Large-Scale Mobile Application Classification Using Large Language Models: A Design Science Study in SportsHCI

Juhani J. M. Karjalainen[1], Marko I. M. E. Moilanen[1,*]

[1]INTERACT Research Unit, University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

The rapid growth of mobile sports and health & fitness applications has created challenges for large-scale collection and classification, which have traditionally relied on labor-intensive manual processes and feature-engineered or supervised machine learning approaches. This study addresses this challenge by developing a scalable method for constructing structured datasets of mobile applications. Following a design science research approach, we develop a large language model-based artifact that integrates automated data collection with LLM-based classification. The artifact achieves high agreement with human classification while significantly reducing processing time, demonstrating that LLM-based approaches enable scalable and flexible classification without predefined features or labeled training data. The study contributes methodologically by demonstrating the feasibility of LLM-based artifacts for large-scale classification, and empirically by enabling systematic analysis of application ecosystems in SportsHCI. To support transparency and reuse, the dataset (~47,000 applications) and artifact implementation are available for public access.

Keywords

Artificial Intelligence, Large Language Models, Data Classification, Digital Ecosystems, Sports Applications, Design Science Research, SportsHCI

1. Introduction

Mobile Sports and Health & Fitness (SHF) applications form a large and increasingly diverse domain, with over 100,000 applications available across major application stores. For example, the Google Play Store alone contains over 80,000 applications in the Health & Fitness category and nearly 40,000 in the Sports category [1]. While this scale creates opportunities for Sports-related Human–Computer Interaction (SportsHCI) research, it also introduces challenges for systematically collecting and structuring large-scale application datasets, particularly when relying on manual approaches.

SHF applications can be broadly classified into two main types: tracking applications that measure users' physical activity, and instructional applications that support individual training [2]. Modern smartphones are also equipped with built-in sensors, which enable applications to capture physiological and biometrical data without additional equipment [3]. In the health context, applications are used as tools for behaviour change and for tracking physiological and lifestyle data [4, 5]. From a stakeholder perspective, earlier work distinguishes between competitive and non-competitive actors [6], while more recent research identifies broader stakeholder groups, including competitors, support staff, supporters, and governing entities [7].

Within this diverse field, the classification of applications remains challenging, particularly due to limitations in app store categorization, where applications are assigned static categories that do not reflect evolving application domains [8]. Existing approaches include multiple taxonomy-based methods for classifying applications [9, 10, 11], but a unified classification framework for e.g. health-related applications is lacking, and current approaches remain fragmented [12].

TKTP 2026: Annual Doctoral Symposium of Computer Science, 8.—9.6.2026 Turku, Finland

*Corresponding author.

juhanikarjalainen54321@gmail.com (J. J. M. Karjalainen); marko.i.moilanen@oulu.fi (M. I. M. E. Moilanen)

https://github.com/jukarjalainen/AICCSA---AI-tool-for-Collecting-and-Classifying-Sports-Applications (J. J. M. Karjalainen)

0009-0006-9209-253X (M. I. M. E. Moilanen)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-02

CEUR-WS.org/Vol-4309/paper12.pdf

Previous research has proposed several approaches for classifying applications. These include manual screening and filtering of applications [13, 11], machine learning approaches that utilize application description texts for automatic categorization [10], and methods based on word embeddings for representing application descriptions and classifying them into categories [8].

The SportsHCI domain provides a particularly suitable context for this study, as it encompasses diverse functionalities, stakeholder groups, and usage contexts, making it a complex case for large-scale application classification. Existing approaches either rely on predefined taxonomies or require labeled datasets, limiting adaptability in rapidly evolving application ecosystems. This study addresses this gap by leveraging Large Language Models (LLMs) as semantic classifiers that can operate without task-specific training data.

LLMs are capable of producing diverse forms of content, including text [14], and provide capabilities such as summarization and writing assistance, while their usage costs have decreased significantly [15, 16]. Structured instructions (prompting) are used to guide how LLMs interpret tasks and generate outputs [17], and enables task-specific use without additional training data [18]. However, the effectiveness of prompting strategies varies. For example, role prompting has been shown to produce inconsistent effects [19], LLMs are sensitive to irrelevant prompt details [20], and domain priming has been shown to provide more consistent improvements compared to persona-based prompting [21]. In classification tasks, structured outputs (e.g., JSON) improve performance, while reducing reasoning performance [22].

This study addresses the following research problem: How can an LLM-based approach support scalable and accurate classification of sports and health & fitness applications? To address this problem, the study adopts a Design Science Research (DSR) approach [23] to develop and evaluate an artifact that integrates automated data collection with LLM-based classification. The artifact supports multi-dimensional categorization of applications by sport type, purpose, and primary user group.

After data cleaning, the artifact was demonstrated on a dataset of approximately 50,000 applications collected from major app stores and evaluated by comparing Artificial Intelligence (AI)-generated classifications with those produced by a human evaluator. The results indicate accurate classification while reducing the time required for processing large datasets.

This study contributes to the literature on mobile application classification by demonstrating how LLM-based approaches can address limitations of prior methods, particularly in terms of scalability and reduced reliance on labeled data. In addition, it provides a structured dataset and a scalable methodological approach for analyzing the SportsHCI application ecosystem.

2. Background

This section introduces the sports and health & fitness mobile application field, together with application classification approaches and known challenges, after which, automated and LLM-based classification is considered.

2.1. Applications in the Sports and Health & Fitness Categories

Mobile application ecosystems have expanded rapidly in recent decades. Within this ecosystem, SHF applications represent a diverse and evolving category of digital tools designed to support physical activity, training, and overall well-being, with over 100,000 applications available [1].

Research on sports applications can be approached from multiple perspectives. SportsHCI focuses on the human aspects of interaction [24, 25], while information systems research examines organizational and technological dimensions [7, 26]. This study bridges these perspectives by considering user groups together with application sport type and purpose. Here, application purpose refers to its main functionalities rather than to existential or meaning-related aspects of use, which have been explored in prior work [27, 28].

SHF applications can be classified into two main types: those that focus on tracking and analysing users' physical activity, and those that provide instructional content to support individual training [2].

Common features observed in such applications include reminders, notifications, voice instructions, workout diaries, statistical summaries, and social sharing functionalities, with reminders being the most common feature [2]. Modern smartphones are also equipped with built-in sensors, such as accelerometers, gyroscopes, cameras, and microphones, which enable the capture of physiological and biometrical data without additional equipment - enabling applications to generate detailed measurements directly from the device. [3]

Within the broader domain of mobile health (mHealth), these applications are used as tools for behaviour change [4]. Health and fitness applications monitor and shape users' health-related behaviours and enable the tracking of physiological and lifestyle data, such as physical activity, sleep patterns, calorie intake, heart rate, and blood pressure [4, 5]. A distinction is made between disease-focused applications and wellness applications, where the latter support the maintenance of healthy habits and overall well-being [29].

Prior research has identified multiple stakeholder groups involved in sports digitalization, including competitors, support staff, supporters, and governing entities [7]. These stakeholder groups differ in terms of their proximity to the sporting activity, their influence on competition, and their group size [7]. Earlier work distinguishes between competitive actors, such as athletes and teams, and non-competitive actors, such as referees and officials who facilitate sporting activities [6].

2.2. Mobile Application Classification: Approaches and Challenges

Application store categorizations are static and do not reflect evolving application domains [8]. The literature lacks a unified definition or common classification framework for health-related applications, and existing approaches are fragmented and focus on narrow aspects [12].

Several studies have proposed taxonomy-based approaches for classifying mobile applications. A study by [9] developed taxonomies based on application goals and features, identifying functional groups and higher-level categories such as education, diet and nutrition, physical activity, and health. Similarly, [10] introduced a more fine-grained categorization by defining multiple sub-categories beyond those provided by application stores.

In the context of sports applications, [11] proposed a three-dimensional categorization framework based on sport type, application functionality, and analytics. The study further categorizes applications according to analytics types and identifies multiple functionality categories.

The literature includes multiple taxonomy-based approaches for classifying applications [9, 10, 11]. However, a unified classification framework for health-related applications is lacking, and existing approaches are fragmented [12].

2.3. Automated and LLM-Based Classification

Application data is commonly collected from app stores using keyword-based search and manually screening the results [11, 13]. In addition, prior research has explored automated approaches for classifying mobile applications [10, 8]. For example, a framework utilizes application description texts as input and applies machine learning algorithms for classification, where linear Support Vector classification performs best [10]. Similarly, previous research has proposed an approach based on word embeddings for representing application descriptions and classifying them into categories [8].

Generative artificial intelligence and LLMs are capable of producing diverse forms of content, including text, images, and code [14]. Research on LLMs has grown rapidly in recent years, alongside their integration into various research domains and workflows [30]. These models provide capabilities such as summarization, writing assistance, translation, and sentiment analysis, and have become increasingly accessible due to reductions in usage costs [15, 16]. Modern LLMs are typically based on transformer architectures [31], which use self-attention mechanisms [32].

Prompting is used to guide how LLMs interpret tasks and generate outputs through structured instructions [17]. Prompt engineering enables users to adapt LLMs to specific tasks without additional training data and can be viewed as a form of natural language programming [18]. Various prompting

techniques have been proposed, including zero-shot and few-shot approaches [33]. The effectiveness of prompting strategies varies. For example, role prompting has been shown to produce inconsistent effects, sometimes improving performance but also reducing accuracy [19]. This variability is related to the sensitivity of LLMs to irrelevant prompt details, which can introduce noise and affect performance [20]. Domain priming has been shown to provide more consistent improvements compared to persona-based prompting [21].

In addition to prompt design, the structure of model outputs affects performance. A previous research shows that stricter format constraints, such as structured outputs (e.g., JSON), improve performance in classification tasks, while reducing reasoning performance. Format constraints affect LLM performance differently across tasks. [22]

Based on the identified approaches and limitations, the following section presents a research process to design and evaluate a LLM-based artifact for application classification.

3. Research Method

This study follows the Design Science Research (DSR) methodology [34] and the Design Science Research Process (DSRP) model [23]. DSR is well-suited for addressing practical problems through the design and evaluation of artifacts. The following sections present the six steps of the DSRP model as applied in this study.

3.1. Problem Identification

The large and rapidly growing number of SHF applications makes systematic collection and classification challenging while e.g. the Google Play Store alone includes over 100,000 applications in these categories [1]. Prior work on application classification has evolved from feature-based methods [10] to semantic approaches [8], but these still rely on predefined features and labeled datasets. At the same time, the field lacks unified classification frameworks [12], and manual approaches remain difficult to scale [13, 11].

These challenges are particularly evident in SportsHCI, where diversity in purposes, user groups, and sport types complicates classification. To address this, this study investigates whether a LLM-based artifact can support large-scale collection and classification of sports applications.

3.2. Definition of Objectives

Based on the identified problem, the objective is to design an artifact capable of automatically collecting and structuring large-scale datasets of SHF applications. In addition, the artifact needed to accurately classify applications into predefined categories, including user groups, purposes, and sport types, while filtering out irrelevant entries.

The goal was to approximate the quality of manual classification while significantly improving scalability and efficiency, thereby enabling large-scale analysis of the application ecosystem.

3.3. Design and Development

The artifact was implemented as a software tool integrating data collection and LLM-based classification. Data were gathered from mobile application stores using web scraping libraries (google-play-scraper and app-store-scraper). The classification component utilized a large language model (GPT-based API) to categorize applications and filter irrelevant entries.

The system was developed using JavaScript for data collection and Python for the classification pipeline. AI-assisted development tools were used to support implementation efficiency.

3.4. Demonstration

The artifact was demonstrated by applying it to a dataset of application descriptions. The system processed each entry, assigning categories for user groups, purposes, and sport types, while filtering out irrelevant applications. Demonstration is further described in Section 4.

3.5. Evaluation

The artifact was evaluated by comparing its classifications with those produced by a human evaluator on the same dataset. Classification performance was assessed using exact agreement across all classification parameters, as well as agreement at the level of individual parameters.

3.6. Communication

The results of this study, including the identified problem, artifact design, and evaluation, are communicated through this paper and a related conference submission.

4. Artifact Design

This section presents an overview of the artifact, how it works, together with how data collection and LLM-based classification are designed.

4.1. Overview of the Artifact

The developed artifact consists of two main modules: a data collection module and a LLM-based classification module. The data collection module gathers application data from mobile application stores at scale, while the classification module assigns structured categories and filters out irrelevant applications.

The artifact was designed to address the challenge of processing large-scale datasets, where manual classification of tens of thousands of applications would be impractical. Together, these modules form a pipeline that enables the automated construction of a structured dataset of SHF applications.

The artifact design evolved iteratively through multiple design and evaluation cycles (Table 1), where insights from preliminary results and practical constraints informed subsequent refinements of both data collection and classification components.

4.2. Data Collection Design

The data collection module was designed to enable large-scale and diverse retrieval of chosen application fields from mobile application stores. Data were collected using two complementary strategies: (1) retrieving applications from predefined store lists (e.g., top charts) and (2) executing keyword-based searches across a wide range of topic-related terms.

To increase coverage and reduce regional bias, data collection was conducted across multiple countries. Due to practical constraints related to API limits, the final design focused on 30 countries and a reduced set of search queries, balancing dataset diversity and collection efficiency.

The combined use of list-based and search-based approaches resulted in overlapping results, which were addressed through deduplication based on unique application identifiers. In addition, data quality issues in application descriptions were handled by re-collecting missing or corrupted fields.

This process produced a large-scale dataset of over 53,000 applications, forming the basis for subsequent classification.

Table 1
Iterative Development Phases of the Artifact

Phase	Issue Identified	Iterative Improvements	Outcome
Data Collection	Manual search slow and limited	Switched to automated API-based scraping	Enabled large-scale data collection
Collection Scope	Scraping too slow (118 countries, 424 queries)	Reduced to 30 countries and 190 queries	Completed in ~50 hours
Data Processing	Noisy data and corrupted descriptions	Removed invalid entries and repaired descriptions	Classified 46,862 applications in 3.5 hours
Prompt Iteration	Low initial classification accuracy	Iterative refinement with human-labeled samples	133/144 matches with human classification

4.3. LLM-Based Classification Design

The classification module is designed to automatically assign structured categories based on applications' descriptions. Given the scale of the dataset, a LLM-based approach was selected to enable flexible and scalable classification without reliance on predefined feature engineering or labeled training data.

The classification scheme consists of three main dimensions: (1) user groups, (2) application purpose, and (3) sport type. User group classification was based on stakeholder roles (e.g., athletes, supporters, support staff, and governing entities) as suggested by [7], while purpose categories were derived from prior work by [11] and extended to better capture the diversity of sports and health applications. Sport type labels were assigned using a predefined set of tags, with flexibility for handling multi-sport and general health contexts.

The classification process was guided by a structured prompting approach, where the model was instructed to produce standardized outputs and apply predefined rules for category assignment and exclusion of irrelevant applications. To ensure consistency and scalability, classification requests were processed in batches, enabling efficient handling of the large dataset. This design enabled automated, large-scale classification of applications while maintaining a level of consistency.

The development of the artifact followed an iterative process, where identified limitations were addressed through successive refinements. Table 1 summarizes the key development phases and their outcomes.

5. Evaluation

This section presents the evaluation of the proposed artifact. The evaluation examines the accuracy of LLM-based classification in comparison to human judgment, as well as the scalability and efficiency of the approach in large-scale application classification.

5.1. Evaluation Setup

The artifact was evaluated by comparing its classification outputs with those produced by a human evaluator on the same dataset. The evaluation focused on assessing how closely the LLM-based classification aligns with human judgment. A single expert evaluator with familiarity in sports applications and classification tasks was used to establish a baseline comparison for the artifact. While multi-rater evaluation would strengthen reliability, the goal of this evaluation was to assess alignment with informed human judgment rather than establish inter-rater agreement.

A subset of the collected applications was used for evaluation, resulting in a dataset of 145 applications after removing duplicates and unavailable entries. One application could not be classified by the AI tool, leaving 144 applications for evaluation. Each application was classified across seven parameters: relevance, purpose, sport type, and four user group indicators (athlete, supporter, support staff, and governing entity).

Table 2
Classification accuracy across parameters.

Dimension	Correct	Incorrect	Correct (%)
Relevance	141	3	97.92
Purpose	131	6	95.62
Sport Type	136	1	99.27
Athlete	135	2	98.54
Supporter	135	2	98.54
Support Staff	134	3	97.81
Governing Entity	137	0	100.00

A classification was considered correct only when all seven parameters exactly matched the corresponding human classification. This strict criterion was chosen to ensure that the artifact performs reliably across all classification dimensions simultaneously. In addition, accuracy was analyzed at the level of individual parameters to provide a more detailed view of classification performance.

5.2. Classification Accuracy

The classification performance of the artifact was evaluated by comparing its outputs with those of the human evaluator. Using the strict matching criterion, the artifact achieved exact agreement in 133 out of 144 applications (92.4%), indicating a high level of alignment with human judgment.

In addition to exact matches, classification accuracy was analyzed across individual parameters (Table 2). The results show consistently high agreement across all dimensions, with accuracy ranging from 95% to 100%. The highest accuracy was observed in sport type and governing entity classification, while slightly lower agreement was found in the assignment of application purposes.

These results demonstrate that the artifact can reliably classify applications across multiple dimensions, supporting complex categorization tasks beyond simple single-label classification.

5.3. Scalability and Efficiency

The data collection process initially retrieved over 53,000 applications from mobile application stores, demonstrating the artifact's capability for large-scale classification. After removing duplicates and unavailable entries, the final dataset consisted of 46,862 applications, demonstrating the scalability of the approach.

The classification phase was completed using batch processing, enabling the full dataset to be processed within 3.5 hours. In contrast, manual classification of a dataset of this size was estimated to require several months of work.

In terms of resource efficiency, the classification process required 868 API calls and processed approximately 19 million tokens, with a total cost of 7.4€. These results indicate that the artifact enables cost-effective large-scale classification without the need for manual labeling or predefined feature engineering.

5.4. Error Analysis

The majority of classification errors occurred in the assignment of application purposes, where distinctions between closely related categories (e.g. team management vs. league management, betting vs. fantasy sports) led to inconsistencies. These errors reflect the inherent ambiguity and overlap between certain application functionalities.

Additional errors were observed in cases where application descriptions were limited or unclear, reducing the amount of information available for accurate classification. Minor discrepancies also appeared in user group assignments, particularly in edge cases where applications targeted multiple stakeholder groups.

Overall, the error patterns suggest that while the artifact performs reliably in most cases, classification challenges remain in semantically overlapping categories and contexts with low-information.

These results suggest that LLM-based classification can achieve high agreement with human judgment while addressing key limitations of prior approaches, particularly in terms of scalability and reduced reliance on predefined features or labeled datasets.

6. Discussion

This study addressed the research and design problem of how an LLM-based approach can support scalable and accurate classification of sports and health & fitness applications. To address this, a design science research approach was adopted to design and evaluate an artifact that integrates automated data collection with LLM-based classification. The artifact enables multi-dimensional categorization of applications based on sport type, purpose, and user groups.

The results suggest that the LLM-based artifact achieves performance comparable to that of the human evaluator, with an overall accuracy of 92.4% and 95–100% accuracy across individual classification dimensions. These findings indicate strong agreement with human judgment in a complex, multi-label and multi-dimensional classification task.

In addition to accuracy, the artifact demonstrates strong efficiency and scalability. The classification of 46,862 applications was completed in approximately 3.5 hours at a total cost below 8€, supporting recent claims regarding reduced usage costs of LLM-based solutions [15, 16]. Although the data collection phase required approximately 50 hours due to API rate limitations, this still represents a substantial improvement compared to months of manual work.

The artifact also exhibited classification errors, primarily in the assignment of application purposes. These errors were associated with overlapping categories, ambiguous application descriptions, and edge cases involving multiple user groups. Importantly, these errors were not random but systematic and explainable, suggesting that they can be addressed in future iterations. At the same time, these challenges are not solely technical but also reflect inherent limitations in the domain, where application stores often contain overlapping semantics and incomplete descriptions. This highlights a broader design challenge beyond the artifact itself.

A key advantage of the LLM-based approach is its adaptability. Unlike rigid classification systems, the artifact's categories and prompts can be modified to reflect evolving classification needs. This is particularly relevant if standardized classification systems emerge in the future, as asked by [12].

Overall, the findings indicate that LLM-based approaches provide a viable and flexible alternative for application classification, extending prior work of e.g. [10, 9, 8, 11]. The artifact demonstrates that application descriptions can be effectively interpreted and structured without extensive preprocessing or training data, representing a shift from traditional classification pipelines toward prompt-driven approaches.

6.1. Contributions

This study makes three main contributions to the literature on mobile application classification and SportsHCI.

First, the study provides a methodological contribution by demonstrating how LLMs can be used for large-scale, multi-dimensional classification of mobile applications. Unlike prior approaches that rely on predefined feature engineering or labeled datasets, the proposed approach leverages LLMs as semantic classifiers, enabling flexible classification directly from application descriptions. The results show that this approach can achieve high agreement with human judgment while significantly improving scalability.

Second, the study contributes an artifact in the form of a two-module pipeline that integrates automated data collection with LLM-based classification. The artifact enables efficient processing and structured categorization of large application datasets across multiple dimensions, including user groups, purposes, and sport types. This extends existing classification approaches by

addressing the complexity and scale of real-world application ecosystems in the SportsHCI domain. To support transparency, reproducibility, and future reuse, the artifact has been published as open-source software and is publicly available through GitHub: https://github.com/jukarjalainen/AICCSA---AI-tool-for-Collecting-and-Classifying-Sports-Applications

Third, the study provides an empirical contribution by producing a large-scale, structured dataset of 46,862 sports and health & fitness applications. The dataset combines application metadata with LLM-generated multi-dimensional classifications, enabling systematic and reproducible analysis of the SportsHCI application landscape. To support transparency, reproducibility, and future reuse, the dataset has been published through the Fairdata Etsin research data service and is publicly available at: https://etsin.fairdata.fi/dataset/9eb5dacd-9444-4bb0-8794-ad139250a48d

Together, these contributions demonstrate how LLM-based approaches can advance mobile application classification by improving scalability, flexibility, and reducing reliance on manual labeling.

6.2. Limitations and Future Research

This study has several limitations that suggest directions for future research. First, the evaluation was conducted on a relatively small sample of applications compared to the full dataset. While the results indicate high agreement with human classification, a larger evaluation set or multiple human evaluators would provide a more robust assessment of reliability. Future work could extend the evaluation by incorporating inter-rater agreement and broader validation datasets.

Second, the classification process relies on prompt-based interaction with a LLM, making the outcomes sensitive to prompt design and model configuration as noted by [19, 20, 21]. Although the prompt was iteratively refined, alternative prompting strategies, model choices, or modifications to prompt structure and ordering may yield different results. Future research could systematically compare models, prompting techniques, and prompt components to better understand their influence on classification outcomes.

Third, the classification of application purposes revealed inherent ambiguity between closely related categories. This highlights a conceptual limitation in defining clear boundaries between application types, particularly in complex domains such as SportsHCI. Future work could refine classification taxonomies, explore multi-label approaches that better capture overlapping functionalities, and investigate edge cases, excluded applications, and the primary reasons why certain entries are filtered out during the classification process.

Fourth, the dataset was collected from the Google Play Store and Apple App Store using publicly available APIs, which impose limitations on data access and coverage. As a result, the dataset does not represent the full population of available applications. Future research could explore alternative data collection methods and extend the approach to additional application domains.

Finally, while the study demonstrates the feasibility of LLM-based classification for large-scale datasets, further research is needed to examine its generalizability across different contexts and classification tasks. This includes applying similar approaches to other domains, evaluating long-term robustness as application ecosystems evolve, and exploring interactive classification approaches where categories can be adapted or supplemented by user-defined inputs.

7. Conclusion

This study addressed the challenge of large-scale collection and classification of mobile sports and health & fitness applications, which has traditionally relied on labor-intensive manual processes as well as feature-engineered and supervised machine learning approaches. To overcome this limitation, a large language model-based artifact was developed following a design science research approach, integrating automated data collection with LLM-based classification.

The artifact achieved relatively high agreement with human classification while significantly reducing the time required to process large datasets. These results demonstrate that LLM-based approaches

can support scalable and flexible classification of mobile applications without reliance on predefined features or labeled training data.

Overall, the study shows that LLM-based artifacts enable systematic analysis of complex application ecosystems and provide a foundation for future research in SportsHCI and related domains. To support transparency, reproducibility, and future reuse, both the developed artifact and the resulting dataset have been made publicly available.

Acknowledgments

This research was financially supported by a scholarship from the ITEE Faculty and INTERACT Research Unit, together with a doctoral scholarship from the University of Oulu Graduate School (UniOGS).

Declaration on Generative AI

The authors used ChatGPT for improving writing style, grammar and spelling check, together with AI-assisted software development tools for code generation and programming assistance when implementing the research artifact. After using these tools, the authors reviewed and edited content manually and take full responsibility for the final content.

References

[1] AppBrain, Android app market statistics, 2025. URL: https://www.appbrain.com/stats/android-market-app-categories.

[2] A. Tubek, M. Duplaga, The assessment of functionalities of mobile applications supporting physical activity, Rozprawy Naukowe Akademii Wychowania Fizycznego we Wrocławiu 61 (2018).

[3] M. P. Shaw, L. P. Satchell, S. Thompson, E. T. Harper, C. Balsalobre-Fernández, D. J. Peart, Smartphone and tablet software apps to collect data in sport and exercise settings: Cross-sectional international survey, JMIR mHealth and uHealth 9 (2021). doi:10.2196/21763.

[4] A. A. Pinem, D. I. Sensuse, R. R. Suryono, K. Kautsarina, A. N. Hidayanto, Mobile fitness application quality analysis: Kano model satisfaction or dissatisfaction, Journal of Infrastructure, Policy and Development 8 (2024). doi:10.24294/jipd.v8i8.5960.

[5] I. K. Buntoro, R. Kosala, Experimentation of gamification for health and fitness mobile application, in: Proceedings of the 2019 International Congress on Applied Information Technology (AIT), IEEE, 2019, pp. 1–8. doi:10.1109/AIT49014.2019.9144842.

[6] L. Goebeler, W. Standaert, X. Xiao, Hybrid sport configurations: The intertwining of the physical and the digital, in: Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS), 2021, pp. 5841–5850. URL: https://hdl.handle.net/10125/71328.

[7] L. Haffner, I. Oshri, J. Kotlarsky, Directions for future is research on sports digitalisation: A stakeholder perspective, Journal of Strategic Information Systems 34 (2025). doi:10.1016/j.jsis.2025.101905.

[8] F. Ebrahimi, M. Tushev, A. Mahmoud, Classifying mobile applications using word embeddings, ACM Transactions on Software Engineering and Methodology 31 (2022). doi:10.1145/3474827.

[9] M. V. Villasana, I. M. Pires, J. Sá, N. M. Garcia, E. Zdravevski, I. Chorbev, P. Lameski, F. Flórez-Revuelta, Mobile applications for the promotion and support of healthy nutrition and physical activity habits, The Open Bioinformatics Journal 12 (2020) 50–71. doi:10.2174/1875036201912010050.

[10] Q. Xu, G. Ibrahim, N. Archer, R. Zheng, Toward automated categorization of mobile health and fitness applications, in: Proceedings of the International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc), 2014, pp. 49–54. doi:10.1145/2633651.2633658.

[11] B. Chembakottu, H. Li, F. Khomh, A large-scale exploratory study of android sports apps in the google play store, Information and Software Technology 164 (2023) 107321.

[12] M. Höhn, U. von Jan, T. Framke, U.-V. Albrecht, Classification of health related applications, Studies in Health Technology and Informatics 226 (2016) 139–142.
[13] M. Kebede, B. Steenbock, S. M. Helmer, J. Sill, T. Möllers, C. R. Pischke, Identifying evidence-informed physical activity apps: Content analysis, JMIR mHealth and uHealth 6 (2018). doi:10.2196/10314.
[14] L. Banh, G. Strobel, Generative artificial intelligence, Electronic Markets 33 (2023). doi:10.1007/s12525-023-00680-1.
[15] R. Gozalo-Brizuela, E. C. Garrido-Merchán, A survey of generative ai applications, arXiv preprint arXiv:2306.02781 (2023). URL: http://arxiv.org/abs/2306.02781.
[16] N. Maslej, et al., Artificial intelligence index report 2025, 2025.
[17] J. White, Q. Fu, S. Hays, et al., A prompt pattern catalog to enhance prompt engineering with chatgpt, arXiv preprint arXiv:2302.11382 (2023). URL: http://arxiv.org/abs/2302.11382.
[18] A. Thakur, The Art of Prompting: Unleashing the Power of Large Language Models, 2024.
[19] M. Zheng, J. Pei, L. Logeswaran, M. Lee, D. Jurgens, When a helpful assistant is not really helpful: Personas in system prompts do not improve performance of large language models, arXiv preprint (2024).
[20] P. H. Luz de Araujo, P. Röttger, D. Hovy, B. Roth, Principled personas: Defining and measuring the intended effects of persona prompting, arXiv (2025).
[21] X. Bai, A. Holtzman, C. Tan, "you are a brilliant mathematician" does not make llms act like one, in: Proceedings of the Open Conference on AI Agents for Science, 2025, pp. 1–22. URL: https://openreview.net/pdf?id=sVaRgmH8FE.
[22] Z. R. Tam, C.-K. Wu, Y.-L. Tsai, C.-Y. Lin, H. Lee, Y.-N. Chen, Let me speak freely? a study on the impact of format restrictions on performance of large language models, arXiv preprint arXiv:2408.02442 (2024). URL: http://arxiv.org/abs/2408.02442.
[23] K. Peffers, T. Tuunanen, M. A. Rothenberger, S. Chatterjee, A design science research methodology for information systems research, Journal of Management Information Systems 24 (2007) 45–77. doi:10.2753/MIS0742-1222240302.
[24] F. Mueller, D. Young, 10 lenses to design sports-hci, Foundations and Trends in Human-Computer Interaction 12 (2018) 172–237. doi:10.1561/1100000076.
[25] M. I. M. E. Moilanen, N. Iivari, L. Arhippainen, A. Lanamäki, Guidelines for disc golf applications and design principles for sportshci: A human-centered approach, in: Proceedings of the International Conference on Mobile and Ubiquitous Multimedia, 2024, pp. 396–411. doi:10.1145/3701571.3701584.
[26] X. Xiao, J. Hedman, F. T. C. Tan, C.-W. Tan, T. Clemmensen, Sports digitalization: A review and research agenda, in: Proceedings of the 38th International Conference on Information Systems (ICIS), 2017, pp. 1–21.
[27] E. D. Mekler, K. Hornbæk, A framework for the experience of meaning in human-computer interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–15. doi:10.1145/3290605.3300455.
[28] M. I. M. E. Moilanen, N. Iivari, L. Arhippainen, A. Lanamäki, Defining hci-ikigai: Designing for existence and meaning in human-centered design, in: Proceedings of the 28th International Academic Mindtrek Conference, 2025, pp. 326–339. doi:10.1145/3757980.3758008.
[29] P. Olla, C. Shimskey, mhealth taxonomy: A literature survey of mobile health applications, Health and Technology 4 (2015) 299–308. doi:10.1007/s12553-014-0093-8.
[30] L. A. Cotfas, A. Sandu, C. Delcea, P. Diaconu, C. Frasineanu, A. Stanescu, From transformers to chatgpt: An analysis of large language models research, IEEE Access 13 (2025) 146889–146931.
[31] S. Mohammadabadi, B. C. Kara, C. Eyupoglu, C. Uzay, M. S. Tosun, O. Karaku, A survey of large language models: Evolution, architectures, adaptation, benchmarking, applications, challenges, and societal implications, Electronics (2025).
[32] M. Kalra, V. Sharma, Impact of low rank adaptation with parameter efficient fine-tuning techniques on transformer models, 2025 6th International Conference for Emerging Technology (INCET) (2025) 1–6. URL: https://api.semanticscholar.org/CorpusID:281150724.

[33] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, P. S. Dulepet, S. Vidyadhara, D. Ki, S. Agrawal, C. M. Pham, G. C. Kroiz, F. Li, H. Tao, A. Srivastava, H. D. Costa, S. Gupta, M. L. Rogers, I. Goncearenco, G. Sarli, I. Galynker, D. Peskoff, M. Carpuat, J. White, S. Anadkat, A. M. Hoyle, P. Resnik, The prompt report: A systematic survey of prompting techniques, ArXiv abs/2406.06608 (2024). URL: https://api.semanticscholar.org/CorpusID:270380093.

[34] A. R. Hevner, S. T. March, J. Park, S. Ram, Design science in information systems research, MIS Quarterly 28 (2004).