

# Leveraging Model Context Protocol to Enhance AI Educational Agents: The STEAMBrace Tester Case

Ander Arce<sup>1, \*, †</sup>, Aitziber Sagastizabal<sup>1, †</sup>, Javier Portillo<sup>1, †</sup>, Urtza Garay<sup>1, †</sup>

<sup>1</sup> University of the Basque Country, Leioa Barrio Sarriena 48940, Basque Country

## Abstract

Large Language Models (LLMs) such as GPT-4 are rapidly entering classrooms as conversational tutors, yet their closed architectures leave four critical gaps for learning analytics: context is lost between sessions, interaction logs are unstructured, a single chatbot performs all functions without specialised traceability, and all data are stored on commercial servers outside institutional oversight. This paper explores how the open-standard Model Context Protocol (MCP) can bridge these gaps through a conceptual redesign of *STEAMBrace Tester*, an assistant that helps secondary-school teachers and educators refine STEAM activities. The proposed version replaces the cloud-hosted GPT with Claude Desktop running on institutional servers, links creator and evaluator agents through MCP, and records every exchange as xAPI statements in a local Learning Record Store. This configuration preserves each teacher's history, enables longitudinal analyses of activity quality, isolates the contribution of each agent, and ensures full data sovereignty under GDPR. Illustrative scenarios show how the enriched traces permit investigation of feedback uptake, the evolution of equity-oriented prompts, and the optimal number of iterative cycles. The resulting architecture offers a transferable pathway for educational institutions to reclaim analytic value from LLM-driven assistants while maintaining rigorous privacy and governance standards.

## Keywords

Educational Large Language Models, Model Context Protocol, Learning Analytics.

## 1. Introduction

The emergence of large language models (LLMs), such as GPT-4, has popularized conversational assistants capable of generating and evaluating content of all types, including educational. These models offer insightful responses, but their "black box" architecture poses four critical barriers to the Learning Analytics (LA) discipline: Ephemeral Memory, Limited Instrumentation, Monolithic Agent and Centralized Storage [1,2].

The fourth barrier is particularly relevant, since all "insights" on usage patterns, product improvements or detection of training needs remain within the company hosting the LLM. External educational institutions such as schools or universities, which own the teaching practice, receive at most basic dashboards without access to raw data, which limits educational research and evidence-based decision-making.

In 2024, Model Context Protocol (MCP) emerged as an open standard that acts as a "USB-C port" for AI, enabling bidirectional and secure connections between LLMs and external services [3,4]. MCP introduces an intermediate layer (middleware, MCP clients and servers) that solves previous problems by enabling: Persistent context, Exhaustive data collection, Multi-agent collaboration and Institutional middleware. These capabilities open a pathway to overcome barriers for LLMs in educational settings, where decision making must be based on longitudinal and contextualized evidence [5,6].

\* Corresponding author.

† These authors contributed equally.

 [ander.arce@ehu.eus](mailto:ander.arce@ehu.eus) (A. Arce); [aitziber.sagastizabal@ehu.eus](mailto:aitziber.sagastizabal@ehu.eus) (A. Sagastizabal); [javier.portillo@ehu.eus](mailto:javier.portillo@ehu.eus) (J. Portillo); [urtza.garay@ehu.eus](mailto:urtza.garay@ehu.eus) (U. Garay).

 [0000-0002-2172-6025](https://orcid.org/0000-0002-2172-6025) (A. Arce); [0009-0002-7093-8822](https://orcid.org/0009-0002-7093-8822) (A. Sagastizabal); [0000-0002-0265-9277](https://orcid.org/0000-0002-0265-9277) (J. Portillo); [0000-0001-7298-9274](https://orcid.org/0000-0001-7298-9274) (U. Garay)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

To illustrate this potential, this paper presents the STEAMBrace Tester case study, a customized GPT created in the STEAMBrace project (<https://steambraceproject.eu/>), funded by the European Horizon call, to help secondary school teachers and other educators improve STEAM activities. The current version, STEAMBrace Tester GPT, provides immediate value, but it is still imprisoned by previous barriers. This paper proposes the conceptual design of STEAMBrace Tester MCP, an evolution that leverages MCP to provide the tool with persistent memory, structured logging, and institutional governance of the data, thus expanding the spectrum of Learning Analytics available.

The objectives of this work are first, to analyze the limitations of LLMs with respect to LA and data sovereignty; secondly, to describe how MCP addresses these limitations; lastly, to demonstrate, using STEAMBrace, the pedagogical and analytical improvements obtained.

## 2. Theoretical Framework

### 2.1. Barriers of LLMs in Learning Analytics

The incorporation of large language models (LLMs) into teaching has led to conversational tutors, material generators, and automated evaluators that provide near-instantaneous and differentiated feedback. The Learning Analytics (LA) literature recognizes this potential, but stresses that the analytics infrastructure remains immature and poorly integrated with educational research processes [7].

In practice, LLMs face several obstacles that reduce their analytical usefulness. First, their working memory is ephemeral: the context window-although expanding-continues to limit continuity between sessions and hinders longitudinal tracking of learning [8]. Second, instrumentation is sparse; chat histories are often stored as free text without standardized metadata, complicating their exploitation by LA techniques [6]. Third, the usual paradigm is that of a monolithic agent attempting simultaneously to generate, evaluate, and recommend, missing the evidence supporting multi-agent designs in intelligent tutoring systems [9]. Finally, there is external data governance: conversational traces are stored and processed on the provider's servers (e.g., OpenAI), so that the educational institution or researcher does not control, nor fully capitalize on, the resulting analytics, perpetuating an asymmetric "datafication" [1].

These limitations are properly understood in light of the architecture of OpenAI's custom GPTs, released in 2023. Each GPT is configured using three layers: (i) an instruction block-system prompt that defines the desired behavior, (ii) a repository of knowledge documents that are attached as additional context, and (iii) a set of tools or actions that empower the model to call external APIs [8,10]. While this design enhances customization and extensibility, all telemetry is greatly hampered because the data (prompts, responses, tool invocations) are hosted on the OpenAI infrastructure, giving the provider a privileged position to exploit usage analytics, while the school receives, at best, aggregated summaries.

### 2.2. MCP Capabilities

To address these limitations, Anthropic introduced in 2024 the Model Context Protocol (MCP), described as a "USB-C for AI" because it standardizes bidirectional connections between models and any external data source or tool, avoiding ad hoc integrations on a case-by-case basis [3].

The protocol is articulated in three logical components [4]. The MCP host is the LLM itself, which originates requests when it needs information or must perform an action. These requests are

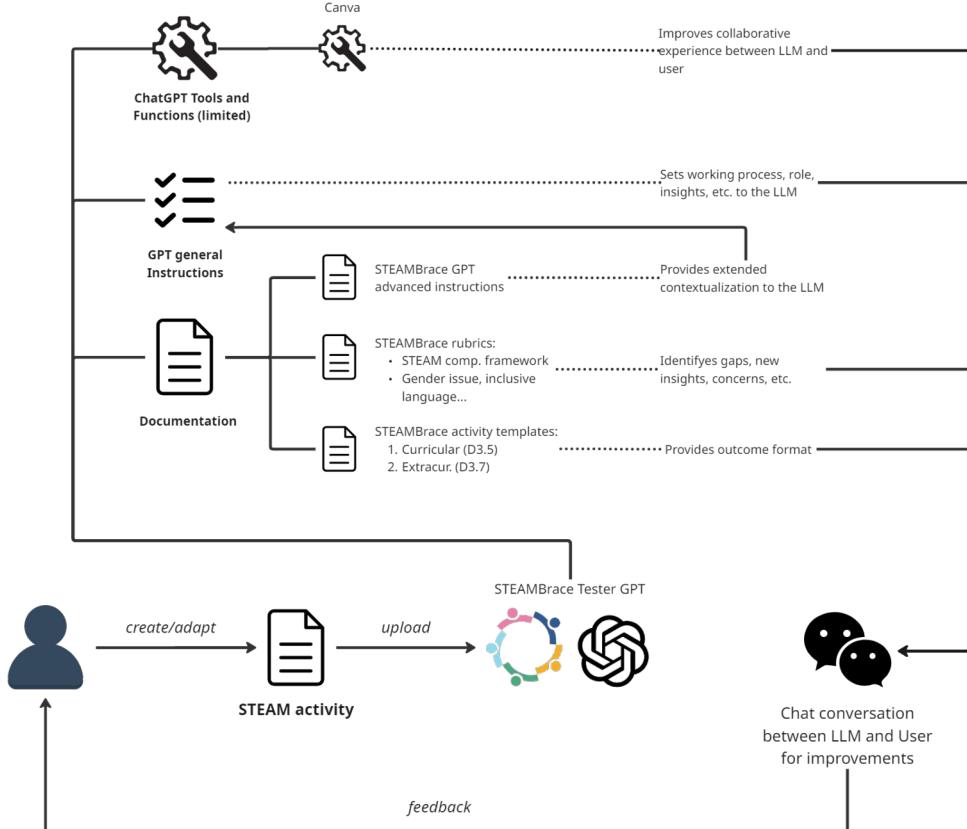
sent to the MCP client, a middleware that validates, formats and executes the request. Finally, one or more MCP servers attend the request: they query a database, retrieve a file from a corporate drive or invoke a teaching API, and return the response to the client, which reintegrates it into the model prompt.

From an analytical perspective, MCP introduces four decisive capabilities. First, it enables persistent context: before generating a response, the model queries profiles, histories or artifacts stored outside its context window, overcoming ephemeral memory [4]. Second, each exchange can be serialized as an xAPI-compliant JSON statement, allowing to record not only satisfaction surveys, but also the entire sequence of messages, reading times, successive versions of an activity, or even emotional indicators inferred from the language. Third, the same standard channel allows multiple specialized agents to collaborate—for example, an activity generator and a rubric-based evaluator—sharing a common state without loss of context. Fourth, by deploying MCP servers on-premises or in the institution's private cloud, data governance and fine-grained permission policies remain under local control, facilitating GDPR compliance and reversing the current asymmetry in the exploitation of conversational traces.

### **3. Case Study: STEAMBrace Tester GPT**

In order to understand the tool itself, first, it is necessary to explain the pedagogical context and objective of STEAMBrace (<https://steambraceproject.eu/>). The European STEAMBrace project aims to reduce the gaps in access and motivation towards STEAM careers in school citizenship. It articulates a hybrid methodology that combines Problem-Based Learning (PBL), Challenge-Based Learning (CBL), and Gagné's nine-phase principles methodology. It is also aligned with the principles of Design Thinking and Maker methodology for learning and prototyping orientation.

Within this framework, the project has developed the STEAMBrace Tester GPT tool, a conversational assistant designed for teachers and other educational agents teaching secondary school students who wish to share their existing STEAM activities and expand their proposals, identify biases (e.g. gender), make improvements, or adapt their proposals to curricular or extracurricular environments.



**Figure 1:** Workflow of STEAMBrace Tester GPT. Own elaboration.

In the current architecture, the tool was built with OpenAI's Create a GPT function, which is based on three configurable layers: system instructions, knowledge documents and external tools. In the case of STEAMBrace GPT (see Figure 1), the system instructions define the evaluator role, the interaction process, and reference different shared files: the knowledge documents, which is composed of: the STEAM competency assessment rubric, templates of curricular and extracurricular activities and a rubric aligned with gender equity and inclusive language. Finally, the only tool enabled is the canvas mode, as the web search function has been omitted to eliminate hallucinations or errors in the processing of the system documents.

The workflow is as follows. In the *Step 1, Language and Description*, the teacher chooses the language in which to interact and receive feedback; and explains whether their activity is curricular or extracurricular, for the GPT to decide which template to use. In *Step 2, Share*, the teacher or educational agent shares the description of his/her activity (or creates it on the spot). In *Step 3, Prompt -Chain*, the GPT analyzes the text, checks it against the rubrics and assigns preliminary scores. In *Step 4, Feedback*, the GPT generates suggestions for improvement (variants, resources, evaluation indicators), and shares them with the user. Lastly, in *Step 5, Fine-Tuning*, the teacher iterates until a satisfactory version is obtained and may request a resource pack (e.g., link to Canva, Maker materials list).

Despite its initial adoption planned for Q2-2025, the prototype exhibits the same constraints that affect commercial LLMs, as previously discussed. The assistant loses its internal memory upon logout, meaning it does not retain previously reviewed activities, which complicates any effort to track a teacher's progress over time. Additionally, all interaction logs—including messages, timestamps, and revisions—are stored as unstructured plain text on OpenAI servers. Analyzing the sequence of interactions either at the individual or collective level would require manual extraction

and structuring, which is incompatible with xAPI standards. Furthermore, the model operates as a single agent: it generates and evaluates content but lacks the capacity to delegate tasks to specialized agents or to share task status across models. This restricts the potential for collaborative evaluation processes involving agents focused on specific areas such as PBL, equity, or particular STEM disciplines. Finally, data governance remains in the hands of the provider. Project administrators only receive summary reports, while full interaction traceability is not accessible to the research team, thus hindering the ability to conduct rigorous impact assessments. For instance, although the tool is currently being used by over 50 educators across Europe, aggregate-level information on user interactions remains unavailable.

These limitations motivate the design of the *STEAMBrace Tester MCP* version, described in the following section.

#### 4. Theoretical extension: STEAMBrace Tester MCP

The STEAMBrace Tester MCP version represents a conceptual evolution of the educational assistant, replacing the cloud-hosted GPT with a Claude Desktop model deployed on in-house servers, connected via the Model Context Protocol (MCP) to internal services managed by the institution. While the user experience remains that of an accessible and friendly conversational assistant, operations that previously occurred opaquely within the commercial provider's infrastructure are now distributed in independent components that favor personalization, data control and analytical exploitation from the Learning Analytics framework.

The following is a summary (see Table 1) of the main phases of the STEAMBrace Tester MCP workflow, together with the analytical functionalities that are activated in each of them:

**Table 1.**

Workflow phases and corresponding analytical functionality. Source: own elaboration.

Workflow Phase	User action	CCM Component	Analytical functionality enabled
Step 1. Language and Description	Selection of language and type of activity (curricular and extracurricular)	Local server	Consultation of teaching history, retrieval of previous activities, activation of contextual profile
Step 2. Share	Submission or writing of the activity by the teacher	Host Claude internal storage	+ Start of interaction logging, activation of conversational logs
Step 3. Prompt Chain.	Generation of suggestions and review of the activity	Host Claude + LRS	xAPI structured logging (message, version, metadata, times)
Step 4. Feedback	Recommendations, improvements, examples and	Host Claude + Evaluator Server	Separate evaluation, agent type-specific tracing, inter-model

	resources	(multi-agent)	comparison
Step 5. Final review ( <i>fine-tuning</i> )	Final iteration, acceptance or adjustment of recommendations	Host Claude + LRS	Final version logging, report export, generation of longitudinal data

From the first step of the interaction, specific functionalities are activated. When the teacher selects the language and specifies whether the activity is curricular or extracurricular, the system accesses a local memory base where it consults previous activities linked to the same user. This persistent memory allows the model to adapt its recommendations according to the history, generating pedagogical continuity and longitudinal data that can be analyzed later to study, for example, the evolution of quality criteria or the progressive incorporation of equity elements.

In the second step, when the teacher shares his activity (either written from scratch or adapted), the main analysis flow is activated. From that moment on, each exchange of messages, settings or questions is automatically recorded in a Learning Record Store (LRS) using statements structured under the xAPI standard. This includes not only the text of the interactions, but also metadata such as the response time, the type of suggestion requested or the latency between actions. This traceability makes the wizard an instrumented tool, capable of providing evidence at both the individual and institutional level.

In the following steps, when the model generates feedback and suggestions, this content is linked to intermediate versions of the activity, which are also stored and tagged in the LRS. In this way, not only the final version of the proposal is preserved, but also the entire iterative process that produced it. This information makes it possible to study the number of improvement cycles, the evolution of quality according to the STEAMBrace rubric, or even the sequence of actions that generate the most added value.

Finally, in the final review phase, the system continues to record any future interaction. Moreover, having functionally separated the agents, it is possible to refer the final proposal to a different evaluation model (e.g., a Llama-3 refined to apply the rubric or detect gender bias), allowing a specialized and complementary evaluation to the creative assistant. Both interventions—creation and evaluation—are recorded separately, facilitating their comparative analysis.

This architecture also allows non-invasive collection of complementary traces: clicks on suggested external resources, permanence at each step, language changes, use of templates... All these variables are securely stored in an on-premises environment, encrypted and under institutional control, which facilitates GDPR compliance and reinforces the sovereignty of educational data.

## 5. STEAMBrace Tester advantages and challenges

The architecture presented in the previous section allows for a qualitative leap in the analysis and understanding of teaching work in the specific context of STEAM. Beyond solving the technical limitations already identified, STEAMBrace Tester MCP opens the possibility of developing new lines of analysis and functions that were previously unattainable.

For example, with the data collected by the LRS it is possible to reconstruct the complete cycle of design, revision and improvement of a STEAM activity. This makes it possible to analyze how many iterations a teacher performs, what kind of suggestions he/she accepts and how his/her score on the rubric evolves after each change. With this traceability, customized mentoring programs could be designed based on real improvement trajectories, or identify critical moments where creativity or the integration of approaches such as CBL stagnates.

Another possibility lies in the aggregate analysis by type of teacher or center. The database could be linked to variables such as educational level, teaching seniority, or rural/urban context, making it possible to study whether certain profiles respond differently to certain suggestions, or whether the intensive use of the assistant correlates with a greater diversity of activities or with a sustained improvement in equity criteria.

From a pedagogical point of view, new forms of support could be activated. For example, if a teacher is developing an activity with a Maker approach but shows doubts or setbacks in the versions, the system could suggest specific support resources or connect him/her with another user who has overcome similar difficulties. These recommendations could be managed by specialized agents, integrated through MCP, without the user perceiving a greater complexity.

In terms of educational research, the repository of structured and contextualized data would make it possible to advance in questions that have not yet been explored, such as the impact of AI assistants on teacher self-regulation, or the way in which different design sequences -for example, generation first, evaluation later, or vice versa- affect the final quality of the activities. In addition, the use of inclusive language, the representation of female referents in the activities or the explicit attention to equity issues could be studied, extracting patterns from the processed texts themselves.

Although there are currently initiatives such as Playlab.ai [11], which allow teachers to create their own conversational assistants and access basic usage metrics, these solutions are still subject to significant structural constraints. Operating within a closed platform, it is not possible to integrate agents with proprietary databases or deploy them in institutional environments. In addition, the analytics available are limited and are produced on external servers, which prevents schools or universities from exercising full governance over the data generated. In terms of privacy, control and analytical capabilities, these tools do not yet offer the depth and sovereignty required for a rigorous, GDPR-aligned Learning Analytics approach.

Therefore, in addition to being a technical solution to previous barriers, STEAMBrace Tester MCP is configured as a living analytical infrastructure, capable of generating applicable educational knowledge and facilitating informed pedagogical decisions. This shift from a closed and opaque architecture to an extensible, governed and learning-oriented system represents a relevant contribution to the debate on the responsible use of AI in education.

## Acknowledgments



Funded by  
the European Union

Funded by the European Union—European Innovation Council STEAMBrace project—Grant Agreement nr. 101132652. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Innovation Council. Neither the European Union nor the granting authority can be held responsible for them.

## Declaration of Generative AI

During the preparation of this work, the authors used DeepL and ChatGPT-4 in order to: grammar, spelling and overall improvements of translation. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References

- [1] Wang, Xin, Tapani Ahonen, and Jari Nurmi. "Applying CDMA technique to network-on-chip." *IEEE transactions on very large scale integration (VLSI) systems* 15.10 (2007): 1091-1100.
- [2] P. S. Abril, R. Plant, The patent holder's dilemma: Buy, sell, or troll?, *Communications of the ACM* 50 (2007) 36–44. doi:10.1145/1188913.1188915.
- [3] S. Cohen, W. Nutt, Y. Sagiv, Deciding equivalences among conjunctive aggregate queries, *J. ACM* 54 (2007). doi:10.1145/1219092.1219093.
- [4] J. Cohen (Ed.), Special issue: Digital Libraries, volume 39, 1996.
- [5] D. Kosiur, *Understanding Policy-Based Networking*, 2nd. ed., Wiley, New York, NY, 2001.
- [6] D. Harel, First-Order Dynamic Logic, volume 68 of *Lecture Notes in Computer Science*, Springer-Verlag, New York, NY, 1979. doi:10.1007/3-540-09237-4.
- [7] I. Editor (Ed.), The title of book one, volume 9 of The name of the series one, 1st. ed., University of Chicago Press, Chicago, 2007. doi:10.1007/3-540-09237-4.
- [8] I. Editor (Ed.), The title of book two, The name of the series two, 2nd. ed., University of Chicago Press, Chicago, 2008. doi:10.1007/3-540-09237-4.
- [9] A. Z. Spector, Achieving application requirements, in: S. Mullender (Ed.), *Distributed Systems*, 2nd. ed., ACM Press, New York, NY, 1990, pp. 19–33. doi:10.1145/90417.90738.
- [10] B. P. Douglass, D. Harel, M. B. Trakhtenbrot, Statecharts in use: structured analysis and object-orientation, in: G. Rozenberg, F. W. Vaandrager (Eds.), *Lectures on Embedded Systems*, volume 1494 of *Lecture Notes in Computer Science*, Springer-Verlag, London, 1998, pp. 368–394. doi:10.1007/3-540-65193-4\_29.
- [11] D. E. Knuth, *The Art of Computer Programming*, Vol. 1: *Fundamental Algorithms* (3rd. ed.), Addison Wesley Longman Publishing Co., Inc., 1997.
- [12] S. Andler, Predicate path expressions, in: *Proceedings of the 6th. ACM SIGACT-SIGPLAN symposium on Principles of Programming Languages*, POPL '79, ACM Press, New York, NY, 1979, pp. 226–236. doi:10.1145/567752.567774.
- [13] S. W. Smith, An experiment in bibliographic mark-up: Parsing metadata for xml export, in: R. N. Smythe, A. Noble (Eds.), *Proceedings of the 3rd. annual workshop on Librarians and Computers*, volume 3 of *LAC '10*, Paparazzi Press, Milan Italy, 2010, pp. 422–431. doi:99.9999/woot07-S422.
- [14] M. V. Gundy, D. Balzarotti, G. Vigna, Catch me, if you can: Evading network signatures with web-based polymorphic worms, in: *Proceedings of the first USENIX workshop on Offensive Technologies*, WOOT '07, USENIX Association, Berkley, CA, 2007.
- [15] D. Harel, *LOGICS of Programs: AXIOMATICS and DESCRIPTIVE POWER*, MIT Research Lab Technical Report TR-200, Massachusetts Institute of Technology, Cambridge, MA, 1978.
- [16] K. L. Clarkson, Algorithms for Closest-Point Problems (Computational Geometry), Ph.D. thesis, Stanford University, Palo Alto, CA, 1985. UMI Order Number: AAT 8506171.
- [17] D. A. Anisi, Optimal Motion Control of a Ground Vehicle, Master's thesis, Royal Institute of Technology (KTH), Stockholm, Sweden, 2003.
- [18] H. Thornburg, Introduction to bayesian statistics, 2001. URL: <http://ccrma.stanford.edu/jos/bayes/bayes.html>.
- [19] R. Ablamowicz, B. Fauser, Clifford: a maple 11 package for clifford algebra computations, version 11, 2007. URL: <http://math.tntech.edu/rafal/cli11/index.html>.

- [20] Poker-Edge.Com, Stats and analysis, 2006. URL: <http://www.pkredge.com/statsYYFWWQ.php>.
- [21] B. Obama, A more perfect union, Video, 2008. URL: <http://video.google.com/videoplay?docid=6528042696351994555>.
- [22] D. Novak, Solder man, in: ACM SIGGRAPH 2003 Video Review on Animation theater Program: Part I - Vol. 145 (July 27–27, 2003), ACM Press, New York, NY, 2003, p. 4. URL: <http://video.google.com/videoplay?docid=6528042696351994555>. doi:99.9999/woot07-S422.
- [23] N. Lee, Interview with bill kinder: January 13, 2005, Comput. Entertain. 3 (2005). doi:10.1145/1057270.1057278.
- [24] J. Scientist, The fountain of youth, 2009. Patent No. 12345, Filed July 1st., 2008, Issued Aug. 9th., 2009.
- [25] B. Rous, The enabling of digital libraries, Digital Libraries 12 (2008). To appear.
- [26] M. Saeedi, M. S. Zamani, M. Sedighi, A library-based synthesis methodology for reversible logic, Microelectron. J. 41 (2010) 185–194.
- [27] M. Saeedi, M. S. Zamani, M. Sedighi, Z. Sasanian, Synthesis of reversible circuit using cycle-based approach, J. Emerg. Technol. Comput. Syst. 6 (2010).
- [28] M. Kirschmer, J. Voight, Algorithmic enumeration of ideal classes for quaternion orders, SIAM J. Comput. 39 (2010) 1714–1747. URL: <http://dx.doi.org/10.1137/080734467>. doi:10.1137/080734467.
- [29] L. Hörmander, The analysis of linear partial differential operators. IV, volume 275 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin, Germany, 1985. Fourier integral operators.
- [30] L. Hörmander, The analysis of linear partial differential operators. III, volume 275 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin, Germany, 1985. Pseudodifferential operators.
- [31] IEEE, Ieee tcsc executive committee, in: Proceedings of the IEEE International Conference on Web Services, ICWS '04, IEEE Computer Society, Washington, DC, USA, 2004, pp. 21–22. doi:10.1109/ICWS.2004.64.
- [32] TUG, Institutional members of the TEX users group, 2017. URL: <http://www.tug.org/instmem.html>.
- [33] R Core Team, R: A language and environment for statistical computing, 2019. URL: <https://www.R-project.org>.
- [34] S. Anzaroot, A. McCallum, UMass citation field extraction dataset, 2013. URL: <http://www.iesl.cs.umass.edu/data/data-umasscitationfield>.