2025-09-30

LLMs in R Book

Nic Crane

Executive Summary

This proposal requests $19,080 USD to bootstrap a free and open, community-owned online book on LLMs in R. LLMs are a popular topic at the moment, and both the underlying technologies and tooling in the R ecosystem are moving quickly. Reliable guidance exists, but is scattered across package documentation, vignettes, and blog posts. The book will consolidate best practices, key resources, and accessible explanations into a single structured resource. The outcome will be a living, sustainable resource that strengthens R's role in modern AI and leaves analysts, data scientists, and educators equipped to use LLMs effectively. After the funded phase, hosting is cost-free and governance will be community-led, ensuring long-term value without further ISC expense.

Signatories

Project team

- Project lead and primary initial author - Nic Crane: Nic is an open source maintainer and holds a PhD in Applied Social Statistics from Lancaster University. Nic recently co-authored "Scaling Up with R and Apache Arrow", available online at arrowrbook.com and published by CRC Press in July 2025. Nic has a passion for teaching, with a background across in-person workshops, online courses, with excellent learner feedback on the content and accessibility of course materials.
- Core initial contributor - Luis D. Verde Arregoitia holds a PhD in evolutionary biology from the University of Queensland and is a Biodiversity Data Scientist. He is an active contributor to the R community through R Weekly and has authored the guide "LLMs in R" (available at luisdva.github.io/llmsr-book/), which provides foundational documentation of available tools for integrating large language models with R workflows.

Contributors

- Mauro Lepore, associate editor at ROpenSci and software engineer at Recast
- Christoph Scheuch, founder of Tidy Intelligence

Consulted

- Simon Couch, Posit (reviewed initial idea though not complete proposal due to timing constraints)
- Maëlle Salmon, ROpenSci
- Yanina Bellini Saibene, ISC Committee Member
- Kirill Müller, ISC Committee Member

The Problem

LLMs and AI are having a huge impact on software engineering and data science, and the tools available, model capabilities, and advice and guidance on their use are all rapidly changing. Much of the available guidance is scattered across different locations, such as model vendor websites, R package vignettes, and blog posts by community members.

Existing resources provide valuable solution-oriented information, such as how to use specific functions and packages, but R users are left without a clear framework for understanding the wider problem space. As a result, users must piece together knowledge from multiple sources, with no obvious place to start, risking misapplying methods and wasting time and effort. The problem exists for many R users including data scientists, analysts, and educators who want to integrate LLMs into analysis, teaching, and tooling.

This problem could be solved by consolidating best practices, explanations, and guidance into a single structured resource, with further links to other resources, and clear structure for governance and community contribution. There is some existing work in this space, such as Luis D. Verde Arregoitia's guide to tooling for LLMs in R, but this focuses on available tools rather than concepts or context.

It's unlikely that this kind of resource would be a good candidate for a permanent, static reference like a book published via a traditional publisher, and value lies in creating a living, community-oriented resource that can be updated as packages evolve, new methods emerge, and best practices shift, so it can remain relevant.

This approach aligns with successful R Consortium-funded social infrastructure projects, including translation and internationalisation efforts that have strengthened community accessibility, and educational initiatives that have built lasting knowledge resources for the R ecosystem. Like these precedents, this project prioritises sustainable, community-owned infrastructure over static deliverables.

The proposal

Overview

The proposal is to write a free online book about LLMs in R with associated reusable social and technical infrastructure. It will address the problem by providing a centralised location where people can find resources to help learn about the topic. The benefits to the R community is the ability to quickly upskill in an area which is moving quickly and is broad in nature.

The book will prioritise making concepts accessible to readers by employing educational tactics such as using laddering to build complex ideas one piece at a time, avoiding jargon and carefully defining it where it is necessary, as well as using metaphors to break ideas down. The focus would be

on avoiding unnecessary complexity while also equipping the reader with the relevant information to learn independently.

The book would take a problem-oriented approach, starting from the kinds of problems R users face when working with LLMs, and then discussing resources and solutions. It will follow the diataxis documentation framework's definition of "explanation" rather than "tutorial", "how-to", or "reference", and so avoid duplicating existing resources.

It will include discussion of the ethics, for example, debates around model generation, environmental impact, and how to structure work to reduce cost and environmental impact.

This kind of resource is unlikely to be created via the traditional means of publishing a physical book, due to the speed of change in the LLMs/AI space at the moment. Creating this as an online resource means that there is scope to quickly update content, as new technologies and ideas emerge.

Detail

Social Infrastructure

This project is an investment in the R community's ability to engage with AI and LLM technologies. Like past ISC-funded work on translation, education, and documentation, it will create shared knowledge that lowers barriers to participation. Beyond producing a book, the project will set up reusable infrastructure: clear contribution processes, an editorial model, and a plan for long-term maintenance. By pulling together scattered expertise into a single, community-owned resource, it will strengthen R's role in the AI/LLM landscape while providing governance and contribution patterns other projects can adapt.

Minimum Viable Product

A functional MVP will include 3-4 complete foundational chapters providing immediate value to R users, a working Quarto book infrastructure with CI/CD, community contribution guidelines and templates, and a live website accessible to the R community.

Initial planning and repository structure has already been created at https://github.com/thisisnic/llms-in-r to support rapid MVP development.

Architecture

The project will be stored in a public repository on GitHub, using Quarto to create the book, and GitHub Actions to render it. This technical infrastructure will include:

- Reusable CI/CD pipeline for community-maintained documentation projects including Reusable GitHub Actions workflows for Quarto-based R community books, including automated testing, readability score and spellchecks
- Standardised contribution templates and review processes
- Content freshness monitoring system - Automated CI/CD jobs that flag potentially outdated content based on package version changes, broken links, and time-based triggers
- Editorial workflow infrastructure that can be adapted by other R community resources

Community Infrastructure Deliverables

Beyond the book content, this project will deliver reusable community infrastructure: editorial board governance model and processes, and contribution guidelines and templates for collaborative technical documentation that other R community projects can adopt for similar documentation efforts.

Assumptions

One assumption is that changes in applications of LLMs won't change so rapidly that the book would need to be entirely rewritten in the near future. This risk is mitigated in part by keeping content focused on context and explanation over how-to guidance (which already exist). For example, while available underlying models and their capabilities are likely to see many changes over time, deciding factors for choosing an appropriate model will see less change.

External dependencies

The project relies on GitHub Actions for automated builds and on existing R packages that provide LLM functionality.

Project plan

Start-up phase

Administrative tasks

The project is already set up as a repository on GitHub the with appropriate license and initial GitHub Actions CI/CD enabled. Remaining administrative tasks during start-up phase:

- purchase URL for the project which will point to the published book
- set up additional CI jobs to check for readability etc on new pull requests
- expand on first drafts of templates for content

Content planning

In the start-up phase we plan to conduct a detailed review of available resources, and conduct a community survey to identify key topics of most importance to the wider R community. We will use the outcome of this stage to refine the list of high-level topics and focuses.

Technical delivery

Dec 31 2025 - Milestone 1 - Foundation: Community survey conducted and scope finalised. Infrastructure templates and contribution guidelines established. Domain registered and additional CI/CD.

Jan 31 2026 - Milestone 2 - Launch: Initial chapters published. Public launch and community outreach initiated.

Feb 28 2026 - Milestone 3 - Community Infrastructure: Editorial governance established. Community contribution workflows operational. Additional content developed based on survey feedback.

Mar 31 2026 - Milestone 4 - Content Development: Substantial content expansion. Community contributions integrated. Infrastructure templates documented for reuse.

April 30 2026 - Milestone 5 - Complete Draft: Comprehensive first version online. All core topics addressed per community survey.

May 30 2026 - Milestone 6 - Community Transition: Editorial board fully operational. Governance model documented and transferred.

Other aspects

We will publicise the work by posting about it on LinkedIn, Mastodon, and Bluesky upon realisation of each milestone, and actively encouraging community feedback. We will also write content to be

shared quarterly on the R Consortium blog to ensure wide reach, as well as posting on personal blogs and submitting the links to R Weekly, and submitting talks to major R conferences.

We also would like to note that LLMs will not be used to write the contents of this book - human understanding and engagement is key to writing materials which are genuinely useful to others. Any use of LLMs will be for peripheral tasks such as ensuring conformance to defined principles (e.g. lack of use of jargon, readability), creating templates, and checking for grammar/spelling.

Budget & funding plan

We would like to request financial support of $19,080 from the ISC to develop the book materials and transition the project to a community-owned resource. The labour costs use an hourly rate of $100, based on proposals for previous R Consortium sponsored projects.

Milestone	Date	Cost
1: Foundation	Dec 31 2025	$1,580
2: Launch	Jan 31 2026	$3,500
3: Community Infrastructure	Feb 28 2026	$4,000
4: Content Development	Mar 31 2026	$4,000
5: Complete Draft	April 30 2026	$4,000
6: Community Transition	May 30 2026	$2,000

Total funding requested: $19,080 (includes $80 for 3-year domain registration)

Success

Definition of done

- Success means delivering a public, freely available book website on LLMs in R that includes a clear outline and scope of the book agreed through community input, a working CI/CD pipeline, an editorial board in place and contribution guidelines published.

Measuring success

- Traffic to book webpage and number of contributors is steadily increasing.

Future work

After the funded phase, the book can be extended by adding new chapters as R packages evolve or new methods emerge, translating the book into other languages to broaden reach, using the book as a base for workshops, tutorials, or course materials, and maintaining an open contribution model so the community can update, refine, and extend content long term.