Ollama, No Drama: Step-by-Step Guide of Practical Local AI

Dimitar Hristov
Department of Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences
dimitar@dcl.bas.bg

Abstract

The proposed tutorial introduces participants to Ollama, an open-source tool for running large language models locally. It covers installation, model selection, command-line interaction, and practical integrations, giving attendees the skills to use local LLMs for experimentation and small-scale research with minimal setup.

Keywords: Ollama, large language models, on-device AI, hands-on tutorial

1 Topic and goals

This tutorial introduces participants to Ollama, a lightweight and user-friendly framework for running large language models locally. The session covers installation, model selection, loading and switching between models, running prompts and chats directly in the terminal, and exploring simple integrations — all without requiring programming experience. The tutorial emphasizes hands-on interaction, transparency, and demystification of local AI workflows.

The primary goal is to give beginners—including first-year computer science and engineering students and participants with non-engineering backgrounds—the knowledge and confidence to use local LLMs for learning, experimentation, and small-scale research tasks. By the end of the tutorial, attendees will be able to:

- Install and configure Ollama on their own machines
- Understand the differences between available open models
- Run prompts and interactive chats in the terminal
- Configure models for specific tasks and experiments
- Use local LLMs for practical tasks such as summarization, brainstorming, and text exploration
- Connect Ollama to external tools or interfaces
- Develop intuition about how local AI fits into broader NLP and research workflows

The session is designed intentionally code-light, focusing on conceptual clarity and practical utility rather than software engineering.

2 Relevance

Local LLMs are becoming increasingly important in computational linguistics and NLP research due to their accessibility, privacy benefits, reproducibility, and low barrier to entry. For students and researchers who may not have access to cloud resources or large-scale infrastructure, tools like Ollama provide a practical way to explore model behaviour, experiment with prompts, and prototype ideas.

3 Tutorial type and target audience

The proposed tutorial is introductory, aimed at students and young researchers and professionals of various backgrounds, without necessary prior experience in code-writing or command-line interfaces. Some experience with common AI chatbots and assistants is beneficial.

The attendees will be expected to bring their own computers for the hands-on part of the tutorial.

The tutorial is expecting 15-25 attendees.

4 Outline

The tutorial is planned for a 4-hour session, split in 2 main parts and a slot for additional questions and discussions.

- **Part A:** Foundations (2 hours)

1. Introduction – LLMs, chatbots, assistants and cloud services (10 mins)
2. The local alternative – Overview of Ollama (10 mins)
3. Installation and setup (20 mins)
4. The Ollama model library – choosing and preparing the right model for the task (20 mins)
5. Running and interacting with models on the command line (30 mins)
6. Common application integrations with Ollama (15 mins)
7. Questions on the theoretical part (15 mins)

- **Part B:** Practice (1.5 hours)
  1. Ollama installation and setup (30 mins)
  2. Model choice, setup and execution (30 mins)
  3. GUI installation, setup and use with Onyx (or alternatives) (30 mins)
- **Part C:** Additional questions and discussions (30 mins)

5 Relation to previous research

The tutorial is based on a technical report (IfGPT, 2025) produced from the research conducted for the project Infrastructure for Fine-tuning Pre-trained Large Language Models (IfGPT) and related work (Hristov, 2025). It distils the practical findings of this work into a hands-on format, extending its coverage with the various newly available integrations.

6 Covered software and related material

- Ollama[1]
  - Official documentation[2]
  - IfGPT Techical Report (in Bulgarian) (IfGPT, 2025)
- Integrations with:
  - Visual Studio Code (IDE)[3]
  - Onyx (Open Source AI Platform for Work)[4]
  - OpenClaw (Personal AI Assistant)[5]

[1] https://ollama.com/
[2] https://docs.ollama.com/
[3] https://code.visualstudio.com/
[4] https://docs.onyx.app/
[5] https://openclaw.ai/

7 Tutor background

Dimitar Hristov is a graduate of the Sofia High School of Mathematics, with a BSc from the University of Southampton and an MSc from Sofia University. He is currently a PhD student at the Department of Computational Linguistics at the Bulgarian Academy of Sciences, focusing on the development, fine-tuning, and optimisation of large language models, including their specialisation for on-device execution. He has long collaborated with the Department on research in NLP, corpus development, and WordNet resources for English and Bulgarian. Dimitar also has experience leading educational sessions in both non-formal and professional settings, including workshops in student organisations and technical tutorials for apprentices in cleversoft Bulgaria's training programme.

8 Teaching material

All tutorial attendees will receive access to the presentation slides, example configurations, and demonstration scripts used during the session. Supplementary materials such as code snippets, workflow diagrams, and reference documentation will also be provided to support independent practice. If available, video recordings of the tutorial will be shared with registered participants after the event.

The core tutorial materials—including slides, example configurations, and supplementary documentation—can be made publicly available on the CLIB 2026 website. Any components that cannot be shared due to licensing or project-specific restrictions will be clearly indicated, but the main instructional content will be fully accessible.

References

Dimitar Hristov. 2025. Large language models for lexical resource enhancement: Multiple hypernymy resolution in wordnet. In *Proceedings of the 9th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing*, pages 20–26, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

IfGPT. Infrastructure for fine-tuning pre-trained large language models. https://ifgpt.dcl.bas.bg/en. Institute for Bulgarian Language, project BG-RRP-2.017-0030-C01, accessed 2026-02-15.

IfGPT. 2025. Ollama. Technical Report Version 1, Institute for Bulgarian Language. Accessed 2026-02-15.