A World Leading SFI Research Centre

Science Foundation Ireland sfi For what's next

OKLLM - Online Knowledge Search for LLM innovations

Huan Chen, Andy Donald

Insight

SFI RESEARCH CENTRE FOR DATA ANALYTICS

HOST INSTITUTIONS

DCU Dublin City University

Ollscoil na Gaillimhe University of Galway

UCC University College Cork, Ireland

UCD Dublin

PARTNER INSTITUTIONS

Maynooth University

Tyndall National Institute

Trinity College Dublin

University of Limerick

FUNDED BY:

European Union European Regional Development Fund

Science Foundation Ireland sfi For what's next

Introductions

Huan Chen: Research Associate @ Insight Centre for Data Analytics, University of Galway, Ireland

Andy Donald: Research Fellow & Unit lead @ Insight Centre for Data Analytics, University of Galway, Ireland

Paul Buitelaar: Professor in Data Analytics @ Insight Centre for Data Analytics, University of Galway, Ireland

Insight Overview

DCU Ollscoil Chathair Bhaile Átha Cliath Dublin City University

Ollscoil na Gaillimhe University of Galway

UCD Dublin

UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh

Maynooth University National University of Ireland Maynooth

Tyndall National Institute

Trinity College Dublin Coláiste na Tríonóide, Baile Átha Cliath The University of Dublin

University of Limerick Ollscoil Luimnigh

ATU

MTU Ollscoil Teicneolaíochta na Mumhan Munster Technological University

DUBLIN Ollscoil Teicneolaíochta Bhaile Átha Cliath Technological University Dublin

TUS Midlands Midwest

Dundalk Institute of Technology

4
Co-Lead Universities
9 partner institutions

Built on **20** years of research in Data Analytics and AI

450+
Academics, Postdocs, PhDs, RAs

3400+
Scientific conference and journal papers

175+
Funded collaborations with industry partners

350+
Research Awards

16
Spin out companies
72 license agreements

135+
H2020 consortia, 500+ collaborations, 40+ countries

1,137+ school visits, 28,000 students

276
PhDs graduated

Insight

OKLLM - An Introduction

LLMs are extremely powerful and are revolutionising the search and discovery process in their ability to deliver more precise, individualized and context-aware results. However, they currently have numerous issues such as:

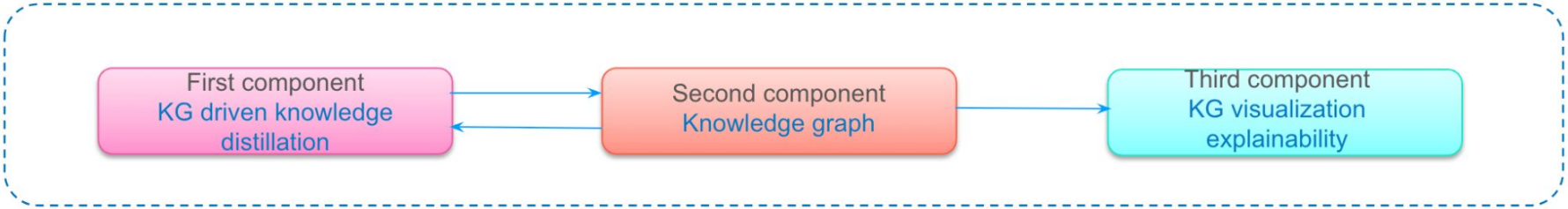
- Intense computational needs
- Bias detection and remediation
- Hallucination
- Explainability of results

OKLLM - Overview

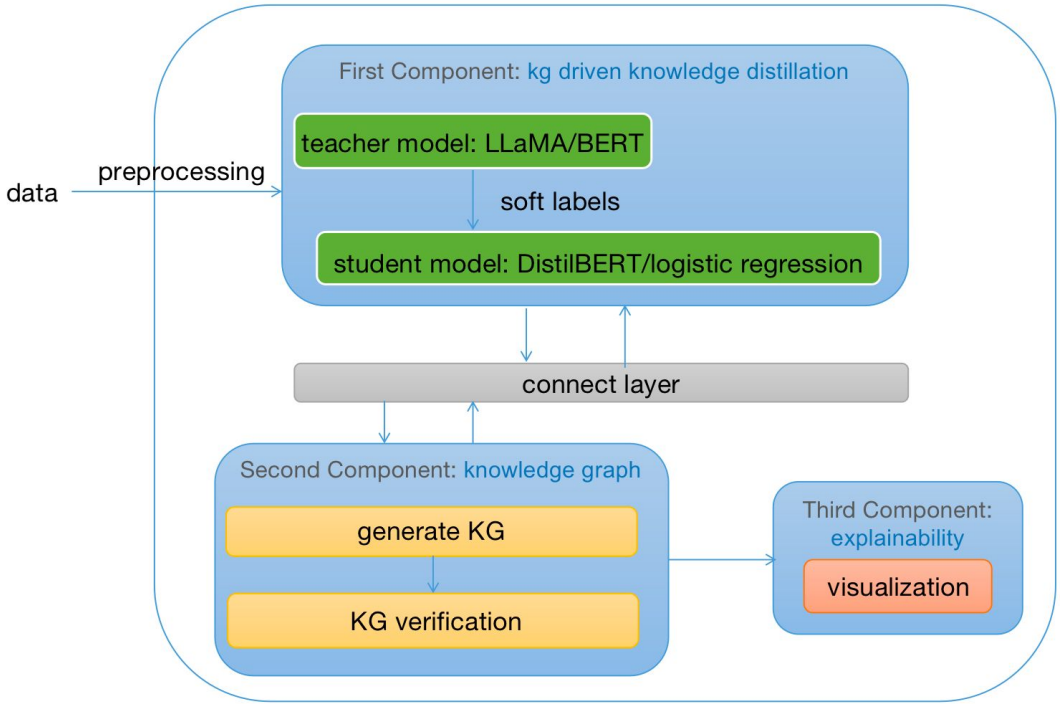
OKLLM is proposed to be developed across three distinct components, each of which will look to solve specific gaps mentioned previously. These components are:

- Knowledge distillation - Handles bias detection and computational issues through knowledge transfer
- Knowledge graph automated generation & verification
- Result explainability

OKLLM - Pipeline



OKLLM Architecture



OKLLM Datasets

- Search Engine Result Pages (SERP) Datasets: The term "SERP Data" refers to information and data gathered from search engine results pages (SERPs), which may contain details about a website's position in search results, the amount of searches for particular keywords, and other search engine optimization (SEO) related metrics.

- MS MARCO Passage Ranking Dataset: A popular dataset in the fields of natural language processing (NLP) and information retrieval is the Microsoft MAchine Reading Comprehension (MS MARCO) dataset. The purpose of the MS MARCO Passage Ranking dataset is to rank passages (short text excerpts) in response to a query.

Related Projects

Saffron
Knowledge Extraction Framework

- knowledge extraction from text
- term extraction
- taxonomy extraction
- knowledge graph generation
- https://saffron.insight-centre.org/

Customer Interaction Data
Bias Detection Research

- Donald, A., Galanopoulos, A., Curry, E., Muñoz, E., Ullah, I., Waskow, M.A., Dabrowski, M. and Kalra, M., 2023. Bias Detection for Customer Interaction Data: A Survey on Datasets, Methods, and Tools. IEEE Access. [3]

PEERS
Knowledge-base Website

- PracticE Ecosystem for standaRdS(PEERs) EU Horizon project
- generation of a knowledge dashboard allowing analysis of project source documents and data
- https://peers.universityofgalway.ie/

Development & Component Release

- September 2024 - Initial draft components (GitHub)
- February 2025 - First complete delivery (GitHub)
  - Politics domain focus

Thank You

Questions?