Most African NLP datasets **lack clear copyright and licensing information**, creating **governance risks** that could undermine **responsible AI** development

Our analysis of 249 African NLP projects revealed significant gaps in copyright transparency, licensing clarity, and dataset governance, highlighting the need for stronger documentation standards and evidence-based AI governance frameworks.

Copyright Compliance and Licensing Patterns in African NLP Datasets

A Data-Driven Governance Analysis of 249 NLP Projects

INTRO

- African NLP research is growing rapidly, with datasets increasingly used to train language technologies and AI systems.
- However, copyright status, licensing terms, and legal permissions are often **unclear** or poorly documented, creating legal and governance **risks**.
- Existing studies largely describe these challenges, but few provide systematic, data-driven evidence of governance **patterns** across African NLP projects.
- This study analysed 249 African NLP projects and 24 governance variables to identify copyright risks, licensing practices, and factors associated with stronger governance.
- The findings provide evidence to support responsible AI development, improved dataset stewardship, and policy-informed governance frameworks.

METHODS

Dataset and Research Design
- Project scoping and research questions
- 249 African NLP project Dataset
- Multivariable

Data preparation
- Data cleaning and Data pre-processing
- Feature Engineering and Variable Selection

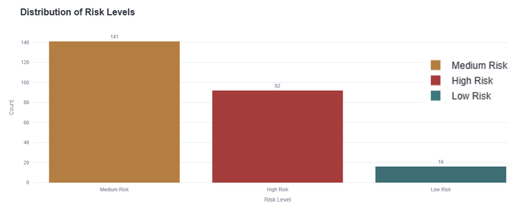
Exploratory Data Analysis
- Frequency distributions, pattern observations
- Problem variables found
- Cross tabulations

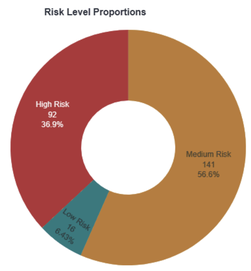
Modelling and Statistical Analysis
- Copyright risk score modelling
- Variable Relationship analysis
  - PCA, MCA, K-means clustering
- Governance and risk patterns

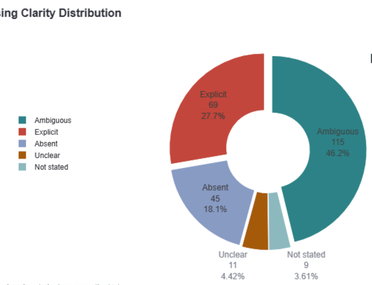
Visualisation and Deployment
- Interactive Copyright Compliance & Licensing Pattern Dashboard
- Risk diagnostics
- Licensing analytics

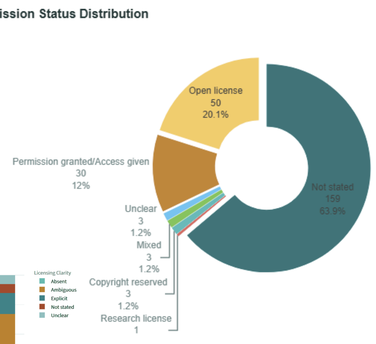
Delwyn Gordon, **Nokhutula** Madekufamba

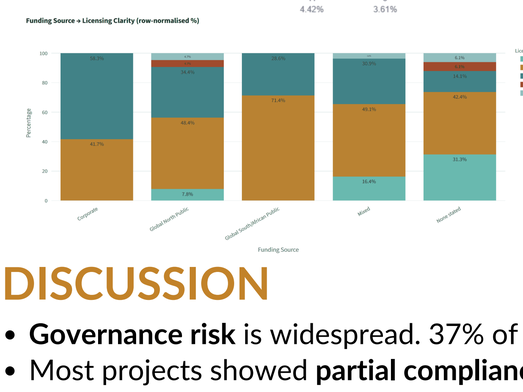
RESULTS











DISCUSSION

- **Governance risk** is widespread. 37% of projects are high risk.
- Most projects showed **partial compliance**, with the majority classified as medium risk, indicating **inconsistent governance** and licensing practices. Unclear permissions, ambiguous licensing, and missing legal justification frequently occurred together.
- **Explicit licensing and permissions** were the strongest indicators of low risk.
- **Commercialisation orientation** and **funding** source were the strongest predictors of licensing transparency, while authorship geography had limited influence.
- Stronger governance supports **responsible AI** adoption. Transparent licensing and dataset stewardship improve trust, reuse, compliance, and the sustainability of African AI ecosystems.
- **AI policy** should prioritise licensing transparency, copyright documentation, and governance standards to support responsible and legally sustainable AI development in Africa.

DASHBOARD

Department of Computer Science

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Capstone Project - MIT 808

Course Coordinators:
Dr. Vukosi Marivate (vukosi.marivate@cs.up.ac.za)
Abiodun Modupe (abiodun.modupe@cs.up.ac.za)

Scan me