2025-09-29

LLMs in R Book

Nic Crane

Executive Summary

This proposal requests $24,080 USD to bootstrap a free, 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.

Over 6 months (Dec 2025 to May 2026), the project will:

- Set up infrastructure (GitHub/Quarto repo, CI/CD, license, domain).
- Conduct a community survey to refine scope and priorities.
- Draft and publish at least ten substantive chapters on relevant topics.
- Publicly release a complete first draft.
- Establish an editorial board and governance model to support future contributions.

Success will be measured by delivery of milestones, publication of 10 chapters, 100 survey responses, published community contributions, a transition to community governance, and a functioning editorial board.

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

Nic Crane 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, having taught “introduction to R workshops” to hundreds of learners earlier in their career while a consultant at Mango Solutions. Nic also has previously developed an online course on package development for Data Camp, and delivered in-person workshops at Posit Conf in 2023 and 2024, with excellent learner feedback on the content and accessibility of course materials. Nic’s experience in authoring online R books, as well as developing educational materials makes them well-suited to lead this project.

Luis D. Verde Arregoitia holds a PhD in evolutionary biology from the University of Queensland and is a postdoctoral researcher specializing in mammalian diversity and phylogenetic comparative methods. 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. Luis’s combination of research experience, R programming expertise, and prior work on LLM documentation makes him a valuable addition to this project team.

Contributors

- Mauro Lepore (reviewed initial proposal)
- Christoph Scheuch (reviewed initial proposal and may become an active contributor at a later point)

Consulted

- Simon Couch, Posit (reviewed initial idea though not completed proposal due to timing constraints).
- Initial feedback from Yanina Bellini Saibene
- Feedback from Maëlle Salmon, ROpenSci

The Problem

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

While existing resources provide valuable solution-oriented guidance on topics, such as how to use specific functions and packages, R users are left without a clear framework for understanding the problem space. This involves decisions about whether and when to use LLMs; for example, in using them to explain results to stakeholders, embedding code in analysis pipelines while maintaining data privacy, and generating R code.

As a result, users must piece together knowledge from multiple sources, with no obvious place to start. There is real risk of 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. This would give R users confidence in having a reliable foundation for experimenting with LLMs and strengthen R’s position as a community invested in open, inclusive and responsible use of technology.

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.

Due to the speed at which things are moving in LLMs/AI, 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. Instead, the value lies in creating a living, community-oriented resource that can be updated as packages evolve, new methods emerge, and best practices shift. By structuring it as an open, ongoing work rather than a fixed text, the book can remain relevant to R users and avoid the obsolescence that traditional publications in this domain are likely to face.

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

My proposal is to write a free online book on LLMs in R. 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".

It will also discuss the ethics, for example, debated 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

The project represents a critical social infrastructure investment in the R community's capacity to effectively engage with AI tech. Like translation efforts, educational initiatives, and documentation frameworks previously funded by the ISC, this book will establish foundational knowledge systems that enable broader community participation. The project goes beyond content creation to build reusable community infrastructure: standardised contribution processes, editorial governance models, and sustainable maintenance frameworks that can be adapted by other R community projects. By consolidating fragmented knowledge into a structured, community-owned resource, this project creates intellectual infrastructure that strengthens R's position in the evolving AI landscape while establishing reusable governance and contribution frameworks for future community projects.

Minimum Viable Product

An initial MVP would include the book repository, the chapter outlines with headings and subheading, and collated resources for each chapter.

As the approach to creating each chapter would be to create the aforementioned outline for each before adding additional context, it should be trivial to deliver such an MVP with minimal effort.

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, accessibility validation, 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
- Automated content validation and consistency checking systems

Community Infrastructure Deliverables

Beyond the book content, this project will deliver reusable community infrastructure: - Editorial board governance model and processes - Contribution guidelines and templates for collaborative technical documentation that other R community projects can adopt for similar documentation efforts - Automated publishing and maintenance workflows - Community engagement and feedback collection systems

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. However, I plan to mitigate this risk 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

GitHub Actions to build the book. Various R packages for working with LLMs in R.

Project plan

Start-up phase

Administrative tasks

I will complete the following administrative tasks:

- Creating the project repository on GitHub with appropriate license

- Setting up the project as a Quarto project which uses GitHub actions to publish the book to GitHub pages
- Purchasing URL for the project which will point to the published book

Content planning

In the start-up phase I 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. I will use the outcome of this stage to refine the list of high-level topics.

Technical delivery

Dec 31 2025 - Milestone 1: Repo created, contribution guidelines created, scope and outline finalised based on community survey

Jan 31 2026 - Milestone 2: Introductory chapters published and initial publicity sought

Feb 28 2026 - Milestone 3: Major content release 1

Mar 31 2026 - Milestone 4: Major content release 2

April 30 2026 - Milestone 5: Full first draft release

May 30 2026 - Milestone 6: Editorial board established and transition to community governance

Other aspects

The intention is to release the contents under a "Creative Commons Zero v1.0 Universal" license. The project is to be stored on GitHub on a public repository.

I will publicise the work by posting about it on LinkedIn, Mastodon, and Bluesky upon realisation of each milestone, and actively encouraging community feedback. I will also write content to be shared quarterly on the R Consortium blog to ensure wide reach, as well as posting on my personal blog and submitting the links to R Weekly.

While I intend to develop the initial contents myself to simplify the process of getting started, once an initial draft is underway and a style guide and contribution format template has been established, I'll be opening it up to community contributions. Clear templates and guidance will be established, and I'll solicit input from members of the community on how to best govern this.

I also intend to submit a talk to major R conferences like UseR! 2026 on the process of establishing the book and growing a community around it. I am open to feedback on other relevant avenues too.

I also would like to note that LLMs will not be used to write the contents this book - human understanding and reasoning is key to writing materials which are genuinely useful to others and maintaining a consistent tone. Any use of LLMs will solely be for ensuring conformance to defined principles (e.g. lack of use of jargon, readability) and checking for grammar/spelling.

Budget & funding plan

I would like to request financial support of 24080 USD from the ISC to develop the book materials, and transition the project to a community-owned resource. The labour costs mentioned below use an hourly rate of $100, based on proposals for previous R Consortium sponsored projects.

Milestone 1: Initial Setup

Target date: Dec 31 2025

Three-year domain costs: $80 Labour costs: $1,500

Work and outcomes:

- GitHub repository created with open license (CC0), Quarto structure, and contribution templates.
- Domain registered and pointed to GitHub Pages.
- Community survey completed and results analysed.
- Initial scope and outline document published.
- Draft contribution guidelines and templates prepared (not live yet).
- Initial MVP delivery

Milestone 2: Introductory chapters published and initial publicity sought
Target date: Jan 31 2026
Labour costs: $3,500
Work and outcomes:

- At least two foundational chapters fully drafted and published.
- Project site launched publicly with domain.
- Publicity campaign run (socials + R Consortium blog post).
- Invitation for feedback and expressions of interest in joining an editorial board.

Milestone 3: Major content release 1
Target date: Feb 28 2026
Labour costs: $5,500
Work and outcomes:

- 2-3 substantial chapters added
- Editorial board established (3–5 members).
- Contribution guidelines finalised and published.
- Repository officially open for community PRs.
- At least 1–2 external contributions merged by milestone date.

Milestone 4: Major content release 2
Target date: Mar 31 2026
Labour costs: $5,500
Work and outcomes:

- 2-3 additional chapters drafted.
- First wave of community contributions reviewed and integrated.
- Second publicity push with highlights of community involvement.

Milestone 5: Full first draft release
Target date: April 30 2026
Labour costs: $5,000
Work and outcomes:

- 1-2 additional chapters drafted.
- Complete version 1 of all planned chapters online.
- Ongoing updates based on community feedback.

Milestone 6: Editorial board established and transition to community governance

Target date: May 30 2026

Labour costs: $3,000

Work and outcomes:

- Ongoing updates based on community feedback.
- Consistency edits across chapters (tone, formatting, terminology).

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.
- At least 10 chapters published.
- Working CI/CD pipeline.
- An editorial board in place and contribution guidelines published.
- A governance model and roadmap documented, ensuring the project can continue after the initial period without further direct funding.

Measuring success

Success could be measured in the following key areas:

1. Milestone completion: Deliverables in the six-milestone plan achieved on schedule.
2. Content published: At least 10 complete chapters live online by April 2026.
3. Community engagement: Minimum of 100 survey responses; 3 community PRs merged by project end.
4. Governance established: Editorial board of at least 3 members active by May 2026, with review process tested.
5. Reach/visibility: Website traffic metrics and engagement on announcement posts; two R Consortium blog posts published and 1 conference submission made.

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.
- Maintaining an open contribution model so the community can update, refine, and extend content long term.