2025-09-21

LLMs in R Book

Nic Crane

Executive Summary

This proposal requests $16,000 USD to bootstrap a free, community-owned online book on LLMs in R. R users face a fast-moving ecosystem where reliable guidance is scattered across package docs, package vignettes, and blog posts. The book will consolidate best practices, examples, and explanations into a single structured, accessible resource.

Over six months (Dec 2025 - 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 with reproducible code.
- 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, 50 survey responses, 3 merged community contributions, and a functioning editorial board.

The outcome will be a living, sustainable resource that strengthens R's role in modern AI and equips analysts, data scientists, and educators 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 complete this project.

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

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 like how to use specific functions, 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. The problem exists for 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, examples, and guidance around tooling into a single structured resource. This would give R users a reliable foundation for experimenting with LLMs and strengthen R's position as a platform for responsible, reproducible AI. By focusing on framing the problem first, there is a clear path to connect readers to tools and workflows relevant to their work.

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 navigating a fast-changing landscape while avoiding the obsolescence that traditional publications in this domain are likely to face.

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 necessary, and 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.

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

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.

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 specific enough to provide value, but high-level enough to allow for flexibility. For example, while available underlying models and their capabilities are likely to see many changes over time, methods of choosing an appropriate model to work with are less likely to see as much change.

External dependencies

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

Project plan

Start-up phase

Admininstrative 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.

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.

Budget & funding plan

I would like to request financial support of 16080 USD from the ISC to develop the book materials, and transition the project to a community-owned resource.

For full transparency, I am about to begin a project involving writing a course on machine learning and AI for modellers in R. As part of my preparation for this work, I am reviewing resources on LLMs in R, some of which will form the contents for the MVP delivery. The budget and time estimates below relate only to work done above and beyond this initial preparatory work which will be completed regardless.

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: $2,500

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: $3,500

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: $3,500 - 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: $3,000

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: $2,000

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 substantive chapters covering introductory through intermediate use cases.
- 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: Each deliverable 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; at least one R Consortium blog post 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 key chapters 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.