



WARWICK INSTITUTE *for*
EMPLOYMENT RESEARCH



**Trialling an Evaluation Framework for LLM-Powered
Careers Advice
Inception Report - Final**

A report prepared

By

**Peter Dickinson, Chris Percy, Deirdre Hughes,
Gianni Anelli and Vincent Ramos**

September 2026

**Warwick Institute for Employment Research
University of Warwick**

Contact details:

Peter Dickinson, Associate Professor
Warwick Institute for Employment Research
University of Warwick
Coventry CV4 7AL

+44 (0)24 76 524420
p.dickinson@warwick.ac.uk
<https://warwick.ac.uk/fac/soc>



**UNIVERSITY
OF WARWICK**

Acknowledgements

The Nuffield Foundation is an independent charitable trust with a mission to advance social well-being. It funds and undertakes rigorous research, encourages innovation and supports the use of sound evidence to inform social and economic policy, and improve people's lives. The Nuffield Foundation is the founder and co-funder of the Nuffield Council on Bioethics, the Ada Lovelace Institute and the Nuffield Family Justice Observatory. This project has been funded by the Nuffield Foundation, but the views expressed are those of the authors and not necessarily the Foundation. Find out more at: nuffieldfoundation.org.

Table of contents

1	Introduction.....	1
2	Project overview and motivation	1
2.1	Main aims, objectives and process	1
2.2	Goal and audience of the Evaluation Framework.....	4
3	Proposed evaluation methods.....	7
3.1	Method 1 overview: Panel review	8
3.2	Method 2 overview: Case Study Comparison.....	8
3.3	Method 3 overview: Randomised Control Trial (RCT)	9
3.4	Objectives to be evaluated.....	9
4	Proposed Evaluation Framework.....	10
4.1	Evaluation Framework Document Outline	10
5	Pilot Evaluation Scoping	12
5.1	Method 1 pilot scope: Panel review (see Section 2 in the complementary Methodology paper) 12	
5.2	Method 2 pilot scope: Case Study Comparison (see Section 3 in the complementary Methodology paper)	12
5.3	Method 3 pilot scope: Randomised Control Trial (see Section 4 in the complementary Methodology paper)	12
5.4	Managing ethical risks to student participants in the field trial.....	13
5.5	Technology selection: Long-list and selection criteria	14
5.6	Managing conflicts of interest.....	21
	Appendix 1: Advisory Group Terms of Reference and members	22
	1. Purpose and Role of the Advisory Group	22
	2. Terms of Reference	22
	3. Membership	23
	4. Meetings and Commitment	23
	5. Confidentiality.....	23
	6. Travel and Expenses	24
	7. Duration.....	24
	8. Review.....	24
	Advisory Group Members	24
	Appendix 2: Project timeline overview	25
	Appendix 3: Risk Management	26

1 Introduction

This is the Inception Report for the Nuffield Foundation funded project Trialling an Evaluation Framework for LLM-Powered Careers Advice. Versions of its developing contents have been reviewed at various points at the three Advisory Group meetings on the 15th April, 9th June and 21st July 2026.

The purpose of the Inception Report is to describe the project and its key deliverables (an Evaluation Framework and Pilot Evaluation Results), as well the project plan and ways of working.

2 Project overview and motivation

2.1 Main aims, objectives and process

This project will develop and pilot an Evaluation Framework to evaluate LLM-powered careers advice tools, where LLM refers to Large Language Models, such as OpenAI's GPT models, Anthropic's Claude models, and Google's Gemini models, among others.

Careers advice is a critical component of UK education and work transitions, particularly for disadvantaged groups who face uneven support. AI has the potential to enhance access, quality, and sustainability – but only if due care is taken to ensure that such tools are only recommended in appropriate circumstances and are quality assured.

Research carried out in 2024 showed that whilst 33% of students are already using LLMs for personalised guidance, only 5% of careers advisers are aware of this usage.¹ More recent evidence confirms that this divergence has widened sharply. By early 2025, the Prospects Luminate Early Careers Survey found that nearly one in five respondents had used generative AI tools for careers advice, rating them overwhelmingly as helpful; and 39% had used AI to edit or write job application materials.² HEPI's Student Generative AI Survey 2025 found that overall student AI usage had risen from 66% to 92% in a single year.³

Yet institutional awareness and practitioner development have not kept pace with changing student habits: Jisc found that fewer than half of FE institutions had delivered any staff development on AI by 2025,⁴ while AGCAS careers leaders identified the need for high-quality AI guidance as one of their most pressing priorities (Toogood, 2025).⁵ This pattern of unsupported, unmonitored student uptake is particularly concerning given persistent equity

¹ Hughes, D., Percy, C., & Tolond, C. (2024). LLMs for HE careers provision. Bristol, UK: Jisc. <https://luminate.prospects.ac.uk/large-language-models-for-careers-provision-in-higher-education>

² Smith, M-S., Greaves, L. and Mason, D. (2025). Early Careers Survey 2025. Jisc 2025. <https://luminate.prospects.ac.uk/early-careers-survey>,

³ Freeman, J. (2025). Student generative AI survey 2025 (HEPI Policy Note 61). Higher Education Policy Institute; Kortext. <https://www.hepi.ac.uk/reports/student-generative-ai-survey-2025/>

⁴ Jisc (September 2025). Staff perceptions of AI 2025. National Centre for AI. <https://nationalcentreforai.jiscinvolve.org/wp/2025/09/02/staff-perceptions-of-ai-2025/>

⁵ Toogood, C. (2025). *What are the priorities for higher education careers and employability in 2025?* AGCAS/Prospects, Luminate <https://luminate.prospects.ac.uk/what-are-the-priorities-for-higher-education-careers-and-employability-in-2025>

gaps: an estimated 9.7 million people in the UK currently lack access to meaningful careers guidance,⁶ and young people from lower socio-economic backgrounds are less likely to have access to human careers support, leaving them more reliant on AI and digital tools to fill the gap.⁷ Usage in all groups has likely increased sharply since 2024/25, but a lack of understanding regarding appreciation of the sophisticated usage of (even reliance on) AI tools is likely to have continued.

Policy makers, service managers, and practitioners have no high-quality evidence on the performance of generic and career-specific tools despite a significant number of students using LLMs for personalised career guidance. The careers sector is not currently equipped to manage this trend, with low awareness and mixed technical expertise among advisers. Without evidence-based assessments, there is a risk of AI tools being adopted based on cost or convenience rather than efficacy.

The study is (1) undertaking an evaluation of LLM-powered AI tools (the first of its kind) but also (2) developing a framework to inform and support future evaluations.

The research team aim to address this gap by creating an evidence base to guide responsible use of LLM-powered AI tools, ensuring they enhance accessibility, quality, and outcomes in career advice. The following questions will be answered:

- What are the nature and characteristics of an Evaluation Framework that the careers sector would accept for assessing careers advice powered by LLMs?
- Within the Evaluation Framework, what is the appropriate balance between evaluation methods (e.g. panel reviews, case study comparisons, and field trials)? What thresholds (e.g. levels of accuracy, usefulness, appropriateness etc.) are appropriate to inform practical guidelines and specific policies on tool usage?
- Based on Pilot Evaluation studies of a specific set of AI-powered tools (selected according to criteria such as relevance, appropriateness, tone accuracy, and actionability), what are the operational requirements for future evaluations?
- Based on the Pilot Evaluation studies, what hypotheses might be formed on the degree of relative performance between different types of careers interventions, and between different subgroups?

The research is being completed in four phases:

- Form an advisory group of 15-25 stakeholders to co-design the Evaluation Framework and pilot evaluation plan.
- Co-develop the Evaluation Framework and evaluation materials, including structured procedures for panel reviews, case studies, and field trials.

⁶ Phillips, A. (2024). *Open door policy: Why the new government should introduce an employment advice guarantee*. Demos, p.6. https://demos.co.uk/wp-content/uploads/2024/07/Open-Door-Policy_2024_July.pdf in Hughes, D., & Percy, C. (2024). *Careers 2035: What does it mean for the future of AI and career guidance?* p.7. Morrisby. <https://www.morrisby.com/careers-ai>.

⁷ BDO (March 2026). Young people still face unequal access to careers advice and work experience. <https://www.bdo.co.uk/en-gb/news/2026/young-people-still-face-unequal-access-to-careers-advice-and-work-experience>

- Implement the Pilot Evaluations, including a field trial with up to 160 students in FE College settings aged 16-19 who have varying levels of certainty about their career choices.
- Refine the evaluation framework, publish findings, and share outputs through open-access platforms, stakeholder surveys, and academic publications.

For the purposes of this project, LLM-powered careers advice is provisionally defined as the use of artificial intelligence (AI) systems built on Large Language Models (LLMs) trained on extensive bodies of text data and other media to generate personalised, conversational responses/outputs for careers-related queries, in ways that go beyond information retrieval to engage users in contextually sensitive, open-ended dialogue about their career development needs.

Such systems may provide labour market information, explore occupational options, support self-reflection, assist with application materials, and signpost users to further resources. They are capable of operating at scale, on demand, and without a human adviser present at the point of interaction, but the quality and limitations of such usage remain poorly understood at present.

LLM-powered careers advice (generic or dedicated) tools are constantly being updated and agreement needs to be made not only about which tools but also about the versions to be included in the study. In addition, how the tools are used can also be critical. For instance, whether questions are asked in voice- or text-mode can produce results of different quality.

Three distinctions are likely to be important as we develop the Evaluation Framework:

- **LLM-powered advice vs. earlier digital guidance tools.** Previous generations of careers technology – interest inventories, course search platforms, job matching algorithms - retrieved or ranked pre-existing content. LLMs generate novel, tailored responses in real time. This shifts both the opportunities (responsiveness, personalisation, scale) and the risks (hallucination, bias, lack of accountability) for generated content.
- **Curated vs. general-purpose LLM deployment.** A young person or adult, including careers advisers, using a general-purpose LLM such as ChatGPT is engaging with a fundamentally different proposition from one using a purpose-built, quality-assured careers tool with curated content, guardrails, and defined scope. Any evaluation framework must therefore be explicit about its deployment context: either it is designed for a specific context, or it is intended to apply across multiple contexts – and if the latter, it must articulate how it does so differentially. This distinction has direct implications for what counts as a meaningful quality standard, and for what can legitimately be compared with what.

The above considerations point collectively to the need for clarity on evaluation scope before any evaluation framework is applied. Three questions are foundational: What type of deployment is being evaluated – curated and purpose-built, or general-purpose? What quality metrics exist for the human advice against which LLM-generated guidance will be compared? And is the evaluation designed to assess implementation and process fidelity, outcomes and impact, or both? Without explicit answers to each, findings risk being methodologically incoherent – comparing unlike with unlike or drawing conclusions that are valid only within a narrower context than claimed.

- **Careers information vs. careers guidance.** In professional and policy terms, a distinction has long been drawn between careers information (factual content about occupations, routes, and opportunities), careers advice (more personalised steers based on an individual's situation), and careers guidance (a professionally mediated, ethically accountable process of supporting individuals to make informed decisions). LLMs can deliver the first (subject to an accuracy caveat); the extent to which they can deliver the second and third – and under what conditions – is precisely the question this project is designed to investigate. The working definition uses the term advice deliberately, as a neutral starting point, without prejudging that question. Because of this distinction, the focus of this study is on careers guidance and the more personalised-end of advice, rather than information provision or generic advice alone. We refer to the target end of the provision spectrum in this report as 'career guidance'.

2.2 Goal and audience of the Evaluation Framework

This Inception Report is the planning instrument for producing the Evaluation Framework in consultation with the Advisory Group and the project funder the Nuffield Foundation. The Evaluation Framework itself will function as methodological guidance – explaining how evaluation elements should be designed and conducted, without prescribing a single approach. Within the Evaluation Framework there will be an evaluation toolkit that will serve as an evaluation study-level document through which a researcher or evaluator records their specific design decisions before beginning.

The Evaluation Framework is primarily a technical document for technical users who might request, fund, or conduct evaluations of AI tools to support the careers sectors, e.g. operational leads in sector funders, policy organisations, large scale providers with research or evaluation priorities, and evaluation organisations. It should be able to be used across the UK, respecting the different careers systems in the Devolved Authorities.

The Evaluation Framework defines how evaluations are to be carried out, what information is to be gathered, and how that information can be used to make decisions about appropriate tool usage in particular learner contexts (including risk mitigation). It provides guidelines, worked examples, and template documents to support these exercises, setting out sector-supported minimum and recommended standards to support actionable conclusions from the evaluations.

The Evaluation Framework should include explicit guidance on acceptable use of LLM-powered tools in careers contexts. This should articulate where such tools demonstrably add value (for example, in broadening access, supporting self-exploration, or handling high-volume initial enquiries), and where they cannot substitute for the professional knowledge, relational skills, and ethical judgement of a qualified careers adviser. Criteria, indicators, and standards decisions should be defined in light of this scope. The Evaluation Framework will identify ways to assess:

1. Quality of careers guidance, this may include:
 - Accuracy of labour market information (LMI) – is the information factually correct, current, and appropriately localised (e.g. local/regional job market data)?
 - Relevance and personalisation – does the advice reflect the individual's stated circumstances, aspirations, qualifications, and constraints?

- Breadth and balance – does it present a range of pathways rather than defaulting to stereotyped or narrow options (e.g. avoiding occupational gender bias)?
 - Depth of support – does it go beyond information-giving to support reflection, decision-making, and next steps?
 - Alignment with professional standards – does the advice conform to recognised quality frameworks (e.g. CDI Code of Ethics, QCG standards etc), Gatsby Benchmarks,?
2. Safety and Harm Avoidance, this may include:
- Absence of harmful advice – does the system avoid directing users towards unsuitable, risky, or exploitative pathways?
 - Appropriate escalation – does it recognise when a user's needs exceed what an AI tool can safely address (e.g. mental health, complex SEND needs) and signpost accordingly?
 - Bias and equity – does it produce equitable outputs across gender, ethnicity, disability, socioeconomic background, and geography? (Requires disaggregated testing)
3. User Experience and Accessibility:
- Usability - is the interface accessible and navigable for the intended user group (young people, adults, those with low digital literacy)?
 - Comprehensibility – is the language appropriate to the user's age and reading level?
 - User satisfaction – do users feel heard, supported, and helped?
 - Trust and transparency – do users understand they are interacting with an AI, and is uncertainty or limitation acknowledged honestly?
4. Adviser and Professional Confidence:
- Practitioner trust – do qualified careers professionals regard the outputs as safe to use alongside or in place of initial guidance?
 - Complementarity – does the tool support rather than undermine the role of the human adviser?
 - Referral appropriateness – does the system correctly identify cases that require human intervention?
5. User Experience and Accessibility:
- Usability – is the interface accessible and navigable for the intended user group (young people, adults, those with low digital literacy)?
 - Comprehensibility – is the language appropriate to the user's age and reading level?
 - User satisfaction – do users feel heard, supported, and helped?
 - Trust and transparency – do users understand they are interacting with an AI, and is uncertainty or limitation acknowledged honestly?

6. Longer-Term Outcomes (where measurable):

- Decision quality – do users report making better-informed decisions as a result of the interaction?
- Progression – do users take positive next steps (enrolment, application, training) following use?
- Reduced NEET risk – over time, does use of the tool correlate with reduced disengagement for target populations?
- Equity of access – does the tool reach and benefit disadvantaged groups who would otherwise receive no guidance? Do users (disadvantaged and non-disadvantaged) continue to make use of professional human advice when needed.

The Framework also needs to explore what might constitute success for an AI tool following review under the Evaluation Framework. Success criteria would need to be defined at multiple levels:

- **Minimum threshold (safety floor):** e.g. The tool produces no harmful, biased, or factually dangerous advice; it correctly identifies and escalates high-risk interactions; it is transparent about its AI nature.
- **Functional success:** e.g. Outputs are judged accurate and appropriate; user comprehension and satisfaction scores meet pre-agreed benchmarks; the tool performs equitably across demographic groups in structured testing.
- **Comparative success:** e.g. The tool performs at least as well as, or demonstrably complements, a comparator (e.g. a standard online information resource, or a brief initial adviser appointment) on agreed quality dimensions.
- **Systemic success:** e.g. The tool increases access to career guidance for underserved groups without displacing professional capacity from those with greatest need; it supports rather than erodes professional standards in the sector.

An Evaluation Framework for LLM-powered careers guidance sets out, in advance, what the tool is expected to achieve, how achievement will be measured, and what level of performance constitutes an acceptable standard. It combines outcome measures – the results the tool should produce for users – with process measures, which assess how those results are reached. Outcome measures will include the accuracy and relevance of the information and advice given, its alignment with professional standards, its safety and equity across different user groups, and the extent to which it supports positive career decisions and progression. Process measures cover transparency, accessibility, appropriate escalation to human support, and the absence of bias. Success is currently defined as meeting pre-agreed thresholds on each dimension – thresholds that should be set with input from careers professionals, users, and researchers before evaluation begins, rather than after results are available.

In conjunction with its primary audience of technical users, the Evaluation Framework's goal is to enable the broad sector to generate high-quality evaluations and translate the evidence from those evaluations into practice. Key stakeholders across this broad sector includes education providers who incorporate careers provision (schools, colleges, sixth forms, universities, apprenticeship and training providers), direct careers support providers (for young

people, out-of-work adults, or in-work adults), relevant professional and sectoral bodies, and those who commission or direct careers support services.

The Evaluation Framework needs to be fit for purpose across a broad range of AI technologies and careers support contexts, as well as reflect different evaluation types across a broad cost/quality spectrum.

The Evaluation Framework will also report on the process, results, and lessons learned of specific Pilot Evaluations, i.e. exercises to evaluate actual AI tools in a real world setting.

The Pilot Evaluations will provide a detailed worked example of three types of methodology that aspire to meet the Evaluation Framework's standards, support exploratory, hypothesis setting research, as well as provide indicative quantitative conclusions that can narrow the range of plausible effect sizes to anticipate for AI-powered guidance (including cases where the uncertainty range includes no effect).

The Evaluation Framework will not prescribe that every evaluation must incorporate all available elements – such as panel review, case studies, or field trials. Rather, it will describe how different elements should be designed, conducted, and reported in order to produce robust and comparable results. This includes indicative quantitative parameters to inform sample size selection, and sufficient methodological detail to support interpretation and replication across studies. Recognising that evaluations must make trade-offs between budget, speed, and rigour depending on context, the framework will provide practical guidance across a range of scopes and cost levels for different evaluation methods.

One key role of the Advisory Group is to provide ideas of how best to engage a diverse set of technical users of the evaluation framework and broader sector engagement with subsequent evaluations.

3 Proposed evaluation methods

Three evaluation methods are proposed for elaboration and piloting during this research:

- Method 1: Panel review
- Method 2: Case Study Comparison
- Method 3: Randomised Control Trial

The Evaluation Framework will explain how established evaluation methods can be applied specifically to the assessment of LLM-powered career advice across a diverse range of careers settings and contexts.

The Pilot Evaluation will apply these methods on a modest scale within a defined and narrow range of careers settings. Its purposes are threefold:

- i. To test and refine the guidance set out in the Evaluation Framework;
- ii. To provide a worked example that makes the framework more accessible to practitioners; and;
- iii. To generate qualitative findings and indicative quantitative hypotheses about the performance of selected AI tools.

In this way, the pilot functions as both a proof of concept and a practical resource for those undertaking subsequent evaluations.

3.1 Method 1 overview: Panel review

An appropriate and diverse panel of careers service practitioners, experts, and 16-19 students is convened to use and assess the technology within a structured rubric (which identifies why and how something is evaluated and identifies performance measures for success). This will cover groups of students with varying levels of certainty about their career choices.

For this method, key areas that the Evaluation Framework should cover include:

- Number of panel members from different groups (for example, trained careers advisers, AI experts and beneficiary representatives, such as students and potentially their parents); what are the minimum and recommended outcome measures for different use cases (i.e. actors, contexts and questions)?
- How should the panel be managed (e.g. what are the guidelines for convening the panel, and what training/support should be provided to panel members)?
- How should panel members use the technology (e.g. individually/jointly, minimum usage duration, preparation/follow-up work)?
- How should panel members assess the technology using a range of outcomes measures (e.g. relevance, appropriateness, usability, tone, accuracy, actionability etc)?
- How could a consensus result be produced and translated into appropriately-framed recommendations given that different users (individuals, advisers, parents/carers) are likely to have different priorities (e.g. how well does a tool need to perform before it might be recommended for a particular use case; how to set a level for actual practice comparison)?

3.2 Method 2 overview: Case Study Comparison

Case study comparison works by drafting distinct personas with specific careers questions, grounded in real world experience of diverse groups, such as lower secondary, upper secondary, and tertiary education students.⁸ The target tool is provided with different personas and questions (determined, for instance, by careers advisers playing the role of the relevant personas), then asked to respond to that scenario, as they would in the real world. Professional careers advisers score the tools' responses to the scenarios on a blind basis using an agreed scoring rubric.

For this method, key areas that the Evaluation Framework should cover include:

- The requisite range and number of personas and scenarios needed to achieve adequate coverage across diverse client profiles.
- A clear explanation of what constitutes a case study in this context, how case study scenarios can be generated, the appropriate scale of individual scenarios, and how multi-step client interactions should be represented.
- Guidance on developing a scoring rubric and associated evaluation framework for evaluating responses to case study scenarios.

⁸ Note that while the generic Evaluation Framework can cover a wide range of groups (such as unemployed adults) and hypothetical AI use cases, the scenarios for our specific Pilot Evaluation are scoped to focus only on the 16–19 Further Education demographic.

- An explanation of how scores are aggregated and compared against defined thresholds to inform decision recommendations.

3.3 Method 3 overview: Randomised Control Trial (RCT)

Students aged 16-19 are recruited and randomly assigned to different interventions, providing a robust basis for understanding the value of the tool.

For this method, key areas that the Evaluation Framework should cover include:

- Explaining the methodological approach i.e. the process, statistical underpinnings and requirements, best practice approaches, budget considerations and challenges.
- Describe different ways that users of the career advice tool (or their representatives) might be recruited and randomly assigned to different interventions.
- Describe different ways the interventions might be deployed to those users, how the results are gathered, and assessed (including possible 'observer effects' or 'self-selection' effects with advisers such that performance is higher than typical).
- Ethical considerations.

3.4 Objectives to be evaluated

In all cases, an evaluation needs to explain what objectives a tool is being evaluated against. It is helpful to structure these objectives in terms of three key decisions:

- **Criteria** - deciding which career-related outcomes matter, and in what contexts;
- **Indicators** - deciding how those criteria are measured, and;
- **Standards** - deciding what levels on outcome measures count as good or better.

It is also important that the language of evaluation does not imply that AI tools are being assessed as potential replacements for practitioners. Throughout the framework, evaluations should be framed as assessing how well a tool supports and extends the work of careers professionals - augmenting human expertise rather than competing with it. This framing reflects the substantive purpose of the tools being evaluated and is essential to maintaining the trust and engagement of the professional community throughout the research. While the Evaluation Framework will allow space for a broad mix and weighting of different criteria, it is likely that common high-level categories of criteria will be relevant across many contexts, such as:

- **Effectiveness** e.g. assessed impact against a careers objective (such as, reduced career uncertainty, greater confidence in a career action plan, access into work or training, beneficiary satisfaction).
- **Equity**, e.g. the accessibility, suitability, and effectiveness of the tool across diverse users and beneficiaries, including variation by demographic characteristics, digital competence, and additional needs:
 - o **Safety and assurance**, e.g. the extent to which the tool operates within appropriate boundaries: recognising when users may have additional needs requiring specialist support (such as wellbeing referral), directing users to a qualified careers adviser when the limits

of the tool have been reached, and being transparent about what it can and cannot reliably do.

- o **User experience**, e.g. how easy is the tool to access, sign up to, navigate, resume a previous session, and save outputs.
- **Efficiency**, e.g. achieving these impacts in a cost-effective manner.
- **Sustainability**, e.g. the extent to which the tool/service is available, likely to be used and how it is used by key stakeholders, and how it aligns with broader system levers/priorities.

In addition to the Pilot Evaluation (which will provide a fully worked example of objective design), the Evaluation Framework will provide several illustrative examples of different criteria, indicators, and standards that might be appropriate for different contexts.

The Pilot Evaluation will not seek to address matters of cost-efficiency or sustainability, but the Evaluation Framework will provide guidance on how evidence from evaluations can be used to inform these important judgements.

These examples will function as a practical reference - analogous to the outcome measure guidance produced by the Education Endowment Foundation for social and emotional skills⁹ - allowing evaluators to select and adapt measures suited to their setting. Across all categories, evaluations should attend not only to outcomes but to the conditions under which those outcomes are achieved, including the quality of delivery and implementation context. The TIDieR framework (Template for Intervention Description and Replication)¹⁰ is recommended as a useful reference for ensuring intervention descriptions are sufficiently detailed to support rigorous and replicable evaluation.

4 Proposed Evaluation Framework

4.1 Evaluation Framework Document Outline

As well as a research pdf, the Evaluation Framework will be available as an online toolkit (hosted, as a static resource, on IER's website) for easy access, search, and navigation across sections. Key sections of the document/toolkit are proposed:

1. Summary: to include a one page diagram of the core Evaluation Framework structure/docs/templates.
2. Introduction: context, motivation, & goals/ambitions for the Evaluation Framework.
3. How to use this Evaluation Framework: different audiences and how we hope they might engage with the Evaluation Framework.
4. Evaluation methods comparison: describe three 'archetypal' methods but briefly noting other methods and variants.
5. Method 1: Panel review: explanation, pros/cons of the method, considerations, process overview, how to set baselines/thresholds, good practice & pitfalls.

⁹ <https://educationendowmentfoundation.org.uk/spectrum-essential-skills-and-non-academic-outcomes>

¹⁰ https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/IPE_Handbook.pdf?v=1780304834

6. Method 2: Case Study Comparison: explanation, pros/cons of the method, considerations, process overview, how to set baselines/thresholds, good practice & pitfalls; 10 specific anonymised client case studies to be included, along with scoring rubric.
7. Method 3: Randomised Control Trial: explanation, pros/cons of the method, considerations, process overview, how to set baselines/thresholds, good practice & pitfalls, ethical considerations.
8. Simple alternative methods: possible lower cost and more operationally feasible versions of the main methods, e.g. non-randomised control trial variants as small-scale trials that might be better suited to individual schools/colleges.
9. How this can be developed further: including incorporating new technologies that are hard to foresee at this stage, being applied to new careers questions, reviewing/developing the evaluation framework itself.

Appendices

- Evaluation Framework generation process: project description; steering group members.
- Example pre-registration template, outcome/activity reporting templates, checklists to include example outcomes to targets and ways to measure them.
- Example sample size calculation.
- Example ethical assessment form to include a privacy impact assessment.

Pilot Evaluation Supplement

- Full Pilot materials and results for the 16-19 webchat advice use case to include recommended revisions to the approach given lessons learned.

A longer-term aspiration of this framework is to support the development of a centralised register of LLM evaluations in careers advice contexts. Such a register would reduce duplication of effort across organisations, build cumulative evidence on LLM performance over time, indicate the breadth of framework uptake, and provide a resource for future research synthesis. Precedents exist in other fields – including trial registries in development economics and pre-registration databases in education research – and a similar model merits exploration as the framework matures and the evaluator community grows.

5 Pilot Evaluation Scoping

Each of the three methods is to be tested as part of a Pilot, to learn about implementation considerations to share, to provide a worked example for the Evaluation Framework, and to provide initial insights for the sector. As an indicative high-level scope for the Pilot:

5.1 Method 1 pilot scope: Panel review (see Section 2 in the complementary Methodology paper)

- **Participants:** Twenty panel members indicatively: 8 experienced and qualified careers advisers with experience in the target use case to be recruited via Advisory Group networks, e.g., the CDI, IEP/ERSA or AoC; 8 target users or user representatives - students aged 16-19 drawn from Further Education Colleges participating in the RCT; 4 AI/technology experts – also to be sourced via the Advisory Group and its networks).
- **Technologies reviewed:** One generic AI tool and one sector-specialised AI tool (see Technology Selection section below).
- **Target use case:** Students aged 16-19 with substantial careers questions.
- **Process design:** By project team with Advisory Group input.
- **Impact to evaluate:** To be specified as part of the next stage of pilot design.

5.2 Method 2 pilot scope: Case Study Comparison (see Section 3 in the complementary Methodology paper)

- **Participants:** Careers advisers to score advice transcripts, with each transcript scored by two different advisers.
- **Technologies reviewed:** One generic AI tool, one sector-specialised AI tool (see Technology Selection section below), plus two professional adviser interventions for comparison (same as Method 3).
- **Use case:** Advice transcripts to be generated for ten client scenarios for students aged 16-19 in different contexts.
- **Process design:** By project team with Advisory Group input.
- **Impact to evaluate:** To be specified. For example, one evaluation criterion might be how well the conversation would help the case study user progress their career readiness, with indicators based on careers advisers scoring the conversation against the set rubric and standards based on provision sufficiently close to that of a professional adviser.

5.3 Method 3 pilot scope: Randomised Control Trial (see Section 4 in the complementary Methodology paper)

- **Participants:** To recruit 160 student participants aged 16-19 (part-time or full-time education) with substantial careers questions, to support participation levels of around 128 participants or more.
- **Random assignment** to one of four interventions to provide support with the participants' careers questions:
 - o One generic AI tool (same as method 1)

- o One sector-specialised AI tool (same as method 1)
- o Professional careers adviser via video call¹¹
- o Hybrid (adviser with access to the sector-specialised AI tool)

(each intervention is allowed up to 60 minutes within a 7-day period; participants may leave early, but will be asked to schedule 60 minutes for it)

- **Participant recruitment:** Participating education provider career services (minimum ten providers, at least two in each of the four UK home nations) will email their current list of students, explaining the voluntary opportunity of participating in the trial. Interested students fill in a sign-up survey, from which we select qualifying students. The sign-up form includes informed consent, basic background characteristics (e.g. gender, ethnicity, current/prior Free School Meals (FSM) eligibility, AI tool familiarity/preferences). Data collection, anonymised and secure data storage, and informed consent will be aligned with Information Commissioners Office (ICO) and BERA (2024) guidelines.
- **Participant breakdown:** Targeting 16 participants from each of the 10 colleges and targeting at least 2 colleges from each of the 4 nations. Targeting at least 20% participants across the whole sample to be FSM eligible. Random selection within each college will be used in the case of over-subscription of potential participants.
- **Impact to evaluate:** All participating students across the four interventions will complete pre- and post-surveys to measure: satisfaction, confidence, trust, usefulness, ease of use, intentions, usage of other sources of careers guidance, and career readiness. Apart from the immediate post-intervention questionnaire, there will also be a 6-8 week follow-up survey on actions taken or initiated and further reflective feedback. Participation in this follow-up is a core part of the commitment asked of participants at the point of recruitment, given that longer-term behavioural outcomes are among the most important indicators of impact. Consent remains an ongoing process throughout, and participants retain the right to withdraw at any stage; however, the importance of the follow-up will be made explicit in recruitment materials and reinforced through appropriate retention protocols to support response rates.

5.4 Managing ethical risks to student participants in the field trial

The main ethical issue revolves around randomly assigning beneficiaries to one of four interventions using a stratified random sampling approach. This could mean they undertake an intervention they would not normally have selected at a time when they could be making important career choices. This issue is mitigated in four ways, in addition to informed consent and voluntary participation:

- First, the trials are constructed so that students can use whatever IAG they normally access, provided they also use the assigned intervention.

¹¹ Where interventions involve a human careers adviser - whether alone or in the hybrid condition - steps will be taken to manage variation in adviser quality and approach. These will include: minimum qualification criteria for participating advisers (Level 6); a standardised briefing protocol defining the scope and boundaries of the session; fidelity checks against session transcripts; and, in the analysis, statistical adjustments for adviser-related features.

- Second, all session transcripts are reviewed for scoring purposes, during which any inappropriate or inaccurate information will be flagged and corrected retrospectively with any affected students as soon as identified.
- Third, all participating students will have additional access to a human careers adviser after the trial concludes (after the follow-up survey) if they have questions.

As a result, all students should have careers support at least no worse than would have been anticipated and typically better than without the trial in place. Note that the field trial will operate in an ecologically valid setting, in that students will be permitted to access their normal sources of IAG during the evaluation period. The trial is testing the additional benefit of careers provision in such a real-world setting.

NB: An application for ethical approval has been made to the University of Warwick's Humanities and Social Science Research Ethics Committee (HSSREC).

A central ethical consideration in any evaluation involving LLM-powered careers guidance tools is compliance with professional codes of ethics – in particular, the requirements of impartiality and the avoidance of bias. Tools developed for or commissioned by a specific institution (such as a sixth form, college, university, training provider, or employer) present a particular risk: their design may structurally favour certain options or pathways, whether by intent or by the boundaries of their curated content. The Evaluation Framework will require that this be declared explicitly. Where a tool is designed to operate within a defined and restricted scope, this must be clearly communicated to users; where a tool presents itself as offering impartial guidance, it must be evaluated against that claim.

Professional standards in this area are not the property of any single body. While the CDI Code of Ethics provides relevant guidance, other professional bodies – including NAMSS, the Institute of Graduate Futures (formerly AGCAS), and the Institute for Employability Professionals (IEP) – each maintain their own codes of ethics and are developing their own AI policies and procedures. The Evaluation Framework is designed to be compatible with this broader professional landscape. Evaluators and commissioners are encouraged to assess compliance with the code or codes most relevant to their practitioner community and deployment context, rather than treating any single body's standards as exhaustive.

5.5 Technology selection: Long-list and selection criteria

Technologies are identified based on their potential to support the specific careers support use case being evaluated in the pilots: students aged 16-19 in full-time or part-time education in the UK, where those students have a 'substantial career question' to explore (i.e. excluding those with only tactical questions such as how to fill in an application form or prepare a CV). In other words, the technologies must be capable of providing flexible information and advice on a broad range of substantial topics, such as what career might suit them, what education and training pathway they might follow, and what actions they might currently take.

The project lead, Peter Dickinson, reviewed the criteria for selecting the AI tools and provided a draft recommendation for the 2nd Advisory Group meeting to consider, reflecting on:

- Explicit focus on careers support that incorporates the use case (i.e. 16-19 year old college students);
- Explicit coverage of the four UK home nations;

- Current relevance within target market (e.g. such that insights on the technology are relevant to more stakeholders/users), and;
- Availability for the evaluation (e.g. accessibility, willingness, and budget-fit).

Following this, a selection of current AI technologies meeting the criteria were presented to the Advisory Group to agree which should be included in the study.

At the current scope, only one generic advice technology and one careers-specialised technology are included in the Pilot Evaluation. Other funders or technology providers are invited to resource an expansion of the scope to include additional technologies in the Pilot and corresponding increase in sample size.

For generic chatbots, Table 1 lists those AI chatbots that are sometimes used for careers guidance. Based on discussion at the 1st Advisory Group meeting, Copilot is the generic AI tool that is most currently used within the target provider group by education institutions and ChatGPT is most commonly used by students outside of an institution-facilitated setting.

Survey evidence identifies ChatGPT from OpenAI as the most common LLM used by teenagers in the UK. For instance, out of the top 10 digital tools or apps named by FE learners in a major Jisc survey, only one of them is an LLM, which is ChatGPT at 4% name-share (for comparison, Google Classroom is 7%; Microsoft Word is 3%).¹² Among children aged 9-17, 64% had used at least one chatbot and out these the most popular was ChatGPT/GPT-4 at 44%, followed by Gemini at 32%, Snapchat AI at 31%, and Copilot at 20%.¹³ Evidence also suggests users are primarily using the free consumer version of these LLMs. For instance, Attewell, Head of AI at Jisc explained that “Most learners/students are using free versions of AI applications”.¹⁴ Indeed, the vast majority of ChatGPT (or other LLM) users are on free versions.¹⁵

Table 1: Commonly used Generic AI chatbots used for careers advice and guidance

Generic AI technologies that are used to provide advice (Frontier LLMs)	OpenAI – ChatGPT model suite
	Anthropic – Claude model suite
	Microsoft – Copilot model suite
	Google – Gemini model suite
	Perplexity AI

Table 2 lists the dedicated career sector AI tools that are currently available in the UK. Details are provided against the above criteria for each of the AI tools. This was undertaken by IER

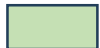
¹² Jisc data analytics (September 2025). Learner digital experience insights survey 2024/25: UK further education (FE) survey findings.

¹³ Internet Matters (July 2025). Me, myself & AI: Understanding and safeguarding children’s use of AI chatbots.

¹⁴ Jisc (May 2025). <https://www.jisc.ac.uk/reports/student-perceptions-of-ai-2025>

¹⁵ <https://www.reuters.com/technology/openai-projected-least-220-million-people-will-pay-chatgpt-by-2030-information-2025-11-26/>

researchers using publicly available information. There is a traffic light system for identifying how well these criteria have been met. Based on current evidence:



Appears to meet the criteria



Appears to meet most criteria but with at least one reservation



Does not appear to meet the criteria

Table 2: Dedicated career sector AI chatbots

Name of chatbot, company and URL	Explicit focus on careers support that incorporates the use case (16-19 year olds)	Explicit coverage of the UK market, including in each devolved authority	Current relevance within target market (e.g. broad usage not a niche product)	Availability for the evaluation (e.g. accessibility, willingness, and cost)
Tool: Ask Nim Company: Stay Nimble URL: https://nim.staynimble.co.uk/access	Strong focus on career support, but the main public positioning is broader career coaching for individuals, teams and adults navigating work transitions. It is described as an AI career co-pilot that supports career questions, skills development, CV/ interview support etc. It could incorporate the 16–19 use case, but the strongest suggests it relates to adults and organisational users	Strong UK basis. Stay Nimble is UK-based and Ask Nim has been deployed through Kingston Council/Connected Kingston, suggesting local public-sector relevance. However, there is no explicit evidence of differentiated coverage across Northern Ireland, Scotland and Wales.	Relevant and active. Stay Nimble describes itself as combining qualified coaches, AI tools and community-based support, with more than 20,000 people supported according to its LinkedIn profile. Ask Nim is also positioned as a 24/7 scalable career development tool.	Potentially available but likely requires provider engagement. There is a gated access route and organisational pricing is visible. The integration package is listed at £600 per month or £6,500 per year, with further support costs. This may fit an evaluation if access can be negotiated.
Tool: Career Coach Company: Apt AI URL: https://www.tryapt.ai/career-coach	Moderate to strong focus on career support. Apt AI presents its tool as an AI career coach and mentor, with a career test followed by a chat with an AI coach. It covers career planning, skill development, interview preparation and professional growth. Is not visibly designed around UK 16–19 learners or education-to-work transitions.	Weak to moderate UK coverage. The public materials emphasise general AI career coaching, but do not provide clear evidence of UK-specific labour market localisation or coverage of devolved authorities.	Potentially relevant as a consumer-facing AI career coach. Apt claims high usage/review figures, but these should be treated cautiously because they are provider-reported and not independently verified in the material reviewed.	Likely accessible for user testing through its public route. The main issue is not access, but relevance: it may be more suitable as a generic career-coaching comparator than as a UK-sector specialised IAG tool.
Tool: Career COACH Company: Rocky AI URL: https://www.rocky.ai/career-coach	Moderate focus on career development, but weaker fit with the specific 16–19 careers guidance use case. Rocky AI is mainly framed around growth mindset, personal development, soft skills, leadership and daily micro-coaching. It includes career development goals and career plans, but not specialist careers guidance.	Weak to moderate UK coverage. Rocky Robots Ltd has a UK presence, but the product is not clearly localised to UK education pathways, UK labour market information, or the four UK nations.	Relevant as a broader AI coaching platform, especially for personal development, workplace coaching and soft skills. Less directly relevant to the target market if the pilot focuses on 16–19 learners with substantial education and career pathway questions.	Publicly accessible in some form and available as an app/platform, with business options and trials. However, for the proposed evaluation it may require careful scoping because the tool is not mainly designed for careers IAG.

Name of chatbot, company and URL	Explicit focus on careers support that incorporates the use case (16-19 year olds)	Explicit coverage of the UK market, including in each devolved authority	Current relevance within target market (e.g. broad usage not a niche product)	Availability for the evaluation (e.g. accessibility, willingness, and cost)
Tool: Cari Company: Careers Wales (details being double-checked with Careers Wales as of 21 May) URL: https://careerswales.gov.wales/	Strong focus on careers support as an organisation, but unable to verify from public sources that “Cari” is currently a distinct AI careers chatbot. Careers Wales supports young people with planning careers, courses, apprenticeships, job information, decision-making tools and adviser access. It has explicit material for young people approaching age 16.	Strong Wales coverage, but not UK-wide. Careers Wales is a Welsh Government-funded national careers service for Wales. It does provide UK and Wales labour market information, but its institutional remit is Wales rather than England, Scotland or Northern Ireland.	Highly relevant within Wales, especially for young people in secondary education and transitions after GCSEs.	Careers Wales services are publicly available.
Tool: CiCi Company: CareerChatUK URL: https://cicichat.co.uk/	Very strong fit. CiCi is explicitly described as an AI careers chatbot for young people and adults, supporting personalised career exploration, labour market information, skills, jobs and career readiness. It includes a link to a human adviser, which is important for safeguarding and professional accountability.	Strong UK coverage. CareerChatUK explicitly refers to schools, colleges, universities, local/combined authorities and compliance expectations across UK inspection and career systems, including Ofsted, Estyn, ETI Northern Ireland and Education Scotland.	Very relevant. It is one of the clearest UK-based specialist AI careers chatbots and has also been profiled by the OECD as a 24/7 careers chatbot for youth.	Likely available but requires conflict-of-interest management and provider cooperation. The draft report already recognises the need to manage conflicts if CiCi is selected.
Tool: COACH Company: CareerVillage URL: https://www.aicareercoach.org/	Strong focus on career support. COACH is described as an AI-powered career coach designed to scale career development, help users explore careers, build skills and improve job readiness. It is relevant to young people and learners, although it is not specifically designed around the UK 16–19 education system.	Weak UK-specific coverage. CareerVillage is a US-based non-profit and COACH appears global, with claimed use across more than 120 countries and 10 languages. There is no clear evidence of UK-specific labour market localisation or devolved-authority coverage.	High relevance as an international benchmark. It reports substantial use, including more than 115,000 users and 35 institutional partners, and states that 90% of learners report improved career readiness. These are provider-reported impact figures and should be treated as indicative rather than independent evaluation evidence.	Strong availability. Individuals can try the tool for free, and there is an institutional route. It would be feasible as a comparator, but less strong as the UK-specialist tool.

Name of chatbot, company and URL	Explicit focus on careers support that incorporates the use case (16-19 year olds)	Explicit coverage of the UK market, including in each devolved authority	Current relevance within target market (e.g. broad usage not a niche product)	Availability for the evaluation (e.g. accessibility, willingness, and cost)
Tool: Morrisby Company: Huxby URL: https://www.morrisby.com/blog/morrisby-and-ai	Strong fit with the 16–19 use case. Morrisby is a careers guidance platform for schools and colleges, built around psychometrics, career exploration, subject choices, apprenticeships, university courses, Gatsby Benchmarks and destinations. Huxby is presented as Morrisby's first AI step to augment adviser-led services and provide support and information when needed.	Strong UK education relevance, but devolved-authority specificity needs checking. Morrisby is widely used by schools and colleges and is aligned with UK careers education needs, but the public source reviewed does not clearly state coverage by Northern Ireland, Scotland and Wales.	Relevant but unvalidated. Morrisby claims 5 million+ career journeys, 1,700+ schools and colleges, and 100,000+ assessments a year. This makes Huxby potentially relevant because it is embedded in an established careers platform rather than a standalone chatbot, but it is unclear whether and how Huxby is currently being used.	Likely available through institutional access or provider cooperation. Public access may be limited compared with open consumer tools, so evaluation would probably require direct contact with Morrisby.
Tool: Outcore Company: Purlos URL: https://www.purlos.com/outcore/	Very strong fit. Outcore is explicitly described as an AI-driven assistant for scaling careers support. It supports learners with career paths, goal-setting, UCAS applications, CVs, meeting booking with careers advisers, multilingual support and learner profiles for staff. This is directly relevant to the 16–19 education setting.	Strong UK institutional relevance, but devolved-authority coverage is not fully explicit. Purlos describes its community as including UK colleges, universities, training providers and combined authorities, and Outcore's dashboard refers to Gatsby Benchmarks, Ofsted and institutional self-assessment reporting.	Very relevant. Purlos states that its destinations and outcomes platform is trusted by over 500,000 students and more than 100 UK education and employability providers. Outcore is therefore likely to be relevant to providers facing scale and capacity pressures in careers support.	Potentially available through demo or free trial routes but likely requires provider cooperation. It may be one of the stronger candidates for formal evaluation because it includes both learner-facing and staff-monitoring functions.
Tool: Rodeo Company: Rodeo URL: https://gorodeo.app/	Strong focus on careers support, especially career planning, job matching and application support. Rodeo is positioned as a personal career agent that helps users plan, find and land their next move. It includes examples relevant to final-year students considering graduate schemes or postgraduate study, so it has partial fit with the 16–19/young adult transition use case.	Strong UK focus. Rodeo only covers the UK at this stage in order to provide more contextualised service.	Partial coverage. Rodeo primarily covers 'job-seekers', which is primarily an adult population. However, some FE users aged 16-19 fall into the 'job-seeking' category.	Rodeo is currently available free of charge to users. It is also possible to contact them directly to invite them to be formal partners.

Name of chatbot, company and URL	Explicit focus on careers support that incorporates the use case (16-19 year olds)	Explicit coverage of the UK market, including in each devolved authority	Current relevance within target market (e.g. broad usage not a niche product)	Availability for the evaluation (e.g. accessibility, willingness, and cost)
Tool: Coach1st Company: Achieve / Career Chatbot URL: https://achieve.coach1st.co.uk/	Moderate fit. Achieve/Coach1st is a digital career and job-search coaching platform with expert-led employability content, AI CV optimisation, a career chatbot and progress tracking. It is more focused on job search, confidence, interviews and CVs than on formal careers guidance for 16–19 learners.	Moderate UK relevance. Coach1st appears UK-based and focused on career/job-search coaching, but there is no clear evidence of UK labour market localisation across devolved authorities.	Moderate relevance. It is useful for mapping AI-supported employability coaching, especially for adults, jobseekers and people in transition. It is less directly relevant than CiCi, Huxby, Outcore or Careers Wales for the school/college use case.	Strong initial availability. The platform is presented as free and includes a 24/7 career chatbot. It could be tested easily, but its relevance to the pilot should be treated as lower priority.
Tool: Maya / Rumbo Company: Rumbo URL: https://www.rumbocareers.com/	Moderate to strong conceptual fit, but early-stage. Rumbo is described as AI career coaching that starts with self-understanding, structured assessments, a personal AI coach, and exploration of fit between personality, interests, skills and market demand. This is relevant to substantial career uncertainty, although it appears aimed at broader career decision-making rather than specifically 16–19 learners.	Moderate UK relevance. Rumbo's LinkedIn page lists London as its headquarters and describes the company as UK-based. However, explicit UK labour market localisation and coverage of devolved authorities are not yet clear from public sources.	Emerging relevance, but currently early-stage. Rumbo appears newly founded in 2026, with LinkedIn posts referring to a launch, a waitlist, 500 sign-ups in the first three days, and an AI coach called Maya. This makes it interesting as an emerging candidate, but not yet as mature as the established providers.	Unclear. The public website appears to exist, but the strongest available evidence is currently from LinkedIn rather than a fully documented product site. It should be treated as a watchlist candidate unless direct access can be confirmed.

Research team colleagues with roles at CareerChatUK Ltd (CCUK) were excused from the short-listing process, beyond providing access to the tool and answering questions that support the project lead in prioritisation.

5.6 Managing conflicts of interest

During the formal technology prioritisation process, CCUK affiliated colleagues left the call/room.

Should CiCi be selected for the Pilot Evaluation, the Advisory Group agreed that conflicts of interest would be mitigated in five ways:

- First, the CiCi team will not be involved in accessing/analysing any evaluation data or conducting the pilot evaluation, beyond providing access to technology. Furthermore, individuals with associations to the technologies under review (including careers advisers) will not participate in analysis or decision-making regarding evaluation outcomes.
- Second, the evaluation will be subject to pre-registration and open data archiving.
- Third, the project is run by IER, who have no stake in CiCi, and all public reports are overseen by diverse sector stakeholders on the Advisory Group.
- Fourth, the evaluation is anticipated to operate at a pilot scale, designed to test the Evaluation Framework and produce hypotheses, rather than to have sufficient sample size to demonstrate conclusive evidence about any of the tools involved.
- Fifth, the tech experts who will review both the dedicated and generic chatbots will be selected by the project PI in consultation with AG members.

Appendix 1: Advisory Group Terms of Reference and members

Project Period: January 2026 - December 2027

Principal Investigator: Associate Professor Peter Dickinson, Warwick University, Institute for Employment Research (IER) – p.dickinson@warwick.ac.uk

Co-Investigators: Associate Professor Deirdre Hughes & Dr Chris Percy, Warwick University IER & Co-Founders, CareerChatUK

Nuffield Foundation Grant Lead: Dr Emily Tanner - etanner@nuffieldfoundation.org

Advisory Group Chair: Professor Huw Morris, University College London, IOE - Social Research Institute – huw.morris@ucl.ac.uk

1. Purpose and Role of the Advisory Group

The Advisory Group will provide strategic oversight, expert guidance, and constructive challenge to ensure the research project delivers robust, ethically sound, and practically applicable Evaluation Frameworks for LLM-powered careers advice services.

2. Terms of Reference

The Advisory Group will:

1. Provide oversight, advice and challenge on the conduct of the research, including the suitability and quality of methods and approaches for assessing LLM-powered careers advice systems and tools.
2. Highlight, discuss and, where possible, resolve ethical considerations that emerge in the course of the research, particularly relating to the use of specific AI tools as part of the research, AI safety, equity of access, data privacy, and the appropriate balance between human expertise and AI assistance.
3. Advise and comment on research outputs, including style and interpretation of results, to help identify implications for policy, practice, and research, and the robustness of an Evaluation Framework that will be made openly available, as part of the research.
4. Help shape and advise on communication and dissemination of project outputs and findings to diverse stakeholder groups including policy makers, careers and employability practitioners, education providers, technology developers and the people they serve.
5. Advise on areas or issues where the research team needs to seek external guidance, whether from the Nuffield Foundation as funder, or from the research, policy, practice, or technology communities.
6. Provide specialist professional expertise to ensure the research reflects current practice and addresses real-world challenges.
7. Support knowledge exchange between the research project and wider developments in AI governance, careers technology innovation, and professional practice standards.

3. Membership

The Advisory Group will comprise members representing a diversity of backgrounds and experiences so that a range of perspectives contribute to the research :

- Policy and government delivery bodies across the UK nations
- Funders with sector expertise
- Careers adviser representatives and professional bodies
- Education providers and sector bodies
- CareerTech providers and AI experts
- Standards-setting organisations
- Voices of beneficiaries of careers advice.

The Principal Investigator will support the Chair working closely with the Co-Investigators. The Nuffield Foundation Grant Lead will be a member of the group.

4. Meetings and Commitment

- Meetings are timed to align with milestones and decision points in the project. All meetings will be hybrid with a preference for in-person meetings (where possible) at the University of Warwick, IER, Coventry, CV4 7AL, England for the first and third meetings mentioned below.
- The Advisory Group will meet up to x7 times over the project period, beginning:
 - o First AG meeting – 15th April 2026 ‘in person/hybrid ’ from 10:00-13:00.
 - o Second AG meeting – 9th June online from 10:00-12:00.
 - o Third AG meeting – 21st July 2026 ‘in person/hybrid’ from 10:00-13:00.
- Members should expect to dedicate approximately 3 hours per meeting, plus preparation time for reading materials circulated in advance
- Meeting notes will be shared with all members following each meeting
- Members unable to attend may submit written comments on key agenda items in advance of the scheduled meeting.
- Formal decisions and recommendations require a quorum of 60% of advisory group members present, either in person or virtually.

5. Confidentiality

All advisory group members must treat as confidential:

- Discussions in advisory group meetings and email correspondence
- Any documents shared, including pre-publication copies of research reports and working papers
- Commercially sensitive information about AI systems under development
- Any personal data or case study materials shared for research purposes.

Members may discuss the existence and broad aims of the project publicly but should not disclose specific findings, methodologies, or discussions without prior agreement from the Chair and Principal Investigator.

6. Travel and Expenses

Members of the advisory group may claim reasonable travel costs incurred in attending formal meetings. Claims should be submitted to IER's Finance Officer Yifan Wang (Yifan.Wang@warwick.ac.uk) and copying in Stef Poole (Stefanie.Poole@warwick.ac.uk).

7. Duration

This Advisory Group will operate from mid-April 2026 to 1st December 2027, aligned with the Nuffield Foundation grant period.

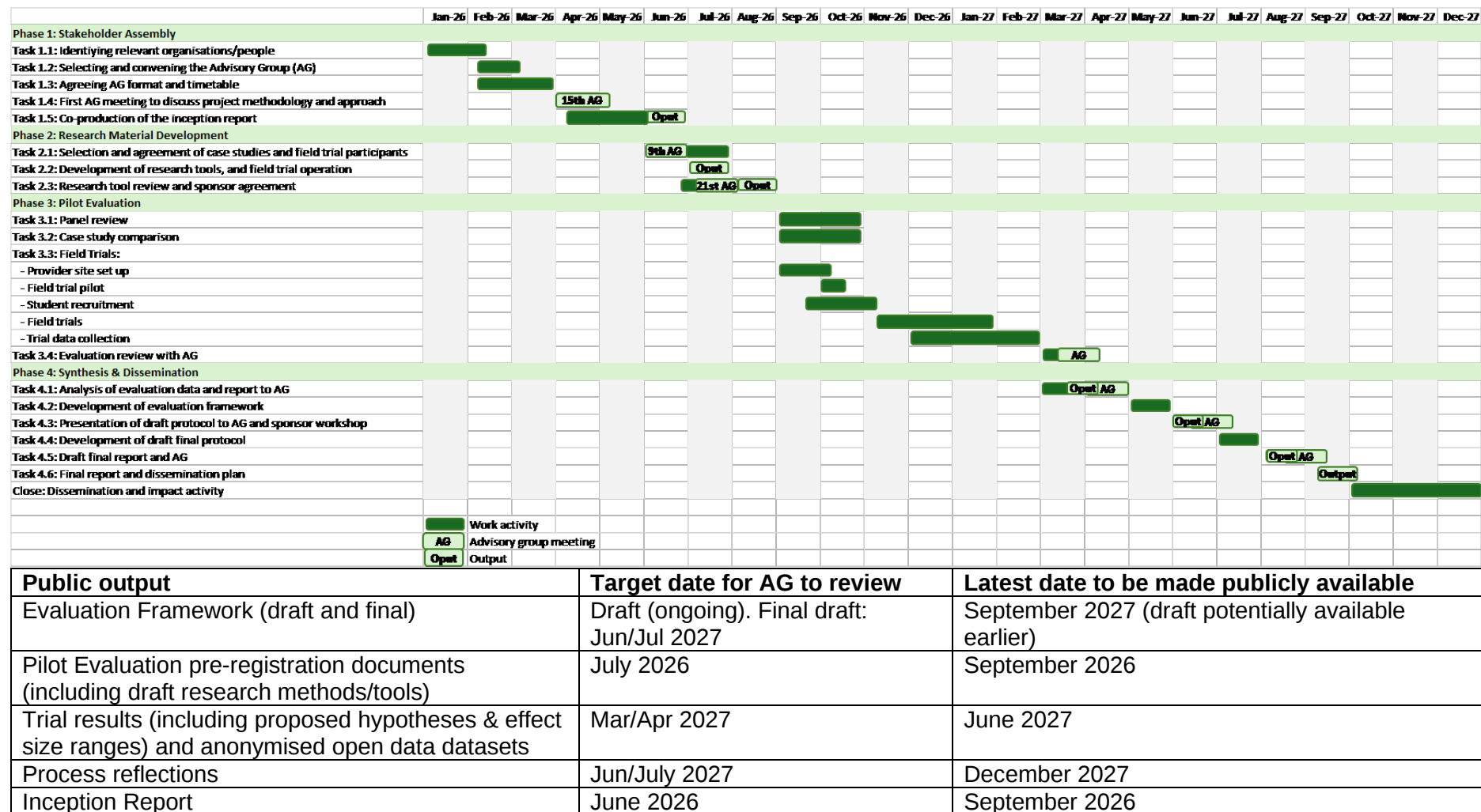
8. Review

These Terms of Reference may be reviewed and amended by agreement between the Chair, Principal Investigator and the Nuffield Foundation Grant Lead, in consultation with the research team and Advisory Group members.

Advisory Group Members

A spreadsheet of Advisory Group members is available on request and will be published along with the Evaluation Framework.

Appendix 2: Project timeline overview



Appendix 3: Risk Management

Risk Category	Risk	Impact	Probability	Mitigation Strategy
Staffing	Loss of key staff	High	Low	All senior members of the IER research team have permanent contracts and have no plans to leave the institution. Furthermore, the IER has a staff of 15 skilled and qualified researchers who can be additionally deployed on the project if required.
Resource Constraints	Project delivery	High	Low/Med	This grant application has been costed to reflect our realistic appraisal of the resources required to deliver the project successfully, including ensuring the appropriate number of staff days, factoring in the need to respond to any particular challenges we face as the research develops. The project budget will be overseen by experienced finance staff.
Difficulty recruiting participants and reviewers for primary research	Project delivery	Med	Med	Initial enquiries suggest that providers and potential reviewers are engaged with the project. We will draw on our extensive experience of qualitative research to secure respondents' in partnership with identified providers. Providers and/or participants who cannot engage with the research can be replaced on a like for like basis, a longlist and shortlist of both will be generated prior to fieldwork. Up to six months has been scheduled for the fieldwork across the three methodologies which is sufficient time to respond to and address any issues that might arise.
Technical Constraints	Systems failure	Med	Low	The research team has been selected to ensure we have the technical skills and expertise, supported by appropriate software and IT systems to conduct the research effectively. The University IT Services team has in place well-embedded disaster recovery plans which detail how the University would respond to an incident affecting the University's information technology provision. These plans cover a wide range of potential issues and are tested periodically. In addition, the University's business continuity and disaster recovery processes (including those within IT services) are regularly audited to ensure that they continue to provide assurance to the University and its stakeholders. Any audit-led recommendations are then managed accordingly.
Confidentiality and security of data	Loss of data	High	Low	The University of Warwick is committed to protecting the privacy rights of individuals who entrust the University with their personal data. The University has policies in place which set out how information security incidents, and breaches to the confidentiality of data, should be managed and to whom should they be escalated. The Data Protection Policy outlines the University's commitment to transparency and accountability and promoting good information governance. The University adopts, where relevant, the principles of recognised international standards, such as ISO 27001. All electronic data is stored within a secure environment on a fileserver. This means that only members of the data team have access to the files. No identifiable personal data is transferred outside consortium members (for example on laptops or memory sticks). The University has a substantial data breach investigation procedure https://warwick.ac.uk/services/legalandcomplianceservices/dataprotection/breaches/guidance

Timing	Meeting milestones	Low	Low	The timetable for this research programme has been carefully constructed to provide realistic time frames for completion of research tasks, based on the specific nature of each task.
Management	Communication with sponsor and Advisory Group	Med	Low	We have developed clear lines of management, communication and responsibility. The research team comprises staff with direct experience of managing research in this and related subject areas.
Ethics	Research with vulnerable groups	Med	Low	Our research will be subject to rigorous ethical approval overseen by University of Warwick Research Code of Practice