{(m)} = \begin{cases} 1, & \text{if } \hat{y}_{i,r}^{(m)} = y_i, \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

For a list of length $K = 5$, the discounted cumulative gain is defined as

$$\text{DCG@5}_i^{(m)} = \sum_{r=1}^{5}\frac{\text{rel}_{i,r}^{(m)}}{\log_2(r+1)}. \tag{21}$$

Since one actually selected field of study is specified for each profile, the ideal list places it in the first position, and $\text{IDCG@5}_i = 1$. Therefore,

$$\text{NDCG@5}(m) = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{DCG@5}_i^{(m)}}{\text{IDCG@5}_i}. \tag{22}$$

In the present formulation, the contribution of an individual profile can be written in simplified form as

$$NDCG@5_i^{(m)} = \begin{cases} \frac{1}{\log_2\left(r_i^{(m)}+1\right)}, & r_i^{(m)} \le 5, \\ 0, & r_i^{(m)} > 5 \text{ or } r_i^{(m)} = \infty. \end{cases} \tag{23}$$

The NDCG@5 metric makes it possible to distinguish between models that included the target field in the output but placed it at different positions. This is particularly important for a DSS, since a recommendation in the second or fifth position does not have the same weight as a recommendation in the first position. Since the present study uses

one observed reference programme for each profile, NDCG@5 characterises not the quality of all alternatives in the list, but the position of the actual field of admission within the top-5 output. If several expert-relevant programmes were available, this metric could be extended to a graded relevance scheme; however, such annotation is beyond the scope of the present study.

The mean reciprocal rank of the first occurrence of the actually selected field of study is additionally used [31]:

$$\mathrm{MRR@5}(m) \;=\; \frac{1}{N}\sum_{i=1}^{N}\frac{1}{r_i^{(m)}}. \tag{24}$$

MRR@5 has a direct interpretation in relation to career guidance: the higher its value, the closer to the beginning of the recommendation list the model systematically places the actually selected field of study. Unlike Accuracy@1, this metric accounts for partial success when the recommendation matching the applicant's actual choice is not in the first position but remains among the top alternatives.

The use of a single positive reference label is a simplified evaluation setting. In real career guidance, several programmes may be simultaneously relevant to the same applicant profile. Consequently, Accuracy@k, NDCG@5, and MRR@5 should be interpreted as benchmark metrics for agreement with the observed reference programme, not as a complete measure of multidimensional educational suitability.

In a DSS, the quality of semantic ranking is a necessary but not sufficient condition for using the model. A local LLM must also generate the result in a form suitable for machine processing and display in the user interface, while the response time must correspond to the advisory scenario.

For each response, the structured validity indicator is defined as

$$v_i^{(m)} \;=\; \begin{cases} 1, & \text{if the model has generated an extractable ranked list of codes,} \\ 0, & \text{overwise.} \end{cases} \tag{25}$$

The aggregated validity of a model is calculated as

$$V(m) \;=\; \frac{1}{N}\sum_{i=1}^{N} v_i^{(m)}. \tag{26}$$

This indicator does not replace the quality metrics: a model may generate a structured response consistently while ranking fields incorrectly. Nevertheless, $V(m)$ must be considered when selecting a DSS component, since failure to extract a recommendation is equivalent to system failure for the corresponding profile.

Model performance is evaluated by the mean time required to obtain a recommendation:

$$\overline{\tau}(m) \;=\; \frac{1}{N}\sum_{i=1}^{N}\tau_i^{(m)}, \tag{27}$$

where $\tau_i^{(m)}$ is the response generation time of model m for profile a_i. The standard deviation of response time may additionally be used:

$$\sigma_\tau(m) \;=\; \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\tau_i^{(m)} - \overline{\tau}(m)\right)^2}. \tag{28}$$

Thus, a model is evaluated according to two classes of characteristics:

$$\mathbf{Q}(m) = (\text{Accuracy@1}(m), \text{Accuracy@3}(m), \text{NDCG@5}(m), \text{MRR@5}(m)), \quad (29)$$

which describe ranking quality, and

$$\mathbf{E}(m) = (V(m), \overline{\tau}(m), \sigma_\tau(m)), \quad (30)$$

which describe the possibility of integrating the model into an applied system.

Since the same set of applicant digital profiles is processed by all models under study, the ranking quality of different LLMs can be compared directly for each profile. Under these conditions, the use of methods for independent samples does not reflect the design of the experiment.

Let $s_i^{(m)}$ be the individual quality score of model m on profile a_i. The position-sensitive NDCG@5 indicator is used as the profile-level score, since it accounts not only for the inclusion of the target field in the list but also for its position in the ranking. For each profile, the values $s_i^{(m)}$ are ranked across all L models, after which the mean rank of each model is calculated:

$$\overline{R}_m = \frac{1}{N} \sum_{i=1}^{N} R_i^{(m)}, \quad (31)$$

where $R_i^{(m)}$ is the rank of model $R_i^{(m)}$ on the i-th profile.

To test the global hypothesis of equal model quality, the non-parametric Friedman test is used, which corresponds to a repeated-measures design. For each profile, models are ranked by their NDCG@5 value, with a smaller rank corresponding to a higher recommendation quality. The null and alternative hypotheses are as follows: H_0 the local models under study do not differ in ranking quality; H_1 at least one model differs from the others in ranking quality.

The decision on the presence of statistically significant differences is made at the significance level. After global differences have been identified, post hoc comparisons are performed between the model with the highest mean NDCG@5 value and the other configurations. The Wilcoxon signed-rank test for paired samples is used for this purpose. The resulting p-values are adjusted using the Holm method to control the error associated with multiple comparisons. The statistical conclusion is considered jointly with the applied metrics. Even if a model demonstrates a high mean rank, the decision on its suitability for a DSS must take into account the accuracy of the first recommendation, the stability of structured output, and inference time.

For the binary indicators Accuracy@1 and Accuracy@3, Cochran's Q test is additionally applied, since the same profiles are evaluated by all models and the result for each model on each profile has a binary form: agreement or disagreement with the actual field of admission at a given top-(k) position. These tests are treated as secondary and are used to provide descriptive support for the main conclusion based on NDCG@5.

In the applied scenario, the first-priority requirement is the ability of the model to correctly identify the primary recommended field; therefore, the basic criterion is the maximisation of Accuracy@1. At the same time, in a scenario where the system is used not to identify a single most suitable field, but to form a set of alternatives for a specialist, Accuracy@3 becomes additionally important. In this case, the model can be considered useful for preliminary consultation even with a lower Accuracy@1, provided that it consistently places the relevant field in the upper part of the list.

All experimental outputs were stored as per-profile structured records containing the model identifier, generated ranked list, extracted programme codes, inference time, and

extraction status. Metric aggregation and statistical testing were performed using author-developed Python scripts. Tabular processing was carried out with pandas and NumPy; statistical tests were implemented using standard Python scientific-computing libraries. Figure 1 was prepared by the authors as a conceptual workflow diagram. Figures 2–5 were generated automatically in Python using Matplotlib (version 3.7.2) from the same profile-level records and aggregated tables used in the Section 4. Figure 2 was checked against Table 3, Figure 3 against Table 4, Figure 4 against the content-group aggregates reported in Section 4.3, and Figure 5 against the mean NDCG@5 ranks reported in Section 4.4.

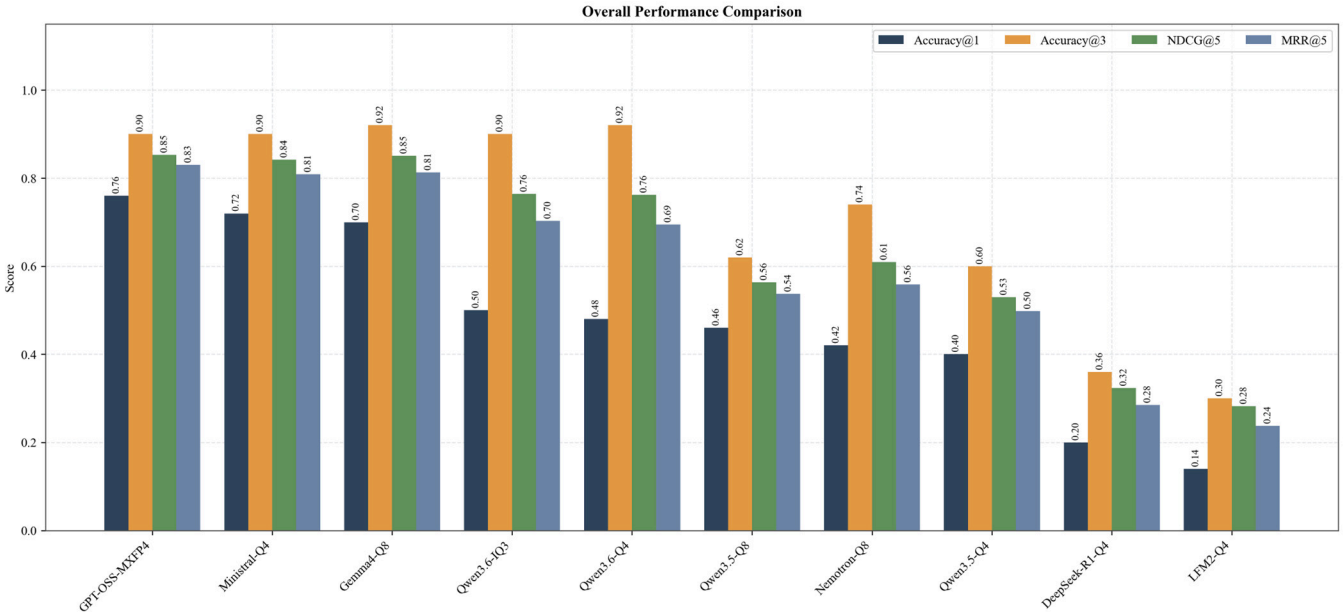


Figure 2. Comparison of local LLMs in ranking degree programmes according to the main metrics.

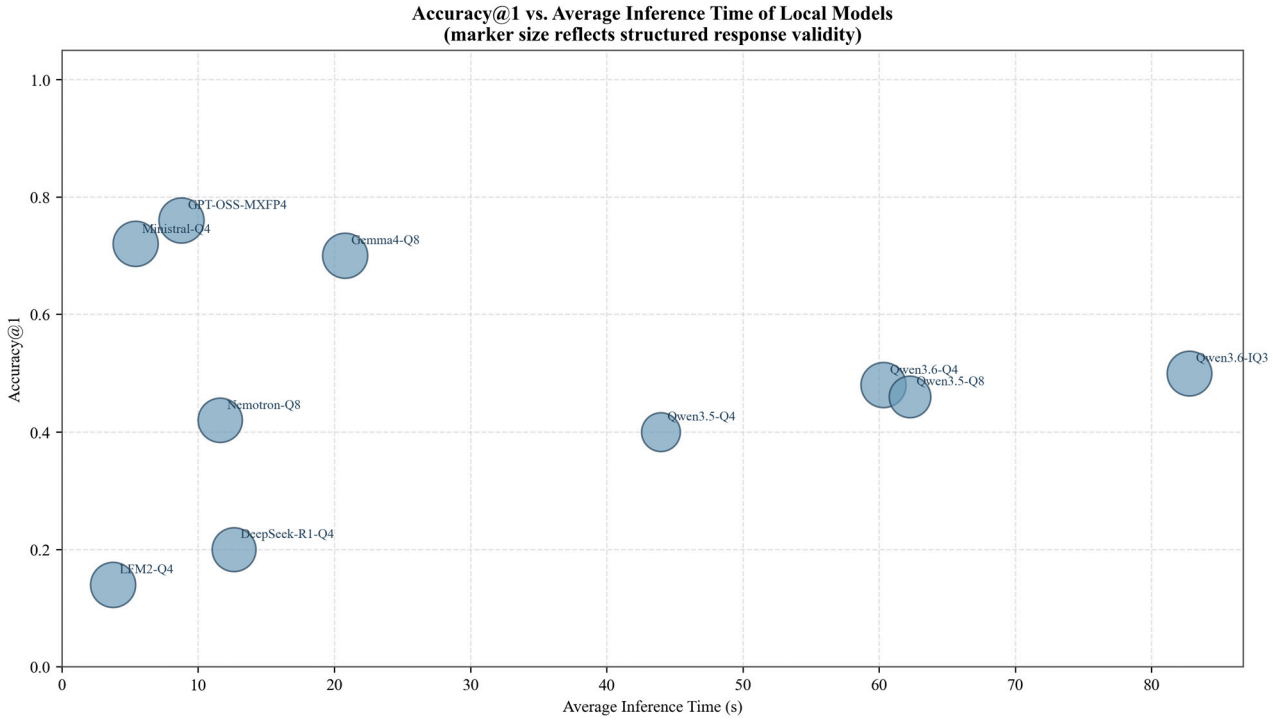


Figure 3. Relationship between Accuracy@1 and mean inference time for local models; marker size reflects structured-output validity.

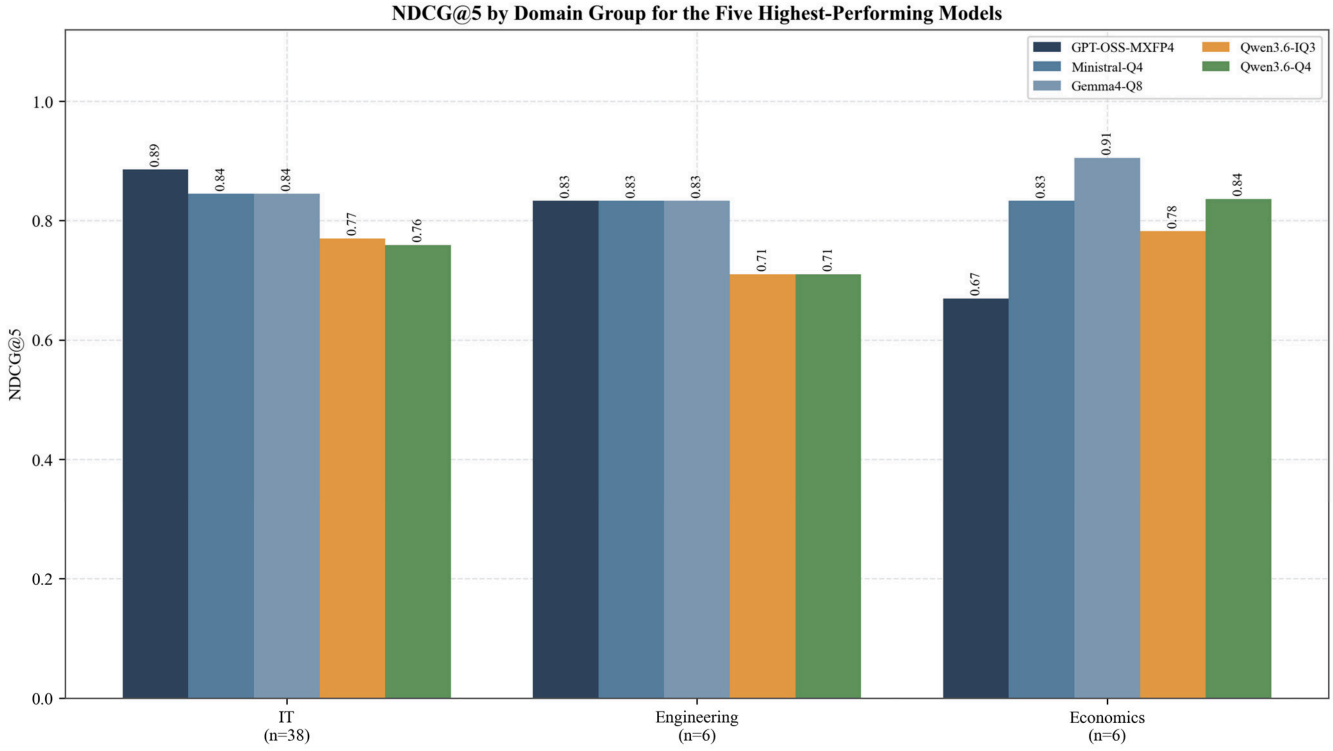


Figure 4. NDCG@5 values by content group of profiles for the five models with the highest Accuracy@1.

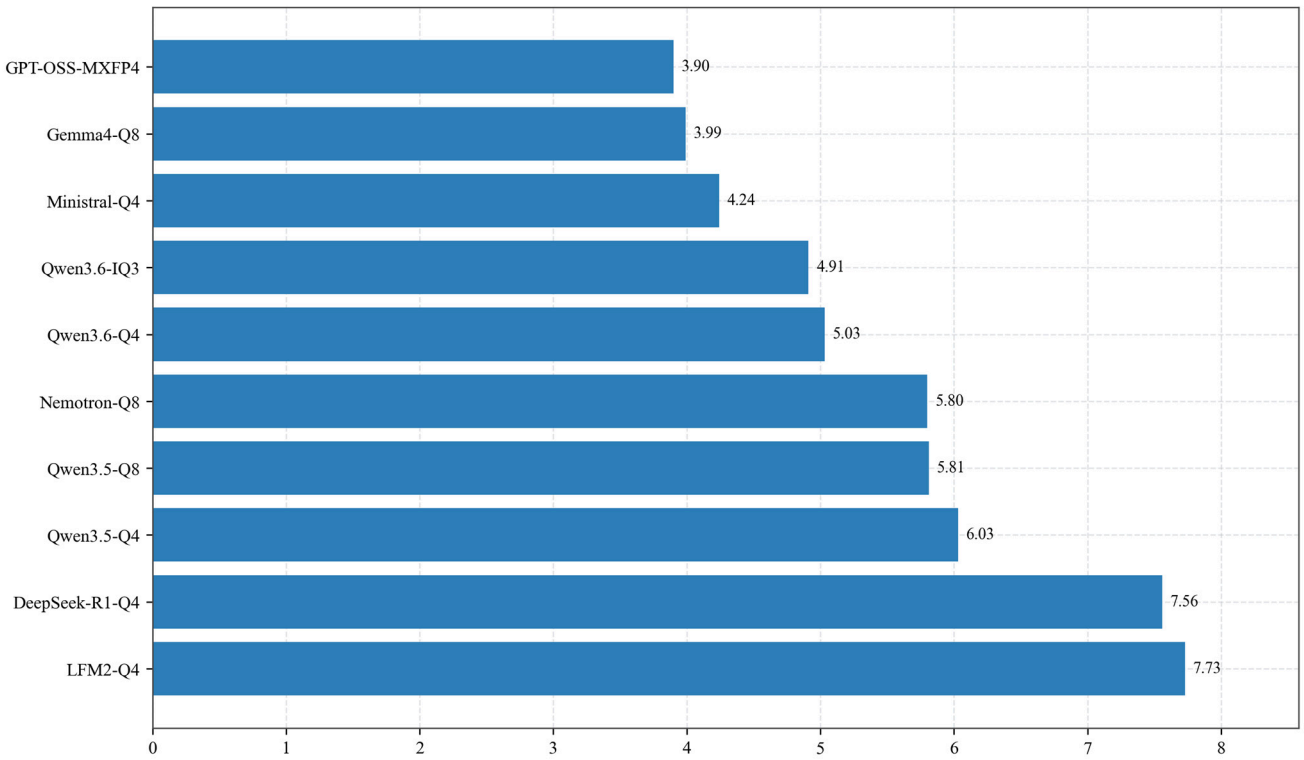


Figure 5. Mean ranks of local LLMs according to NDCG@5; a smaller value corresponds to a better result.

Table 3. Main quality metrics of local LLM ranking configurations.

Model	Accuracy@1	95% CI	Accuracy@3	95% CI	NDCG@5	95% Bootstrap CI	MRR@5	95% Bootstrap CI	Structured Validity
gpt-oss-20b-MXFP4	**0.76**	0.626–0.857	0.90	0.786–0.957	**0.8534**	0.7636–0.9283	**0.8307**	0.7367–0.9140	**1.00**
Ministral-3-14B-Reasoning-2512-Q4_K_M	0.72	0.583–0.825	0.90	0.786–0.957	0.8421	0.7605–0.9168	0.8090	0.7163–0.8950	**1.00**
gemma-4-E4B-it-Q8_0	0.70	0.562–0.809	**0.92**	0.812–0.968	0.8508	0.7768–0.9162	0.8133	0.7283–0.8917	**1.00**
Qwen3.6-27B-UD-IQ3_XXS	0.50	0.366–0.634	0.90	0.786–0.957	0.7644	0.6863–0.8365	0.7033	0.6167–0.7883	0.98
Qwen3.6-35B-A3B-UD-Q4_K_S	0.48	0.348–0.615	**0.92**	0.812–0.968	0.7626	0.6901–0.8319	0.6947	0.6087–0.7780	**1.00**
Qwen3.5-9B-Q8_0	0.46	0.330–0.596	0.62	0.482–0.741	0.5634	0.4359–0.6886	0.5373	0.4107–0.6633	0.84
NVIDIA-Nemotron-3-Nano-4B-Q8_0	0.42	0.294–0.558	0.74	0.604–0.841	0.6096	0.4974–0.7143	0.5583	0.4417–0.6683	0.96
Qwen3.5-9B-Q4_K_M	0.40	0.276–0.538	0.60	0.462–0.724	0.5296	0.4069–0.6543	0.4983	0.3783–0.6233	0.72
DeepSeek-R1-Distill-Qwen-7B-Q4_K_M	0.20	0.112–0.330	0.36	0.241–0.499	0.3232	0.2153–0.4358	0.2847	0.1800–0.3957	0.94
LFM2-24B-A2B-Q4_K_M	0.14	0.070–0.262	0.30	0.191–0.438	0.2821	0.1818–0.3861	0.2373	0.1433–0.3380	**1.00**

CI, confidence interval. Higher values indicate better ranking quality.

Table 4. Inference-time characteristics of the evaluated local LLMs.

Model	Mean Time, s	SD, s	Min, s	Max, s
gpt-oss-20b-MXFP4	8.78	2.09	4.95	14.30
Ministral-3-14B-Reasoning-2512-Q4_K_M	5.38	0.35	4.51	6.33
gemma-4-E4B-it-Q8_0	20.78	2.13	16.35	25.92
Qwen3.6-27B-UD-IQ3_XXS	82.79	9.59	60.80	98.26
Qwen3.6-35B-A3B-UD-Q4_K_S	60.32	6.20	46.54	72.69
Qwen3.5-9B-Q8_0	62.27	11.24	15.92	73.59
NVIDIA-Nemotron-3-Nano-4B-Q8_0	11.62	6.91	5.52	32.32
Qwen3.5-9B-Q4_K_M	43.97	6.08	24.16	50.89
DeepSeek-R1-Distill-Qwen-7B-Q4_K_M	12.62	2.89	9.03	22.45
LFM2-24B-A2B-Q4_K_M	3.74	0.12	3.52	4.02

SD, standard deviation. Lower inference time is preferable.

Figures presenting model quality, inference time, content-group performance, and mean ranks were generated automatically from the same aggregated data tables used in the Section 4. No plotted values were manually edited after figure generation. The validation of the figures consisted of checking that each plotted value corresponded to the aggregated metric values reported in the tables. Structured-output validation was performed before metric calculation: a response was treated as valid only when a non-empty ranked list of study-programme codes could be extracted from the model output and matched to the allowed programme catalogue.

3.6. Additional Robustness Set Based on Processed Applicant Profiles

To examine the robustness of the evaluated approaches under reduced direct lexical correspondence, an additional set was formed from the anonymised data of ten applicants. These profiles were not used in the main 50-profile benchmark. The original profile information was manually processed to make the textual representation more complex and

less terminologically aligned with the programme catalogue, while preserving the actually selected field of admission as the reference label.

The processing included reformulating professional experience in indirect and partly informal language, retaining achievements related to adjacent subject areas, and reducing the use of explicit programme-specific terminology. The purpose was to reproduce a realistic situation in which an applicant describes relevant experience without using the same formal vocabulary as the educational-programme catalogue.

The additional set represented five closely related fields of study, with two applicant profiles corresponding to each field. All ten programmes from the main experiment remained available as ranking alternatives. The processed profiles and their reference labels were fixed before the comparative run and were not modified after observing the outputs of TF–IDF or the LLMs.

The processed-profile robustness profiles were designed according to the following predefined principles: profiles contained achievements from adjacent or potentially confusing domains; relevant experience was described in heterogeneous, partly informal language. The actually selected field of admission was retained as the primary reference programme. A secondary plausible programme was recorded only as an additional diagnostic indicator.

For example, information-security competence was described through unauthorised access to documents, analysis of event logs, separation of user privileges, and investigation of unusual login activity, without directly using terms such as "information security", "intrusion detection", or "cybersecurity". Similarly, business-informatics competence was represented through analysis of organisational losses, process redesign, management indicators, costs, and evaluation of the expected effect of digitalisation.

4. Results

During the experiment, each of the ten locally executable language models generated a ranked list of fields of study for the same set of applicant digital profiles. The results were evaluated against the actually selected field of study using four ranking quality indicators: Accuracy@1, Accuracy@3, NDCG@5, and MRR@5. In addition, the proportion of responses suitable for extracting a structured result was evaluated as an indicator of structured-output validity.

4.1. Evaluation of Model Accuracy

The summary metric values are presented in Table 3. The results show substantial differences between the models according to the selected criteria. Thus, even under a unified task formulation, the same prompt, and the same set of degree-programme alternatives, local LLMs can analyse applicant digital profiles in markedly different ways. The best results for each metric are highlighted in bold.

Among the evaluated local LLMs, the highest observed value of the primary practical metric Accuracy@1 was obtained by the gpt-oss-20b-MXFP4 model: in 76% of the profiles, the actually selected field of study was placed in the first position in the model response. The second-best result for Accuracy@1 was obtained by Ministral-3-14B-Reasoning-2512-Q4_K_M, with a value of 0.72. The gemma-4-E4B-it-Q8_0 model ranked third in terms of primary-choice accuracy, with a value of 0.70. These three models therefore form the upper group in terms of the accuracy of the main recommendation, since the first recommendation is the most influential in the DSS-based selection of a field of study.

The remaining models showed lower results. It should be noted that the tested configurations of the 27B and 35B size classes did not obtain an advantage in terms of Accuracy@1 in this experiment. The lowest results for Accuracy@1 were obtained by

DeepSeek-R1-Distill-Qwen-7B-Q4_K_M and LFM2-24B-A2B-Q4_K_M, with values of 0.20 and 0.14, respectively.

The TF–IDF cosine baseline demonstrated strong performance on the main benchmark, achieving Accuracy@1 = 0.8800, Accuracy@3 = 1.0000, NDCG@5 = 0.9557, and MRR@5 = 0.9400. The correct reference programme was therefore ranked first for 44 of the 50 profiles and appeared within the first three positions for all evaluated profiles.

These results show that the main dataset contained a strong lexical and thematic correspondence between applicant-profile descriptions and educational-programme descriptions. Under such conditions, a lightweight information-retrieval method can provide highly competitive ranking quality without generative inference. Therefore, the main LLM comparison should not be interpreted as evidence that local LLMs are universally superior to traditional content-based approaches.

The next stage of the analysis focused on evaluating how consistently the actually selected field of study remained in the upper part of the recommendation list, since the final choice or applicant consultation may be based on several degree-programme alternatives rather than a single option. At this stage, Accuracy@3, NDCG@5, and MRR@5 are of particular importance. A comparison of the models according to these metrics is shown in Figure 2.

The gpt-oss-20b-MXFP4 model combines the highest Accuracy@1 value of 0.76 with a high Accuracy@3 value of 0.90. Thus, the model not only frequently includes the actually selected field of study in the recommended set, but also predominantly places it in the first position. Similar results were obtained by Ministral-3-14B-Reasoning-2512-Q4_K_M (Accuracy@1 = 0.72) and gemma-4-E4B-it-Q8_0 (Accuracy@1 = 0.70). These three models form the upper group in terms of primary recommendation accuracy.

For Accuracy@3, the highest value was obtained by gemma-4-E4B-it-Q8_0 and Qwen3.6-35B-A3B-UD-Q4_K_S (0.92). The gpt-oss-20b-MXFP4, Ministral-3-14B-Reasoning-2512-Q4_K_M, and Qwen3.6-27B-UD-IQ3_XXS models obtained a close value of 0.90. Therefore, several models quite frequently include the reference programme in the upper part of the list; however, this does not imply the same reliability when selecting a single primary recommendation.

Structured-output validity was 1.00 for gpt-oss-20b-MXFP4, Ministral-3-14B-Reasoning-2512-Q4_K_M, gemma-4-E4B-it-Q8_0, Qwen3.6-35B-A3B-UD-Q4_K_S, and LFM2-24B-A2B-Q4_K_M. This means that, in all 50 cases, a ranked list of recommendations could be automatically extracted from the responses of these models. This indicator characterises the technical extractability of the result, but it does not replace the assessment of the substantive correctness of the ranking.

Finally, the NDCG@5 and MRR@5 metrics were considered. The maximum values of these indicators were again obtained by gpt-oss-20b-MXFP4: 0.8534 and 0.8307, respectively. The Ministral-3-14B-Reasoning-2512-Q4_K_M and gemma-4-E4B-it-Q8_0 models lagged only slightly behind. Thus, all three models placed the relevant field in the upper part of the recommendation list. Among the remaining models, Qwen3.6-27B-UD-IQ3_XXS and Qwen3.6-35B-A3B-UD-Q4_K_S showed relatively high values: NDCG@5 values of 0.7644 and 0.7626, and MRR@5 values of 0.7033 and 0.6947, respectively. These results confirm that the presence of the reference programme among the top three results is not equivalent to high-quality ranking.

4.2. Evaluation of Model Performance

The operational performance of the models was also evaluated because inference time affects hardware requirements and the feasibility of multi-user deployment. The mean

inference time for the models under study is presented in Table 4, and its relationship with Accuracy@1 is visualised in Figure 3.

The quality–inference-time plot makes it possible to identify three practically distinct groups of models. The first group includes gpt-oss-20b-MXFP4 and Ministral-3-14B-Reasoning-2512-Q4_K_M. These models are characterised by both high primary recommendation accuracy and moderate processing time. The second group is represented by gemma-4-E4B-it-Q8_0. Although this model demonstrates high recommendation accuracy and quality, it is substantially slower than the alternative models in the upper group. The third group includes the larger Qwen3.6 models, which show substantial inference times, while their larger size did not lead to improved prediction accuracy. LFM2-24B-A2B-Q4_K_M requires separate consideration: this model showed the minimum mean response time, but also the lowest accuracy results.

The results indicate the absence of a simple linear relationship between model size and ranking quality. Therefore, the choice of an LLM for each considered scenario should be made empirically on the target task, rather than on the basis of the number of parameters or the model file size.

4.3. Analysis of Results by Content Groups of Profiles

To identify possible domain-related heterogeneity in recommendation quality, the five models with the highest Accuracy@1 values were analysed across three content groups of profiles. In the current sample, the IT group includes 38 profiles, the engineering group includes 6 profiles, and the economics-and-information group includes 6 profiles (Figure 4).

In the IT group, the highest NDCG@5 value was obtained by gpt-oss-20b-MXFP4 (0.8856). The gemma-4-E4B-it-Q8_0 and Ministral-3-14B-Reasoning-2512-Q4_K_M models showed close results: 0.8450 and 0.8448, respectively. Thus, the advantage of gpt-oss-20b-MXFP4 in the overall sample is largely consistent with its result in the largest content group.

In the engineering group, gpt-oss-20b-MXFP4, Ministral-3-14B-Reasoning-2512-Q4_K_M, and gemma-4-E4B-it-Q8_0 obtained identical values for the main metrics: Accuracy@1 = 0.8333 and NDCG@5 = 0.8333. However, this group contains only six profiles; therefore, the result should be interpreted as the absence of a detected difference in the available subsample, rather than as evidence of model equivalence for engineering programmes.

In the economics-and-information group, the highest NDCG@5 value was obtained by gemma-4-E4B-it-Q8_0 (0.9051). Ministral-3-14B-Reasoning-2512-Q4_K_M and Qwen3.6-35B-A3B-UD-Q4_K_S obtained values of 0.8333 and 0.8363, respectively, whereas gpt-oss-20b-MXFP4 obtained 0.6696. This result may indicate a dependence of ranking quality on the subject area of the profile, but it requires verification on a larger and more balanced sample.

Thus, the subgroup analysis does not alter the general conclusion regarding the high effectiveness of the three leading models, but it shows that a single aggregated metric may conceal differences in their behaviour across individual content groups.

4.4. Statistical Analysis of Model Performance

Formal statistical testing was performed on the full sample of 50 profiles (Table 5). The primary endpoint was NDCG@5, since it reflects the position of the reference programme in the recommendation list and accounts for the difference between the first and subsequent positions. Accuracy@1 and Accuracy@3 were considered as secondary binary characteristics of applied system-use scenarios.

Table 5. Global statistical comparison of local LLM configurations.

Endpoint	Test	Statistic	df	*p*-Value	Effect Size	Post Hoc
NDCG@5 (primary)	Friedman	119.8099	9	<0.001	Kendall's W = 0.2662	Wilcoxon + Holm
Accuracy@1 (secondary)	Cochran's Q	87.2991	9	<0.001	—	Descriptive support
Accuracy@3 (secondary)	Cochran's Q	146.8689	9	<0.001	—	Descriptive support

For the primary metric NDCG@5, the Friedman test revealed statistically significant differences between the models: $\chi^2_F(9) = 119.8099$. Kendall's coefficient of concordance was $W = 0.2662$. This value indicates a moderate magnitude of differences: the selected local LLM configuration affects the position of the reference degree programme in the output, although the differences between the closest leading models are not sharply pronounced.

Additional tests of the binary indicators also revealed heterogeneity between the models. For Accuracy@1, Cochran's $Q(9) = 87.2991$, for Accuracy@3, Cochran's $Q(9) = 146.8689$. Therefore, the models under study differ both in their ability to place the reference programme in the first position and in their ability to include it among the top three recommendations. These results are used as additional confirmation of differences in applied scenarios, whereas the primary statistical conclusion is based on NDCG@5.

The mean ranks of the models according to NDCG@5 are presented in Figure 5. A smaller mean rank corresponds to higher quality in placing the reference programme in the recommendation list.

The lowest mean rank was obtained by gpt-oss-20b-MXFP4 (3.90). Close values were obtained by gemma-4-E4B-it-Q8_0 (3.99) and Ministral-3-14B-Reasoning-2512-Q4_K_M (4.24). The highest, that is, worst, mean ranks were obtained by DeepSeek-R1-Distill-Qwen-7B-Q4_K_M (7.56) and LFM2-24B-A2B-Q4_K_M (7.73).

To refine the global result, post hoc comparisons were performed between the model with the highest observed NDCG@5 value, namely gpt-oss-20b-MXFP4, and the remaining configurations. The Wilcoxon signed-rank test with Holm correction over all 45 pairwise model comparisons was used. Table 6 presents the subset of comparisons involving gpt-oss-20b-MXFP4 as the model with the highest observed mean NDCG@5.

Table 6. Post hoc comparisons of models with gpt-oss-20b-MXFP4 according to NDCG@5.

Comparison Model	Δ Mean NDCG@5	Wilcoxon W	Holm-Adjusted *p*-Value	Conclusion
LFM2-24B-A2B-Q4_K_M	0.5713	4.5	<0.001	Significant
DeepSeek-R1-Distill-Qwen-7B-Q4_K_M	0.5303	16.0	<0.001	Significant
NVIDIA-Nemotron-3-Nano-4B-Q8_0	0.2439	18.5	0.0029	Significant
Qwen3.5-9B-Q4_K_M	0.3239	46.0	0.0044	Significant
Qwen3.5-9B-Q8_0	0.2900	30.0	0.0119	Significant
Qwen3.6-27B-UD-IQ3_XXS	0.0891	51.0	0.2481	Not confirmed
Qwen3.6-35B-A3B-UD-Q4_K_S	0.0908	93.0	0.5515	Not confirmed
Ministral-3-14B-Reasoning-2512-Q4_K_M	0.0114	55.0	1.0000	Not confirmed
gemma-4-E4B-it-Q8_0	0.0027	52.0	1.0000	Not confirmed

W, Wilcoxon signed-rank statistic.

A statistically significant advantage of gpt-oss-20b-MXFP4 according to NDCG@5 was established over LFM2-24B-A2B-Q4_K_M, DeepSeek-R1-Distill-Qwen-7B-Q4_K_M, NVIDIA-Nemotron-3-Nano-4B-Q8_0, Qwen3.5-9B-Q4_K_M, and Qwen3.5-9B-Q8_0.

The differences between gpt-oss-20b-MXFP4 and gemma-4-E4B-it-Q8_0, Ministral-3-14B-Reasoning-2512-Q4_K_M, Qwen3.6-27B-UD-IQ3_XXS, and Qwen3.6-35B-A3B-UD-Q4_K_S were not statistically confirmed after Holm correction. Therefore, the experimental

result should not be interpreted as evidence of the universal superiority of a single model. The gpt-oss-20b-MXFP4 model is the leader in terms of the observed values of Accuracy@1, NDCG@5, and MRR@5, whereas the final choice of configuration for the DSS should also take into account inference time and the intended recommendation mode.

4.5. Evaluation on the Additional Processed-Profile Robustness Set

The additional robustness experiment evaluated the methods on ten processed anonymised applicant profiles. The underlying profiles originated from actual applicant records, while their textual components were reformulated to reduce direct lexical correspondence with the programme catalogue and to introduce more heterogeneous descriptions of relevant experience. The actually selected field of admission remained unchanged and was used as the reference label.

The results are presented in Table 7. The TF–IDF baseline achieved Accuracy@1 = 0.6000, Accuracy@3 = 0.8000, NDCG@5 = 0.7131, and MRR@5 = 0.6833. In contrast, gemma-4-E4B-it-Q8_0 achieved the highest performance, with Accuracy@1 = 0.9000, Accuracy@3 = 1.0000, NDCG@5 = 0.9500, and MRR@5 = 0.9333.

Table 7. Ranking performance on the processed-profile robustness set.

Method	Accuracy@1	Accuracy@3	NDCG@5	MRR@5
TF–IDF cosine hybrid	0.6000	0.8000	0.7131	0.6833
gemma-4-E4B-it-Q8_0	**0.9000**	**1.0000**	**0.9500**	**0.9333**
gpt-oss-20b-MXFP4	0.6000	**1.0000**	0.8524	0.8000
Ministral-3-14B-Reasoning-2512-Q4_K_M	0.7000	0.8000	0.8448	0.7950

Higher values indicate better ranking quality.

All three LLM configurations returned extractable structured rankings for all ten profiles. Structured-output validity is not applicable to the deterministic TF–IDF baseline.

The gpt-oss-20b-MXFP4 model produced the same Accuracy@1 as the TF–IDF baseline, but placed the reference programme within the first three positions for all profiles and achieved higher NDCG@5 and MRR@5. Ministral-3-14B-Reasoning-2512-Q4_K_M also exceeded the TF–IDF baseline in Accuracy@1, NDCG@5, and MRR@5, although both methods achieved Accuracy@3 = 0.8000.

Compared with its performance on the main benchmark, the TF–IDF baseline decreased by 0.2800 in Accuracy@1, by 0.2000 in Accuracy@3, by 0.2426 in NDCG@5, and by 0.2567 in MRR@5. Thus, it retained 68.2% of its original Accuracy@1, 74.6% of its original NDCG@5, and 72.7% of its original MRR@5.

The profile-level analysis showed that the reduction was not limited to minor reordering of the first two candidates. For profiles C03 and C04, both representing Information Systems and Technologies through indirect descriptions of data integration, shared organisational services, synchronisation, and recovery, the primary programme was absent from the TF–IDF top-five ranking. All three LLMs placed the same primary programme within the top three for both profiles. For profile C07, which described information-security competence indirectly through access-control and event-investigation tasks, TF–IDF placed the primary programme third, whereas all three LLMs placed it first.

The model behaviour was not uniform across all cases. Ministral ranked the primary programme outside the top three for profiles C09 and C10, while gpt-oss placed the correct programme second in both cases. Gemma achieved the highest observed aggregate ranking performance on the processed-profile robustness set, placing the reference programme first for nine of the ten profiles and third for the remaining profile.

5. Discussion

This section interprets the empirical findings in relation to the five research questions. It first discusses differences among local LLM configurations, deployment and output-format trade-offs, and subject-area heterogeneity, and then examines the relationship between lexical retrieval and semantic robustness before considering practical implications and threats to validity.

5.1. Interpretation of the Comparative Analysis Results

The experiment provides initial evidence of the feasibility of using local large language models as semantic ranking components under controlled benchmark conditions. The main result is not proof of the practical applicability or absolute superiority of a particular model, but the identification of several candidate configurations for further DSS testing. At the same time, the main result of the study is not evidence of the absolute superiority of a single model, but the identification of a group of practically applicable local LLMs that differ in the nature of their usefulness for different DSS operating scenarios.

The gpt-oss-20b-MXFP4 model obtained the highest aggregated values for the primary recommendation and ranking quality indicators. Therefore, this model is the most justified candidate for a scenario in which the system is required to identify one primary field of study. However, the results of the post hoc analysis showed that, according to the primary metric, a statistically significant advantage of gpt-oss-20b-MXFP4 over its closest competitors, namely Ministral-3-14B-Reasoning-2512-Q4_K_M and gemma-4-E4B-it-Q8_0, was not established. It is therefore more appropriate to consider these three configurations as the top-performing group of models suitable for further testing.

The practical significance of this result lies in distinguishing between two modes of DSS application. For the primary recommendation mode, the most important characteristic is the model's ability to correctly determine the first position, which is reflected by Accuracy@1. For the preliminary consultation mode, in which an applicant is offered several substantively close options, Accuracy@3 becomes more important. In this respect, the gemma-4-E4B-it-Q8_0 model, which achieved the maximum Accuracy@3 value of 0.92, is of particular interest for forming an extended set of alternatives, even though its first-recommendation accuracy is slightly lower than that of gpt-oss-20b-MXFP4.

The Ministral-3-14B-Reasoning-2512-Q4_K_M model occupies an intermediate but practically significant position. Its ranking quality is close to that of the leading model, while its mean inference time is 5.38 s, compared with 8.78 s for gpt-oss-20b-MXFP4. This ratio makes Ministral-3-14B-Reasoning-2512-Q4_K_M a promising option for an interactive advisory service in which response generation latency is of substantial importance.

The results of the Qwen3.6 model family show why the quality of educational recommendation cannot be evaluated using only one indicator. These configurations quite often included the actually selected field of study in the top three positions, but they were inferior to the top-performing group in terms of first-recommendation accuracy. Consequently, a model may correctly identify the relevant subject area, but may not reliably determine the primary programme among closely related alternatives. For a DSS focused on issuing one priority field of study, this distinction is fundamental.

5.2. Influence of Model Size Class and Requirements for Local Deployment

The results obtained do not confirm the assumption that increasing the size class of a local LLM automatically improves the quality of field-of-study selection. The Qwen3.6-27B-UD-IQ3_XXS and Qwen3.6-35B-A3B-UD-Q4_K_S configurations were inferior to the more compact solutions in the top-performing group in terms of Accuracy@1, NDCG@5, and

MRR@5, while requiring substantially longer response times. By contrast, the 14B size-class model showed quality close to the maximum, with lower computational latency.

This result can be explained by the specificity of the task under study. Ranking degree programmes does not require the maximum breadth of generative capabilities, but rather the ability to follow instructions strictly: to identify significant profile features, match them with a limited list of closely related degree programmes, and return a correctly ordered list. Therefore, the final quality is determined not only by the number of model parameters, but also by instruction tuning, robustness to mixed numerical and textual input, and the ability to comply with output-format constraints.

The result of the LFM2-24B-A2B-Q4_K_M model is particularly illustrative. This model provided the minimum mean inference time, but at the same time obtained the lowest values for the ranking metrics. This means that response speed alone cannot be considered a sufficient basis for selecting the intelligent component of an educational DSS: a rapidly generated but substantively weak recommendation does not provide the required level of decision support.

Local deployment of LLMs is important not only in terms of speed or computational cost. An applicant digital profile may contain academic results, information on achievements, and portfolio materials, the processing of which requires organisational control over access and information transfer. Using a model within the infrastructure of an educational organisation reduces dependence on external cloud services and makes it possible to fix the model configuration and inference parameters used. At the same time, the mere fact of local execution does not guarantee the correctness of the decision: the model must undergo task-specific quality testing and should be used only as a support tool, not as an autonomous mechanism for selecting an educational trajectory.

5.3. Importance of Structured-Output Validity

For the DSS under consideration, an important result is the confirmation that it is necessary to evaluate not only the semantic quality of ranking, but also the suitability of the response for subsequent automated processing. Unlike a free-form conversational assistant, an applied system must obtain an ordered list of admissible programmes that can be stored, displayed in the user interface, and used when generating an explanation of the recommendation.

In the current analysis, a strict validity criterion was applied: a response was considered structurally valid when a non-empty ranked list of programme codes could be automatically extracted from the model output. The top-performing models, namely gpt-oss-20b-MXFP4, gemma-4-E4B-it-Q8_0, and Ministral-3-14B-Reasoning-2512-Q4_K_M, ensured structured-output validity. These three configurations combined high ranking quality with perfect structured-output validity. Several other configurations also achieved a validity value of 1.00, but did not provide the same combination of ranking quality and computational characteristics.

By contrast, some models may demonstrate acceptable inclusion of the relevant field in the upper part of the list, while at the same time violating the formal constraints of the output. For a practical system, this implies the need for repeated requests, external normalisation of the response, or a reserve model. Therefore, structured-output validity should be considered an independent characteristic of the operational reliability of a local LLM.

5.4. Lexical Retrieval and Semantic Robustness

The TF–IDF baseline produced two complementary findings. First, its high performance on the main benchmark demonstrates that applicant achievements and educational-

programme descriptions contained a substantial lexical signal. When the profile explicitly mentioned algorithms, communication protocols, information security, thermal processes, business processes, or other domain-specific terminology, direct content-based retrieval was sufficient to form an accurate programme ranking.

Second, the robustness experiment showed that this advantage decreased when relevant competence was expressed indirectly. The TF–IDF Accuracy@1 decreased from 0.8800 to 0.6000, while NDCG@5 decreased from 0.9557 to 0.7131. This result is consistent with the limitations of sparse lexical representations: TF–IDF assigns high relevance to shared tokens and character sequences but does not explicitly infer that, for example, unauthorised access to another user's document is evidence of information-security competence, or that reducing repeated data entry across organisational units is evidence of information-systems competence.

The selected local LLMs generally preserved ranking quality better than TF–IDF under these processed low-lexical-overlap conditions. Gemma substantially exceeded the TF–IDF baseline on all ranking metrics. The NDCG@5 values of gpt-oss and Ministral also remained close to their values on the main benchmark, whereas the corresponding TF–IDF metric showed a marked reduction. This pattern indicates that selected local LLM configurations can exploit contextual relations between actions, competencies, and programme objectives when direct terminology is reduced. Because the additional set was small, this finding should be treated as a robustness result for the examined profiles rather than as evidence of universal LLM superiority.

However, the robustness results do not establish universal LLM superiority. Gpt-oss achieved the same Accuracy@1 as TF–IDF, and Ministral produced weak rankings for the two Business Informatics profiles. The difference between gemma and the other LLMs also demonstrates that semantic robustness depends on the selected model and cannot be assumed solely from the use of an LLM.

From an applied perspective, the findings motivate the further development and controlled evaluation of a hybrid DSS architecture. TF–IDF or another inexpensive retrieval method can be used as a first-stage ranker for profiles containing explicit professional terminology. An LLM can then be applied to semantically ambiguous cases, to re-rank the retrieved candidates, explain the relationship between applicant experience and programme competencies, or request additional information from the applicant. Such an architecture could preserve the computational efficiency of classical retrieval while using LLM inference where contextual interpretation provides the greatest added value.

5.5. Subject-Area Heterogeneity of Results

The analysis by content groups of profiles indicates that the relative quality of the models may depend on the subject area. In the information technology group, the highest NDCG@5 value was demonstrated by gpt-oss-20b-MXFP4, whereas in the economics-and-information group, the highest result was obtained by gemma-4-E4B-it-Q8_0. In the engineering group, the three top-performing models showed identical values for the main metrics.

This result does not prove the existence of a stable subject-area specialisation of the models, since the sizes of individual subgroups are limited. However, the observed heterogeneity shows that evaluation using a single aggregated metric may conceal important features of model behaviour. In the future, this provides grounds for testing a routed DSS architecture, in which the choice of model depends on a pre-defined profile class or group of degree programmes.

5.6. Comparison with Existing Studies

Existing studies on the educational application of LLMs mainly consider automated assessment of written work, information extraction from textual descriptions of experience, feedback generation, and advisory chatbots. Studies on automated assessment confirm that LLMs are capable of extracting meaningful features from extended educational texts and relating them to assessment criteria [16,17]. Research on the analysis of unstructured descriptions of learning and professional experience demonstrates the possibility of transforming free text into substantively meaningful user characteristics [19].

The present study addresses a different task. The model must not assess a single text against a given scale or generate a response to an informational query, but rather compare a multi-component applicant profile with several competing programmes and determine their relative order. For this reason, the correct identification of a relevant field is insufficient: its position in the recommendation list is important, which justifies the use of Accuracy@1, Accuracy@3, NDCG@5, and MRR@5.

The difference between the present study and chatbots or RAG systems for admissions campaigns is also related to the objective function. Such systems are primarily designed to provide reliable reference information about admission, degree programmes, and admission rules [20,21]. The proposed DSS uses the programme catalogue not only as a data source, but also as a set of alternatives that must be personally ranked with respect to the applicant digital profile.

In relation to educational recommender systems, the present study extends earlier approaches towards the semantic analysis of semi-structured profiles. Unlike the recommendation of a learning resource or an individual course to an already enrolled user, the selection of a field of study is performed before admission, when there is no history of interaction with the information system, and the main sources of personalisation are the applicant's academic results, interests, and portfolio.

5.7. Practical Implications

The results obtained make it possible to propose the following interpretation for the design of a career guidance DSS. Among the evaluated local LLMs on the main benchmark, gpt-oss-20b-MXFP4 achieved the highest point estimates of Accuracy@1, NDCG@5, and MRR@5 and therefore represents the most justified LLM candidate for a primary-recommendation mode. Ministral-3-14B-Reasoning-2512-Q4_K_M provided a faster alternative with similar ranking quality, while gemma-4-E4B-it-Q8_0 was particularly effective for forming an extended list of relevant alternatives and achieved the highest performance on the processed-profile robustness set. All three configurations also achieved perfect structured-output validity.

At the same time, the TF–IDF baseline achieved higher ranking metrics than all local LLMs on the main benchmark. Consequently, the practical choice should not be framed as selecting either a classical or a generative method universally. TF–IDF is preferable for profiles with explicit and standardised terminology, whereas an LLM may provide additional value when applicant experience is described indirectly or ambiguously.

5.8. Limitations and Threats to Validity

Several threats to validity should be considered when interpreting the results.

First, the dataset used in this study is limited in size and institutional scope. It includes 50 applicant profiles and 10 degree programmes from one university. This design is sufficient for a controlled pilot benchmark of local LLM configurations under identical input conditions, but it is not sufficient for broad generalisation to other universities, admission systems, or educational domains.

Second, the distribution of profiles across subject areas is imbalanced. The predominance of IT-oriented profiles reflects the experimental context and the available programme catalogue, but it may affect the aggregated model rankings. Therefore, subgroup results should be interpreted descriptively and require confirmation on larger and more balanced datasets.

Third, the actual field of admission is used as an observed reference label, not as an objective ground truth of educational suitability. The applicant's final programme choice may be influenced by quotas, admission strategy, geographic constraints, tuition conditions, family preferences, and other external factors. The present evaluation therefore measures agreement with the observed reference programme rather than long-term educational success or expert-confirmed suitability.

Fourth, the current formulation uses binary relevance: one programme is treated as the reference label for each profile. In real career guidance, several programmes may be simultaneously appropriate. This limits the interpretation of Accuracy@1, Accuracy@3, NDCG@5, and MRR@5. A more complete evaluation should use graded relevance labels assigned by academic advisors or career guidance specialists.

Fifth, the generated explanations were not independently evaluated by experts. Structured-output validity in this study means that the ranked list could be automatically extracted from the model response. It does not imply that the textual explanation is pedagogically correct, complete, or useful for the applicant.

Sixth, all models were evaluated in a unified zero-shot prompt setting. This improves comparability between models but does not represent the best achievable performance of each configuration. Model-specific prompt engineering, constrained decoding, few-shot examples, or curriculum-augmented retrieval may change the relative ranking of models.

Seventh, the study includes a targeted robustness test based on processed anonymised profiles of ten applicants. The texts were reformulated to reduce direct lexical correspondence, but the underlying applicant information and observed admission programmes were preserved. The test does not cover naturally occurring incomplete profiles, extraction errors, missing portfolio elements, or other forms of real-world noise.

Most applicants entering higher education immediately after completing secondary school are typically approximately 17–18 years old. However, exact age and sex or gender were not recorded or retained in the anonymised dataset used in this study. Consequently, the present analysis does not include demographic subgroup comparisons or a formal fairness audit. This limits the possibility of assessing whether ranking quality differs across demographic groups. Future studies should include appropriately governed demographic variables and evaluate model performance, error patterns, and recommendation consistency separately for relevant applicant subgroups.

Finally, the study does not evaluate longitudinal educational outcomes such as student satisfaction, academic performance, persistence in the selected programme, or subsequent career success. The results should therefore be interpreted as evidence of local LLM feasibility for a constrained ranking benchmark, not as full validation of an autonomous educational recommendation system.

The additional robustness set was based on processed anonymised applicant profiles and was intended to examine ranking under reduced direct lexical overlap. Because it contains only ten profiles, five represented reference programmes, and two profiles per programme, its results should be treated as a targeted robustness analysis rather than as a population-level estimate of performance.

Future work should extend this analysis to larger independently processed samples and formalise the procedure used to reduce lexical overlap. A larger dataset would also enable statistical comparison of the relative performance degradation of sparse retrieval,

embedding-based retrieval, and LLM re-ranking approaches. Future studies should include multi-institutional and cross-country samples, controlled comparison with commercial cloud-based models, qualitative interviews with applicants and career counsellors, and dedicated explainability and demographic-fairness analyses.

6. Conclusions

This study investigated the use of locally executable large language models for ranking degree programmes on the basis of anonymised applicant digital profiles within a career guidance decision-support system. The task was formulated as the ordering of a fixed set of competing degree programmes according to examination results, professional interests, and portfolio achievements. The comparison used Accuracy@1, Accuracy@3, NDCG@5, and MRR@5, while structured-output validity and inference time were analysed as operational characteristics of the local LLMs.

The main experiment included ten local LLM configurations, 50 anonymised applicant profiles, and a fixed catalogue of ten fields of study. Among the evaluated local LLMs, gpt-oss-20b-MXFP4 achieved the highest point estimates of Accuracy@1, NDCG@5, and MRR@5. Ministral-3-14B-Reasoning-2512-Q4_K_M demonstrated similar ranking quality with a lower mean inference time, while gemma-4-E4B-it-Q8_0 achieved the highest Accuracy@3 value and was effective in forming an extended list of degree-programme alternatives. Statistical testing confirmed differences between the evaluated configurations, although the advantage of gpt-oss over the other leading LLMs was not statistically confirmed after correction for multiple comparisons.

The addition of a hybrid TF–IDF cosine baseline substantially changed the interpretation of the main benchmark. On the 50-profile dataset, TF–IDF achieved Accuracy@1 = 0.8800, Accuracy@3 = 1.0000, NDCG@5 = 0.9557, and MRR@5 = 0.9400, exceeding all evaluated local LLM configurations on the ranking metrics. This result demonstrates that classical lexical retrieval remains a highly effective and computationally inexpensive method when applicant profiles and programme descriptions contain explicit and consistent domain terminology.

The additional robustness experiment used processed anonymised data from ten applicants. The profile descriptions were reformulated to reduce direct lexical correspondence with the programme catalogue and to represent relevant experience in more heterogeneous and indirect language, while the actually selected fields of admission were preserved as reference labels. Under these conditions, TF–IDF Accuracy@1 decreased to 0.6000 and NDCG@5 to 0.7131. Gemma-4-E4B-it-Q8_0 achieved Accuracy@1 = 0.9000 and NDCG@5 = 0.9500, while gpt-oss-20b-MXFP4 and Ministral-3-14B-Reasoning-2512-Q4_K_M also exceeded TF–IDF in the position-sensitive ranking metrics.

Taken together, the two experiments show that the relative effectiveness of the methods depends on the representation of applicant information. TF–IDF is particularly effective when the profile explicitly reproduces the terminology associated with the programme catalogue. Selected local LLMs provide additional value when relevant competencies must be inferred from indirect descriptions, interdisciplinary achievements, or contextual evidence. The results therefore support the further development of hybrid systems in which classical retrieval is used for efficient candidate generation and local LLMs are used for semantic re-ranking, explanation generation, and interactive clarification.

The research questions can therefore be answered as follows:

RQ1. The local LLM configurations differed substantially in ranking quality under identical input conditions, with statistically significant global differences in NDCG@5 and the binary top-k indicators.

RQ2. Among the LLMs, gpt-oss provided the strongest observed primary recommendation, gemma was most effective for an extended top-three set and for the processed robustness set, and Ministral offered the most favourable quality–latency compromise.

RQ3. Ranking quality, structured-output validity, model size, and inference time were not monotonically related; deployment therefore requires a multi-criteria choice rather than selection by parameter count alone.

RQ4. Several local LLM configurations are suitable candidates for further DSS testing, but the evidence supports an advisory component with mandatory human oversight rather than autonomous programme selection.

RQ5. TF–IDF was superior on profiles with explicit terminology, whereas selected LLMs preserved more quality when relevant competencies were expressed indirectly; this result motivates a hybrid retrieval and semantic re-ranking architecture.

From a practical deployment perspective, the results suggest a two-stage advisory workflow. A lightweight TF–IDF component can first generate candidate programmes when the applicant profile contains explicit and standardised terminology. When lexical similarity scores are close, the profile is incomplete, or competencies are described indirectly, a local LLM can be used to re-rank the leading candidates and generate a concise explanation of their correspondence to the applicant's experience. The resulting top-three list should then be reviewed by a career guidance specialist together with the applicant. Such an architecture can reduce computational load, preserve local control over applicant data, and reserve generative inference for cases in which contextual interpretation is most useful. This deployment pattern remains a proposal derived from the comparative results and requires prospective validation before operational use.

The findings should nevertheless be interpreted within the limits of the study. The main benchmark contains 50 profiles from one educational organisation and uses the actually selected field of admission as a single observed reference label. The additional robustness set contains only ten processed profiles. The study does not include graded expert relevance for multiple programmes, independent assessment of generated explanations, optimisation of prompts for individual models, or longitudinal evaluation of educational outcomes. Further research should expand the institutional and subject coverage of the dataset, formalise the processing of low-lexical-overlap profiles, evaluate explanations with career guidance specialists, and implement and test the proposed hybrid ranking architecture.

Author Contributions: Conceptualization, A.O. and D.D.; methodology, A.O.; software, Y.N.; validation, A.S. and M.S.; formal analysis, Y.N.; investigation, A.S.; resources, A.S.; data curation, M.S.; writing—original draft preparation, A.S.; writing—review and editing, A.O. and D.D.; visualization, M.S.; supervision, D.D.; project administration, A.O.; funding acquisition, A.O. All authors have read and agreed to the published version of the manuscript.

Funding: The article was carried out with the financial support of the Ministry of Science and Higher Education of the Russian Federation within the framework of the project No. 124032800018-5.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Urdaneta-Ponte, M.C.; Mendez-Zorrilla, A.; Oleagordia-Ruiz, I. Recommendation Systems for Education: Systematic Review. *Electronics* **2021**, *10*, 1611. [CrossRef]
2. Grassini, S. Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. *Educ. Sci.* **2023**, *13*, 692. [CrossRef]
3. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic Review of Research on Artificial Intelligence Applications in Higher Education—Where Are the Educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 39. [CrossRef]
4. Lo, C.K. What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Educ. Sci.* **2023**, *13*, 410. [CrossRef]
5. Rahman, M.M.; Watanobe, Y. ChatGPT for Education and Research: Opportunities, Threats, and Strategies. *Appl. Sci.* **2023**, *13*, 5783. [CrossRef]
6. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 9459–9474.
7. Da Ros, F.; Soprano, M.; Di Gaspero, L.; Roitero, K. Large Language Models for Combinatorial Optimization: A Systematic Review. *ACM Comput. Surv.* **2026**, *58*, 1–53. [CrossRef]
8. Darwin; Rusdin, D.; Mukminatien, N.; Suryati, N.; Laksmi, E.D.; Marzuki. Critical Thinking in the AI Era: An Exploration of EFL Students' Perceptions, Benefits, and Limitations. *Cogent Educ.* **2024**, *11*, 2290342. [CrossRef]
9. Obukhov, A.; Dedov, D.; Krasnyanskiy, M.; Popov, A. A Mathematical Model of Organizing the Developmental Instruction in the System of Professional Education. *Tech. Gaz.* **2020**, *27*, 480–488. [CrossRef]
10. Drachsler, H.; Hummel, H.G.K.; Koper, R. Personal Recommender Systems for Learners in Lifelong Learning Networks: The Requirements, Techniques and Model. *Int. J. Learn. Technol.* **2008**, *3*, 404–423. [CrossRef]
11. Manouselis, N.; Drachsler, H.; Verbert, K.; Duval, E. *Recommender Systems for Learning*; Springer: New York, NY, USA, 2013.
12. Askarbekuly, N.; Luković, I. Learning Outcomes, Assessment, and Evaluation in Educational Recommender Systems: A Systematic Review. *arXiv* **2024**, arXiv:2407.09500.
13. Verbert, K.; Manouselis, N.; Ochoa, X.; Wolpers, M.; Drachsler, H.; Bosnić, I.; Duval, E. Context-Aware Recommender Systems for Learning: A Survey and Future Challenges. *IEEE Trans. Learn. Technol.* **2012**, *5*, 318–335. [CrossRef]
14. Tarus, J.K.; Niu, Z.; Mustafa, G. Knowledge-Based Recommendation: A Review of Ontology-Based Recommender Systems for E-Learning. *Artif. Intell. Rev.* **2018**, *50*, 21–48. [CrossRef]
15. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
16. Pack, A.; Barrett, A.; Escalante, J. Large Language Models and Automated Essay Scoring of English Language Learner Writing: Insights into Validity and Reliability. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100234. [CrossRef]
17. Li, W.; Liu, H. Applying Large Language Models for Automated Essay Scoring for Non-Native Japanese. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 723. [CrossRef]
18. Lira, B.; Gardner, M.; Quirk, A.; Stone, C.; Rao, A.; Ungar, L.; Hutt, S.; Hickman, L.; D'mEllo, S.K.; Duckworth, A.L. Using Artificial Intelligence to Assess Personal Qualities in College Admissions. *Sci. Adv.* **2023**, *9*, eadg9405. [CrossRef] [PubMed]
19. Kondo, T.; Nishigori, H. AI's Accuracy in Extracting Learning Experiences from Clinical Practice Logs: Observational Study. *JMIR Med. Educ.* **2025**, *11*, e68697. [CrossRef] [PubMed]
20. Nguyen, T.T.K.; Hoang, V.N.; Tinh, P.N.; Van, N.H.V. Efficient Chatbot for University Admission Consultation Using Large Language Models. *J. Sci. Technol.—Univ. Danang* **2025**, *23*, 80–87. [CrossRef]
21. Nguyen, L.S.T.; Quan, T.T. URAG: Implementing a Unified Hybrid RAG for Precise Answers in University Admission Chatbots—A Case Study at HCMUT. In Proceedings of the International Symposium on Information and Communication Technology, Hanoi, Vietnam, 13–15 December 2024; pp. 82–93.
22. Ma, B.; Khan, M.A.Z.; Yang, T.; Polyzou, A.; Konomi, S. How Good Are Large Language Models for Course Recommendation in MOOCs? *arXiv* **2025**, arXiv:2504.08208.
23. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288. [CrossRef]
24. Frantar, E.; Ashkboos, S.; Hoefler, T.; Alistarh, D. GPTQ: Accurate Post-Training Quantization for Generative Pre-Trained Transformers. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
25. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–45. [CrossRef]
26. Farrokhnia, M.; Banihashem, S.K.; Noroozi, O.; Wals, A. A SWOT Analysis of ChatGPT: Implications for Educational Practice and Research. *Innov. Educ. Teach. Int.* **2024**, *61*, 460–474. [CrossRef]

27. Lent, R.W.; Brown, S.D.; Hackett, G. Toward a Unifying Social Cognitive Theory of Career and Academic Interest, Choice, and Performance. *J. Vocat. Behav.* **1994**, *45*, 79–122. [CrossRef]
28. Wigfield, A.; Eccles, J.S. Expectancy–Value Theory of Achievement Motivation. *Contemp. Educ. Psychol.* **2000**, *25*, 68–81. [CrossRef] [PubMed]
29. Chen, L.-C.; Chang, K.-H.; Hung, J.-F. WASPAS-Based Natural Language Processing Method for Handling Content Words Extraction and Ranking Issues: An Example of SDGs Corpus. *Information* **2025**, *16*, 198. [CrossRef]
30. Stathopoulos, E.A.; Karageorgiadis, A.I.; Kokkalas, A.; Diplaris, S.; Vrochidis, S.; Kompatsiaris, I. A Query Expansion Benchmark on Social Media Information Retrieval: Which Methodology Performs Best and Aligns with Semantics? *Computers* **2023**, *12*, 119. [CrossRef]
31. Nam, G.; Yang, G. LLMLoc: A Structure-Aware Retrieval System for Zero-Shot Bug Localization. *Electronics* **2025**, *14*, 4343. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.