# Evidence Synthesis Ireland Fellowship Scheme
# Project Identification Form

**Mentor and Host Centre**

| |
|---|
| Ailbhe Finnerty Mutlu; EPPI Centre, UCL |

**Project title**

| |
|---|
| Automated Risk of Bias Assessments using LLMs: Development and Evaluation with DEET |

**Project type and methods**

| |
|---|
| Applied Methodology / AI-assisted evidence synthesis |

**Project details**

Risk of bias (RoB) assessment is a cornerstone of evidence synthesis, and it is one of the most resource-intensive and subjective stages of systematic review. Large language models (LLMs) have demonstrated early promise for RoB assessment, with studies suggesting feasibility for established RoB instruments. However, rigorous, structured evaluation of AI-generated RoB assessments, particularly across study designs, remains limited and inconsistently reported.

We have developed the data extraction evaluation toolkit (DEET) to support reproducible evaluation of AI-assisted data extraction. By extending DEET to RoB, this project will create a reusable and generalisable automation approach for RoB assessment including performance evaluation.

**Aim**

- Develop and refine LLM-based workflows to automate RoB assessment using the Cochrane RoB 2 for RCTs
- Extend DEET to support structured evaluation of AI-generated RoB outputs against human gold-standard assessments

- Evaluate performance at signalling question, domain, and overall RoB judgement levels; quantifying accuracy, agreement, and failure modes
- Quantify workload reduction and identify the conditions under which AI-assisted RoB is reliable enough for supervised or semi-automated use
- Produce guidance on human-in-the-loop checkpoints and performance thresholds for deployment of automated RoB within evidence synthesis workflows

**Outputs**

- Extended DEET module supporting automated RoB evaluation (open-source, documented, reusable)
- Practical guidance document for evidence synthesis teams on AI-assisted RoB assessment including when human oversight is essential
- Peer-reviewed publication(s) reporting development methods, performance benchmarks, and comparative analysis across RoB tools

**Review current status:**

We have several review datasets from the GALENOS project https://www.galenos.org.uk/about that have conducted RoB assessments of RCTs and can be used as a gold standard reference for development and evaluation.

**Any specific/desirable requirements for Fellow**

**Desirable**

- Experience conducting RoB assessments for a range of study designs
- Familiarity with LLMs and prompt engineering
- Experience with or willingness to learn Python

**Estimated start and completion dates**

October 2026 – July 2027