Latent Worlds

How Language Models Build Their Internal Maps

Dr. Yusuf Ahsan

A Kelford Press Original

First published in 2026 by Kelford Press

ISBN 978-1-7398-2201-5 (Digital) ISBN 978-1-7398-2202-2 (Print) ISBN 978-1-7398-2203-9 (Audio)

kelfordpress.com

For the researchers who keep asking what the model is doing, and refuse to accept "it just works" as an answer.

Contents

Interpretability research has matured from a curiosity into a discipline. After years of treating large language models as inscrutable black boxes, a generation of

researchers has begun to map the internal representations that emerge when models are trained on next-token prediction at scale. This book is a field guide to what we have learned, what remains contested, and where the line still falls between mechanistic insight and clever-sounding statistical pattern-matching.