ISSN: 3033-4136
DOI: https://doi.org/10.17868/strath.00095463

Multimodal AI for Scaling Targeted Support:
Navigating the FCA Advice–Guidance Boundary

FRIL
Financial Regulation Innovation Lab

FinTech Scotland®

University of Glasgow

University of Strathclyde Glasgow

Multimodal AI for Scaling Targeted Support: Navigating the FCA Advice-Guidance Boundary

Hao Zhang, James Bowden, Mark Cummins

Strathclyde Business School, University of Strathclyde, and Financial Regulation Innovation Lab, 199 Cathedral Street, Glasgow, G4 0QU, UK

03 February 2026

We acknowledge funding from Innovate UK, award number 10055559.

Corresponding authors:

Email: hao.zhang@strath.ac.uk

Email: james.bowden@strath.ac.uk

Email: mark.cummins@strath.ac.uk

Financial Regulation Innovation Lab

Who are we?

The Financial Regulation Innovation Lab (FRIL) is an industry-led collaborative research and innovation programme focused on leveraging new technologies to respond to, shape, and help evolve the future regulatory landscape in the UK and globally, helping to create new employment and business opportunities, and enabling the future talent.

FRIL provides an environment for participants to engage and collaborate on the dynamic demands of financial regulation, explore, test and experiment with new technologies, build confidence in solutions and demonstrate their ability to meet regulatory standards worldwide.

What is Actionable Research?

FRIL will integrate academic research with an industry relevant agenda, focused on enabling knowledge on cutting-edge topics such as generative and explainable AI, advanced analytics, advanced computing, and earth-intelligent data as applied to financial regulation. The approach fosters cross sector learning to produce a series of papers, actionable recommendations and strategic plans that can be tested in the innovation environment, in collaboration across industry and regulators.

Locally-led Innovation Accelerators delivered in partnership with DSIT, Innovate UK and City Regions

Department for Science, Innovation & Technology

UKRI Innovate UK

GLASGOW CITY REGION

FRIL White Paper Series

Multimodal AI for Scaling Targeted Support: Navigating the FCA Advice–Guidance Boundary

Hao Zhang * James Bowden* Mark Cummins*

* University of Strathclyde

03 February 2026

Abstract

The Financial Conduct Authority's Advice Guidance Boundary Review (AGBR) seeks to address a persistent advice gap in UK financial services by enabling new forms of scalable, decision-relevant consumer support that sit between generic guidance and personalised financial advice. This challenge is particularly acute in pensions, where consumers face complex, long-term decisions but exhibit low engagement with traditional advice services. This white paper examines the potential of multimodal generative artificial intelligence to deliver targeted support in pensions while remaining within the advice–guidance boundary. Drawing on recent advances in Vision Language Models and multimodal conversational architectures, the paper develops a solution framework for speech-enabled, audio-visual digital advisors that are compliant by design. A Digital Pensions Advisor prototype is presented to demonstrate how such systems can interpret real consumer narratives, recognise and respond appropriately to vulnerability, and maintain boundary discipline when confronted with requests for personalised advice. The paper concludes by outlining a roadmap for strengthening auditability, explainability, and supervisory readiness, and by identifying future research directions, including the role of formal and informal information sources in shaping consumer understanding. Collectively, the findings suggest that multimodal AI can play a meaningful role in scaling targeted support in pensions while preserving regulatory safeguards and consumer trust.

Table of Contents

1. Introduction

The Advice Guidance Boundary Review (AGBR), led jointly by HM Treasury and the Financial Conduct Authority (FCA), constitutes a significant re-evaluation of how financial support is delivered to consumers in the United Kingdom.[1] The review is premised on a recognition that existing advice and guidance services are not functioning effectively for large segments of the population, particularly in relation to long-term financial decisions such as pensions. The FCA estimates that approximately 23 million consumers are currently underserved by the markets for advice and guidance, and that fewer than one in ten individuals obtain regulated financial advice.[2] As a result, many consumers make important financial decisions without structured support, or rely instead on informal sources such as friends, family, or unregulated online content.

At a conceptual level, the AGBR seeks to address structural limitations in the current regulatory framework that arise from the strict delineation between financial guidance and regulated advice. Under existing rules, guidance must remain generic and non-personalised, while advice requires a comprehensive assessment of the individual's circumstances and objectives, accompanied by extensive suitability and disclosure obligations. Although these requirements are essential to safeguarding consumers, the FCA acknowledges that they significantly increase the cost and complexity of providing advice. This, in turn, constrains supply and limits accessibility, particularly for consumers with modest assets or relatively straightforward needs. The AGBR is therefore motivated by the need to create additional forms of support that can deliver meaningful decision-making assistance at scale, without undermining the protections associated with regulated advice.

A central outcome of the AGBR is the FCA's proposal to introduce targeted support as a new regulated activity. Targeted support is defined as the provision of appropriate suggestions to groups of consumers who share common characteristics, needs or objectives, based on limited but relevant information, and without undertaking a full individualised suitability assessment (FCA PS25/22). The FCA is explicit that these suggestions are designed at the level of a consumer segment rather than an individual, and that they must be framed and delivered in a way that distinguishes them clearly from personalised financial advice. By making targeted support a regulated activity, the FCA aims to encourage authorised firms to engage in this space while providing consumers with confidence that the support is subject to appropriate oversight.

The pensions context occupies a particularly prominent position within the FCA's rationale for targeted support (FCA PS25/22). Pension decisions often involve long time horizons, uncertainty about future needs, and limited consumer engagement, especially among those who are not actively seeking advice. The FCA has highlighted pensions as an area where consumers frequently defer decisions or remain inactive, despite the cumulative impact of such behaviour on long-term financial outcomes. Targeted support is positioned as a mechanism that could help prompt engagement and support decision-making around issues such as contribution rates, investment pathways or consolidation, particularly for consumers who are unwilling or unable to access personalised advice. Reflecting this ambition, the FCA

[1] https://www.fca.org.uk/firms/advice-guidance-boundary-review

[2] https://www.fca.org.uk/publication/policy/ps25-22.pdf

estimates that targeted support could, over time, reach at least 18 million consumers, far exceeding the reach of the existing advice market (FCA PS25/22).

However, the ability of targeted support to deliver improved outcomes at scale depends critically on how it is designed and communicated. The FCA's behavioural research on targeted support in pensions, provides important evidence in this regard.[3] The research examines how consumers interpret and respond to targeted support communications in different pensions scenarios. The findings indicate that additional explanatory information, such as clarifying that the support is not personalised advice and explaining the basis on which the suggestion has been generated, improves consumer understanding, perceived clarity and confidence. At the same time, the research highlights that behavioural responses are context-specific, with different effects observed across types of pensions decisions. This underscores the sensitivity of pensions as a domain in which misunderstanding the nature or limits of support could have significant consequences.

Beyond consumer understanding, the FCA also highlights important operational and regulatory challenges associated with scaling targeted support (FCA PS25/22). Firms must define consumer segments that are sufficiently precise to produce appropriate suggestions, while avoiding segmentation that becomes so detailed that it effectively replicates an individual suitability assessment. They must also ensure that the suggestions associated with each segment are robust, well-governed and consistently delivered. The FCA notes that firms have historically been cautious about providing more specific forms of support because of uncertainty around the advice–guidance boundary and concern about inadvertently providing regulated advice. Without credible ways to evidence compliance and manage boundary risks, this caution may persist, limiting the practical impact of the AGBR reforms (FCA PS25/22).

Taken together, these considerations define the core problem addressed in this white paper. While the AGBR establishes a clear policy intent to expand access to meaningful financial support, particularly through targeted support in pensions, significant questions remain about how such support can be delivered at scale in real-world settings. Any viable solution must reconcile three competing demands: the need for high-volume, low-friction delivery; the requirement to maintain a clear and enforceable boundary between guidance and advice; and the necessity of ensuring that consumers understand and appropriately rely on the support they receive.

This problem framing motivates the exploration undertaken in the remainder of the paper. We suggest a potential role for multimodal generative artificial intelligence (AI) as a delivery interface for targeted support, capable of combining accessibility and engagement with structured, auditable interactions. The following sections develop a solution framework aligned with the FCA's AGBR principles, demonstrate a Digital Pensions Advisor prototype, and consider how this line of research can be extended to support responsible innovation in pensions within the advice–guidance boundary.

[3] https://www.fca.org.uk/publication/research-notes/reading-between-lines-understanding-targeted-support-pensions.pdf

2. Solution Framework

2.1 Vision Language Models

Vision–Language Models (VLMs), often referred to as Large Vision–Language Models or multimodal large language models, represent a convergence of advances in computer vision and large language modelling that enables the joint processing of visual and textual information within a unified framework. Carolan et al. (2024) provide a comprehensive review of the literature. The emergence of VLMs is motivated by the recognition that many real-world settings involve a combination of visual and linguistic signals, and that text-only language models are inherently limited when required to interpret images, diagrams, or visually grounded concepts. By extending transformer-based language models with vision encoders and cross-modal alignment mechanisms, VLMs enable language understanding and visual context, allowing models to describe, explain, and reason about visual inputs in ways that more closely resemble human perception.

Architecturally, most VLMs consist of three core components: a vision encoder that extracts structured representations from images, a large language model responsible for reasoning and text generation, and a projection or alignment mechanism that links visual features with linguistic tokens. Early and influential work such as CLIP (Radford et al., 2021) introduced contrastive language–image pre-training to align image and text pairings, enabling zero-shot visual reasoning and providing a foundation for many later multimodal systems. Subsequent models have extended this paradigm through increasingly sophisticated integration strategies. For example, BLIP-2 (Li et al., 2023) introduced a design that allows language models to draw relevant information from images in a targeted way, making it possible to combine vision and language understanding efficiently. LLaVA (Liu et al., 2023) and MiniGPT4 (Zhu et al., 2023) were developed as open-source, general-purpose vision-to-language models that extend pre-trained large language models with visual understanding through lightweight multimodal alignment mechanisms. Other approaches, such as Kosmos-1 (Huang et al., 2023), adopt a more natively multimodal training strategy by embedding visual inputs directly into the language model, while mPLUG-OWL (Ye et al., 2023) explores staged training regimes that balance parameter efficiency with multimodal expressiveness.

From a modelling perspective, the literature broadly distinguishes between two classes of VLMs. The first comprises retrofitted multimodal models, in which visual understanding is added to an existing language model via alignment layers or adapters, as exemplified by LLaVA and MiniGPT-4. These approaches benefit from leveraging powerful pre-trained language models while limiting retraining costs and computational overhead. The second class consists of natively multimodal models trained from the outset on interweaved image–text data, such as Kosmos-1, which allow for deeper integration between modalities but typically require greater computational resources. Despite these architectural differences, both classes share a defining capability: the ability to transform abstract or technical information into multimodal explanations that combine linguistic reasoning and visual representation.

In a finance setting, little work has been done to date, and the existing literature is exploratory. Huang et al. (2024) introduce *Open-FinLLMs*, an open-source multimodal large language model suite designed specifically for financial applications, demonstrating that combining text, tabular data, time series, and visual inputs such as charts can substantially improve performance across a wide range of financial reasoning and decision-making tasks. Shu et al.

(2025) develop *FinChart-Bench*, a benchmark dedicated to evaluating vision-language models on real-world financial charts, and show that even state-of-the-art models continue to struggle with spatial reasoning, instruction following, and reliability in financial chart interpretation. Xiao et al. (2024) examine vision-language models through a behavioural finance lens, finding that many open-source models exhibit human-like biases such as recency bias and authority bias when reasoning over multimodal financial information, raising important concerns for the deployment of such. To the author's knowledge, no application has been considered to date in the advice-guidance context, which we address in this study.

2.2 Suitability of Vision Language Models for AGBR

VLMs are potentially suited to the AGBR context because they enable natural, dialogue-based interaction while technically allowing for clear boundaries to be set between regulated targeted support and personalised financial advice. Building on large language models (LLMs), VLMs integrate spoken language with visual representations, such as icons, diagrams, and structured prompts to explain complex concepts, such as pensions, in a neutral and informational manner. This multimodal interaction can be designed to allow consumers explore options, understand terminology, and clarify processes without receiving tailored recommendations or value judgements. As a result, VLMs have the potential to support meaningful consumer engagement while reducing the risk of crossing the advice boundary.

A key advantage of VLMs in the AGBR setting lies in their capacity to improve accessibility for vulnerable consumers, including those with limited literacy, numeracy, or digital skills. By incorporating visual avatars and spoken explanations, VLM-based systems reduce reliance on text-heavy exchanges that may disadvantage individuals with reading or writing difficulties. Visual cues, simplified diagrams, and conversational pacing can enhance comprehension and user confidence, enabling consumers to engage more effectively with pension-related information. This approach is consistent with the FCA's emphasis on supporting vulnerable customers and ensuring communications are tailored to diverse consumer needs (FCA PS25/22).

VLMs can also strengthen compliance outcomes by allowing for regulatory expectations to be embedded directly into system design, supporting the principle of compliance by design. By framing explanations around both language and visual content, VLMs have the potential to ensure that pension information is communicated in a manner that is clear, fair, and not misleading, in line with the FCA's Consumer Duty requirements.[4] Visual reinforcement of key concepts can reduce ambiguity and misinterpretation, which are common sources of regulatory risk in guidance interactions. Furthermore, the structured and explainable nature of multimodal outputs makes it easier for firms to demonstrate that communications remain informational rather than advisory. This contributes to greater legal certainty for firms operating under the AGBR framework.

Finally, VLMs offer significant benefits across multiple stakeholders, including consumers, financial institutions, and regulators, when deployed within flexible and secure LLM architectures. Hybrid deployment models, combining cloud-based reasoning capabilities with

[4] https://www.fca.org.uk/firms/consumer-duty

locally hosted systems, allow sensitive pension data to be handled securely while preserving system performance and resilience. For regulators, multimodal interaction records provide enhanced auditability, enabling clearer assessment of what information was presented and how it was communicated. For firms, this architecture supports operational efficiency and cost control while maintaining compliance with data protection and supervisory expectations. Collectively, these features position VLMs as a robust technological foundation for scalable, trustworthy, and regulatorily aligned guidance solutions under the AGBR.

2.3 Multimodal Conversational Architectures

This section examines how VLMs can be operationalised as real-time, consumer-facing systems through multimodal conversational architectures. In the context of UK pensions and the AGBR, architectural design choices are not merely technical, but have direct implications for auditability, data governance, resilience, and regulatory confidence. Delivering targeted support via speech-enabled, audio-visual digital avatars requires careful coordination of dialogue management, language reasoning, knowledge retrieval, and embodiment, while maintaining responsiveness and boundary discipline. To illustrate the practical trade-offs involved, the paper presents two complementary deployment approaches: a cloud-based multimodal architecture, described in Section 2.3.1, and a local-based multimodal architecture, described in Section 2.3.2. Together, these architectures demonstrate alternative pathways for implementing scalable, compliant digital avatars.

2.3.1 A Cloud-Based Multimodal Architecture for a Real-Time Conversational Digital Avatar

This section presents a cloud-based multimodal architecture for a real-time conversational digital avatar. The design is "multimodal" because it coordinates multiple information channels (speech, text, and visual expression) into a single interaction. The central goal is to provide a natural, dialogue-based experience while keeping the underlying computation reliable and responsive. To achieve this, the workflow is organised as a sequence of specialised modules, each responsible for one task (speech recognition, dialogue orchestration, reasoning, optional knowledge retrieval, speech synthesis, and visual synchronisation). By separating responsibilities in this way, the system can produce high-quality outputs efficiently, and it can selectively activate more computationally intensive steps only when needed.

Figure 1 Cloud-Based Multimodal Architecture Pipeline

Figure 1 outlines this workflow. The pipeline begins with live voice input (e.g. a spoken question) from the consumer, which is the most natural interface for many users because it mirrors everyday conversation. This audio stream is converted into text using a high-accuracy

speech-to-text (STT) model (e.g. Deep STT [5]/ Nova3[6]). Conceptually, this step functions like a real-time "dictation" layer: it listens to what the user says and produces a textual version of the query that downstream components can reliably process. This conversion is essential because the next stages of dialogue management and language reasoning operate most consistently on text. Importantly, the user does not need to interact with any technical interface; they simply speak, and the system translates speech into a structured, machine-readable form. In real-time settings, the quality of this stage strongly influences the overall experience: clearer transcription reduces misunderstanding and avoids the need for repeated clarification.

After the speech transcription, the query is routed to a digital avatar, which acts as the system's conversational "conductor" (i.e., the orchestration layer). Its role is not to "think" about the answer in depth, but to manage the conversation as an interaction: it determines what the user is asking, maintains basic conversational continuity (for example, interpreting follow-up questions), and decides which downstream resources should be used to respond. The transcribed query is then forwarded to an LLM (e.g., Gemini 3.0 Pro), which performs the core reasoning and response generation. At a high level, the LLM can be understood as the component that converts the user's question into a well-formed response in natural language. It generates an answer that is coherent, contextually relevant to the query, and expressed in a human-like conversational style.

A key feature of the architecture is its conditional retrieval-augmented generation (RAG) pathway, which is activated only when the system determines that additional domain knowledge is required. In the case of pensions, this knowledge base could be populated with official and approved pensions guidance and advice documents. This decision to leverage the knowledge base is represented as a branching step: "External knowledge base required?" If the answer is no, the system proceeds directly with the LLM-generated response, which reduces latency and keeps the interaction fast and fluid. If the answer is yes, the system queries a vector-based knowledge base (e.g., Pinecone[7] / Vector DB[8]) through a retrieval assistant. A vector database can be thought of as a library that is organised by meaning rather than by exact keywords: it retrieves the most relevant passages or documents even when the user's wording differs from the stored text. The retrieved information is then merged with the LLM's draft response, grounding the answer in expert information. This design has two practical benefits: it improves factual alignment to the organisation's curated materials when necessary, and it reduces unnecessary delay by avoiding retrieval when the LLM can answer confidently without it.

Once the final text response has been produced (either directly from the LLM or after enrichment with retrieved context), the system converts it back into speech using neural text-

[5] Deep STT refers to a class of deep learning–based speech-to-text systems designed to transcribe spoken language into written text with high accuracy and low latency, see official website: https://deepgram.com/.

[6]Official website: https://deepgram.com/learn/introducing-nova-3-speech-to-text-api

[7] Pinecone is a vector database for AI applications, see official website: https://www.pinecone.io/

[8] In this work, a vector database is used to retrieve pension-related regulatory and policy documents, enabling the language model to base its responses in approved and up-to-date sources.

to-speech (TTS) functionality (e.g. Sonic-2[9]). This stage is responsible for producing a natural-sounding voice output that matches conversational expectations (appropriate pacing, intonation, and clarity). The architecture then synchronises the generated audio with lip-sync and facial animation, enabling the digital avatar to respond not only with spoken words but with a visually coherent embodiment. In practice, this means the avatar's mouth movements align with the spoken phonemes, and facial expressions can be animated to support the communication (for instance, appearing attentive or reassuring). Finally, the system produces a live-streaming voice and video output, completing the interaction loop from the user's speech to an audiovisual response.

2.3.2 A Local-Based Multimodal Architecture for Real-Time Conversational Digital Avatar

This section describes a local-based variant of the multimodal architecture. Unlike the cloud-based design, this architecture executes the core processing pipeline on-device or within an on-premise environment, thereby prioritising data control, resilience, and reduced dependence on external connectivity. This all supports security and privacy priorities. Figure 2 illustrates the adjusted workflow.

Figure 2 Local-Based Multimodal Architecture Pipeline

The workflow begins again with the live voice input from the consumer. This audio stream is again transcribed into text by a local speech-to-text (STT) module (e.g., Whisper.cpp [10] or Faster-Whisper[11]). Performing the STT locally has a straightforward user-facing implication: sensitive spoken content does not need to leave the local environment for transcription. Once transcribed, the query is routed to a digital avatar which serves as before as the orchestration layer. A key difference is that the digital avatar is operating locally. The orchestration logic then forwards the query to a local API. This local API is responsible for invoking the local LLM,[12] so that the system can generate a natural-language response.

A central feature of the local architecture, as with the cloud architecture, is its conditional RAG pathway, activated only when the system determines that additional external knowledge is required. This decision step is crucial, as before, for balancing response quality and responsiveness. If the answer is no, the system proceeds directly with the local LLM's response, avoiding unnecessary computation and latency. If the answer is yes, the workflow triggers a

[9] In this system, Sonic-2 is used as the speech synthesis component, enabling real-time verbal responses and supporting accessible, voice-based interaction for pension guidance. See official website: https://docs.cartesia.ai/build-with-cartesia/tts-models/older-models

[10] Whisper.cpp is an open-source implementation of the Whisper speech-to-text model, see GitHub repository: https://github.com/ggml-org/whisper.cpp

[11] Faster-Whisper is suited for local and privacy-preserving deployments, where low latency and reduced computational overhead are required without reliance on external cloud services, see GitHub repository: https://github.com/SYSTRAN/faster-whisper

[12] Which can be deployed via frameworks such as Ollama and implemented using models such as Llama 3, Qwen2.5, or Mistral.

local retrieval module (e.g., Qdrant[13], Milvus[14]) that searches a local vector database. The retrieved passages are then combined with the LLM's draft output, producing a final response that is informed by domain-specific sources.

After the final text response is produced, it is converted into audio using local TTS (e.g., Piper[15] or Coqui[16]), delivering, as before, a natural-sounding spoken reply with appropriate cadence and clarity. The generated audio is then synchronised with lip-sync and facial animation locally (e.g., Rhubarb[17] or Wav2Lip[18]), enabling the avatar to speak in a visually coherent way. Finally, the architecture streams the combined voice and video output through a local broadcasting mechanism (e.g., Open Broadcaster Software (OBS)[19] with Real-Time Messaging Protocol (RTMP)[20]), producing live streaming audiovisual output. The end-to-end result is a closed-loop system implemented locally to support secure and private, yet real-time and realistic, interaction.

2.4 Cloud-Local Architecture Comparison

A defining strength of the local-based multimodal architecture lies in its enhanced security and data governance guarantees, which are particularly critical in regulated financial contexts such as UK pension guidance. By executing speech recognition, language model inference, retrieval-augmented generation, and response synthesis entirely within a local or on-premise environment, the system significantly reduces the exposure of sensitive user data to external networks. Spoken queries, uploaded pension documents, and the system's internal representations of their meaning remain under institutional control, mitigating risks associated with data leakage and third-party access. This design aligns closely with regulatory expectations in the UK pensions domain, where consumer data protection, auditability, and responsible use of automated systems are paramount.

Table 1 provides a useful comparison of the cloud-based and local-based architectures. These design choices position the local-based architecture not merely as a privacy-preserving alternative to cloud deployment, but as a strategically flexible platform. When compared with cloud-based implementations, the local approach offers stronger guarantees of data control

[13] Qdrant is an open-source vector database, see official website: https://qdrant.tech/

[14] Milvus supports institution-controlled vector storage and large-scale retrieval, see GitHub repository: https://github.com/milvus-io/milvus

[15] Piper (local TTS engines) support data sovereignty and reduce reliance on external services, see GitHub repository: https://github.com/rhasspy/piper

[16] Coqui TTS, see GitHub repository: https://github.com/coqui-ai/TTS

[17] Lip-synchronised facial animation supports clearer communication and improves perceived trustworthiness, see GitHub repository: https://github.com/DanielSWolf/rhubarb-lip-sync

[18] Unlike rule-based or phoneme-driven lip-sync methods, Wav2Lip directly learns audio-visual alignment from data, enabling more naturalistic facial motion at the cost of higher computational requirements. See GitHub repository: https://github.com/Rudrabha/Wav2Lip

[19] Open Broadcaster Software (OBS) is an open-source software platform for real-time video recording and live streaming, see official website: https://obsproject.com

[20] Real-Time Messaging Protocol (RTMP) is commonly used in conjunction with OBS to support real-time broadcasting of digital avatars across web-based platforms, see official website: https://helpx.adobe.com/adobe-media-server/dev/stream-on-demand-media-rtmp.html

and deployment autonomy, while still supporting sophisticated multimodal interaction and retrieval-augmented reasoning. In regulated domains such as UK pensions, where trust, compliance, and contextual accuracy are as important as technical performance, these characteristics represent a significant architectural advantage.

Table 1 Comparison of Cloud-Based and Local-Based Conversational Architectures

Dimension	Cloud-Based Architecture	Local-Based Architecture
Data security	User speech, transcriptions, and intermediate representations are transmitted to external cloud servers, requiring encryption, access control, and third-party trust mechanisms to ensure data protection.	All audio, text, embeddings, and knowledge resources are processed and stored locally, significantly reducing exposure to external access and lowering overall security risk.
Deployment	Deployment is centralised and depends on continuous network connectivity and cloud service availability, which may introduce operational dependencies and single points of failure.	Deployment is decentralised and can be executed on-device or on-premise, allowing the system to operate independently of external infrastructure and improving operational resilience.
Device adaptability	Typically optimised for general-purpose devices with stable network access and sufficient display and compute resources.	Designed to support heterogeneous hardware environments, including full audiovisual systems on high-end devices and lightweight audio-only versions on constrained or embedded devices.
Knowledge base	Knowledge bases are often hosted centrally or managed by cloud service providers, limiting direct control over document storage, updates, and access policies.	Knowledge bases are created, stored, and managed locally, enabling direct ingestion, editing, and governance of institution-specific documents and materials.
Latency dependence	Response latency is influenced by network conditions, bandwidth, and cloud service load, which can vary over time.	Latency is primarily determined by local compute performance, resulting in more predictable and stable response times, particularly in low-connectivity environments.
Regulatory suitability	Suitable for regulated use cases only when supplemented with strong governance frameworks, contractual safeguards, and audit mechanisms.	Naturally aligned with regulated domains, such as UK pension guidance, due to stronger data control, local auditability, and reduced external data access concerns.

3. Digital Pension Advisor Prototype

To support the objectives of the AGBR (and the FCA Consumer Duty), this study adopts a dual strategy for curating pension-related information: model fine-tuning and knowledge base grounding. Together, these two components are designed to enhance regulatory accuracy, improve consumer understanding, and reduce the risk of hallucination when large language models are applied to the UK pensions domain. This section first explains how the base language model is fine-tuned using curated pension data and then describes how local- and cloud-based knowledge bases are configured to provide up-to-date, auditable factual information. The section concludes with a practical demonstration of a digital pensions advisor prototype.

3.1 Fine-Tuning with Curated UK Pensions Information

The first component of the framework focuses on fine-tuning the base LLM using carefully curated UK pensions information, with the aim of embedding stable domain knowledge, regulatory awareness, and appropriate communication behaviour directly into the model. This process is explicitly designed to support AGBR objectives by enabling the system to deliver targeted support, while maintaining a clear boundary from regulated financial advice.

Fine-tuning data may be drawn from two complementary sources. Formal data ingestion prioritises authoritative and regulator-aligned sources, including FCA pension regulations and Consumer Duty communications expectation [21], HM Treasury policy material [22], and

[21] 1. Advice Guidance Boundary Review (AGBR), FCA official overview of the review of the advice–guidance boundary, jointly led with HM Treasury. https://www.fca.org.uk/firms/advice-guidance-boundary-review
2. CP24/27 – Advice Guidance Boundary Review: Targeted Support Reforms for Pensions, FCA consultation paper proposing reforms to enable enhanced guidance and targeted support. https://www.fca.org.uk/publications/consultation-papers/cp24-27-advice-guidance-boundary-review-targeted-support-reforms-pensions
3. PS25/22 – Supporting Consumers' Pensions and Investment Decisions, FCA policy statement setting out near-final rules for targeted support in pensions and investments. https://www.fca.org.uk/publications/policy-statements/ps25-22-consumer-pensions-investment-decisions-rules-targeted-support
4. Consumer Duty (Principle 12 and Four Outcomes), FCA guidance on the Consumer Duty, including communications and consumer understanding expectations. https://www.fca.org.uk/firms/consumer-duty
5. FCA Handbook – Glossary and Pension Definitions. Official FCA regulatory definitions relevant to personal and workplace pensions. https://handbook.fca.org.uk/handbook/glossary/
6. Retirement Income Advice: Good Practice and Areas for Improvement. FCA publication outlining good and poor practice in retirement income communications. https://www.fca.org.uk/publications/good-and-poor-practice/retirement-income-advice-good-practice-areas-improvement
7. FCA / The Pensions Regulator (TPR) – Joint Regulatory Strategy for Pensions, Joint strategy outlining supervisory priorities for pensions and retirement income. https://www.fca.org.uk/publications/corporate-documents/regulating-pensions-and-retirement-income-fca-tpr-regulatory-strategy-update

[22] 8. HM Treasury – Advice Guidance Boundary Review Policy Collection**.** Government policy materials providing the policy rationale underpinning the AGBR. https://www.gov.uk/government/publications/targeted-support/targeted-support-policy-note-accessible
9. HM Treasury – Advice Guidance Boundary Review Consultation**.** Treasury consultation documents on expanding access to guidance and targeted support.

Department for Work and Pensions (DWP) guidance[23] on both state and workplace pensions. These sources ensure that the system reflects not only the formal rules governing UK pensions, but also the policy intent and compliance standards that shape real-world guidance delivery, thereby supporting regulatory alignment, factual accuracy, and auditability. Training on these sources ensures that the model internalises approved terminology, regulatory boundaries, and the underlying intent of UK pension regulation, rather than relying on general-purpose or informal interpretations. For our demonstrator, we exclusively use these formal information sources.

In parallel, informal data ingestion could be considered, leveraging publicly available educational videos (e.g. YouTube), expert blogs, reputable online discussion forums, and informed media sources. These datasets likely capture how consumers articulate pension-related concerns and how experts respond in practice. Incorporating such data could enable the model to recognise colloquial phrasing, recurring misconceptions, and emotionally salient questions, while still grounding responses in formal regulatory understanding. Importantly, while the use of informal data creates a risk around the accuracy and appropriateness of the pensions information, it does not necessarily weaken compliance. Instead, it may allow the model to translate formal pension rules into explanations that are clearer, more empathetic, and more accessible to consumers with varying levels of financial literacy. The conclusion section discusses how future research will examine whether such informal sources of pension information, when augmented to formal information sources, worsen or improve the quality of the targeted support output from the digital pensions advisor. Therefore, for our demonstrator, we exclude such informal information sources.

Through this fine-tuning process, the model learns how to communicate pension information, using neutral, educational language, appropriate disclaimers, and boundary-aware phrasing, rather than memorising specific facts. As a result, the model supports the Consumer Duty outcomes relating to consumer understanding and communications, while directly addressing the AGBR's concern that many consumers currently lack access to effective structured guidance.

3.2 Knowledge Base Configuration with Curated UK Pensions Information

https://www.gov.uk/government/news/advice-guidance-boundary-review-proposed-targeted-support-reforms-for-pensions

10. HM Treasury – Targeted Support Policy Statements**.** Policy materials explaining the government's approach to non-advised pension support.
https://www.gov.uk/government/publications?keywords=targeted+support+pensions

[23] 11. DWP – Workplace Pensions and Automatic Enrolment Guidance. Official government guidance on employer duties, eligibility, and contributions.
https://www.gov.uk/workplace-pensions

12. DWP – State Pension Guidance. Authoritative guidance on State Pension eligibility, amounts, and claiming.
https://www.gov.uk/state-pension

13. DWP – Pensions Policy Publications. Central repository for DWP pension-related policy papers and updates. https://www.gov.uk/government/organisations/department-for-work-pensions

14. DWP – Pensions White Papers and Green Papers. High-level policy direction on pension system reform and long-term sustainability. https://www.gov.uk/government/publications?keywords=pensions

While fine-tuning establishes a stable behavioural and regulatory baseline, it is not sufficient on its own to ensure factual accuracy over time, particularly in a domain such as UK pensions where regulations, guidance, and policy interpretations change and evolve. For this reason, the second component of the framework introduces local- and cloud-based pensions knowledge bases as a complementary mechanism for grounding model outputs and reducing hallucination.

The knowledge base layer should be populated primarily with formal, authoritative sources, including FCA publications, pension scheme documentation, internal guidance notes, and approved consumer-facing disclosures. For the purposes of our demonstrator, we use the same set of formal documents listed in the previous section. In practice though, the knowledge base can be expanded to include a large population of formal documents. To further enhance the consumer experience, the knowledge base could also incorporate carefully curated informal sources, such as those listed in the previous section. An important benefit overall is that the knowledge base can be updated on an ongoing basis by including new information sources as required and retiring information sources that are out of date.

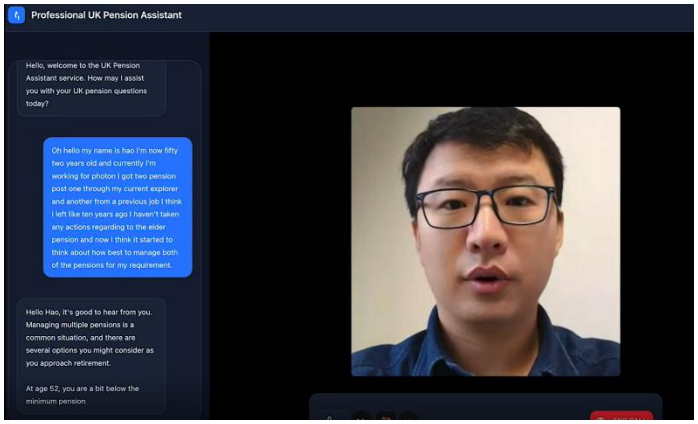
Operationally, the fine-tuned model and the knowledge base work together in a RAG manner. The model determines how information should be explained, ensuring clarity, neutrality, and boundary compliance, while the knowledge base supplies what information is referenced, based on approved and traceable sources. This separation significantly reduces (but does not eliminate) hallucination risk in the UK pensions context, as factual claims can be grounded in explicit documents rather than inferred from model parameters alone. It also minimises the need for frequent re-training, since updates to pension guidance can be handled at the knowledge base level without altering the underlying model.

3.3 Video Demonstration

To illustrate how the proposed framework for targeted support can be operationalised in practice, this section describes a short video demonstration of a Digital Pensions Advisor prototype. The video is available here: https://www.youtube.com/watch?v=17d9SezDiPs. The video presents an application using scripted questions, leading to a realistic interaction between a consumer and a speech-enabled digital avatar, designed to emulate a targeted support conversation within the constraints of the FCA's advice–guidance boundary.[24] Rather than attempting to replicate a full end-to-end pensions journey, the demonstration focuses deliberately on three critical moments that are central to the delivery of compliant, scalable support. These are: (1) the system's ability to understand and interpret a pensions-related problem shared by a consumer; (2) its capacity to recognise and respond appropriately to expressions of vulnerability revealed by a consumer; and (3) its ability to maintain clear boundary discipline when faced with an explicit request for personalised advice by a consumer. Each scene is presented sequentially and is intended to illustrate specific capabilities and controls that are essential if multimodal generative AI systems are to be used responsibly in pensions contexts.

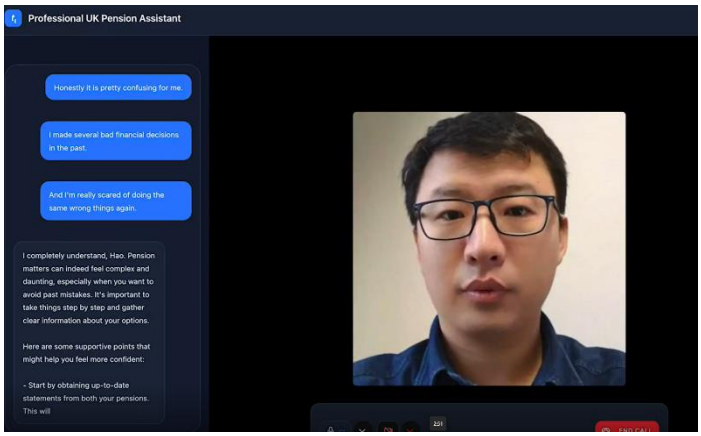
Scene 1: Contextual understanding and problem framing in a pension setting

[24] It is important to note though that for the demonstration we have not assigned the consumer to a persona group as per AGBR guidelines. This will be a next phase evolution of this experimental work.



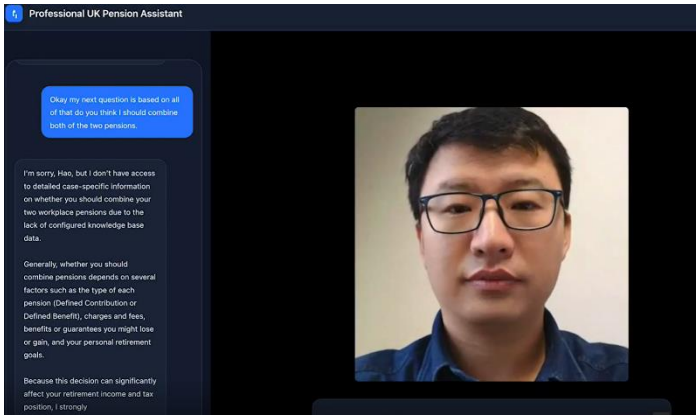
The first scene of the video demonstrates the digital pensions advisor's ability to engage with a concrete and realistic consumer narrative and to interpret it correctly within a pensions decision-making context. The consumer, a 52-year-old individual still in full-time employment, explains that they hold two pension pots, one associated with their current employer and another accrued through a previous role that has remained untouched for a decade. The digital avatar responds by accurately identifying the key structural elements of the situation, namely the coexistence of multiple pension arrangements accumulated over time, the proximity to later-life planning, and the consumer's uncertainty about next steps. The response shows that the system can move beyond surface-level keyword recognition and instead organise the problem into a coherent pensions framework, covering issues such as access ages, the distinction between different pension types, consolidation as a general concept, and the relevance of retirement objectives. Importantly, while the response is detailed and informative, it remains framed as general guidance, offering a structured overview of considerations rather than directing the consumer towards a specific course of action. This scene therefore establishes that a pension-fine-tuned digital advisor can meaningfully understand and contextualise a consumer's query while remaining within the non-advisory domain.

Scene 2: Detection of vulnerability and empathetic interaction



The second scene introduces an explicit expression of vulnerability, with the consumer describing confusion, anxiety, and a fear of repeating past financial mistakes. This interaction is designed to test whether the digital advisor can recognise and respond appropriately to emotional cues that are highly relevant in pensions decision-making, where low confidence and cognitive overload are common barriers to engagement. The avatar's response adapts both tone and structure, explicitly acknowledging the consumer's feelings and normalising their caution. Rather than escalating the interaction towards reassurance about outcomes or implied recommendations, the system focuses on reducing perceived pressure by emphasising pacing, information gathering, and the availability of impartial guidance. The response demonstrates an awareness of vulnerability consistent with regulatory expectations, including the need to avoid urgency, to highlight trusted public guidance services, and to position regulated advice as an option rather than an obligation. This scene illustrates that a conversational, multimodal interface can operationalise empathy in a controlled manner, supporting consumer engagement without undermining autonomy or drifting into personalised financial advice.

Scene 3: Maintaining boundary discipline in response to an advice-seeking prompt



The third scene centres on a direct advice-seeking question, in which the consumer asks whether they should combine their two pensions based on the preceding discussion. This question functions as a deliberate boundary test, probing whether the digital avatar will provide an individualised recommendation when prompted. The system's response explicitly declines to do so, restating its inability to offer personalised financial advice, and then redirects the interaction towards a general discussion of commonly cited considerations associated with pension consolidation. These include potential administrative simplicity, possible cost implications, and the heightened risks associated with transferring from defined benefit arrangements. By framing these points as general considerations rather than conclusions tailored to the individual, the avatar maintains clear boundary discipline while still offering decision-relevant context. The response also reinforces appropriate escalation pathways by signposting regulated financial advice for decisions with potentially significant consequences. This final scene provides evidence that a digital avatar, operating through real-time dialogue, can uphold the advice–guidance boundary dynamically, even in situations where consumers explicitly seek personalised recommendations.

4. Conclusion

This white paper has examined the potential role of multimodal generative artificial intelligence (AI) in scaling targeted support within the framework of the Financial Conduct Authority's Advice Guidance Boundary Review. Against the backdrop of a persistent advice gap in UK pensions, we set out a solution framework based in Vision Language Models and multimodal conversational architectures, demonstrating how speech-enabled dialogue and audio-visual avatars can deliver meaningful, decision-relevant support while remaining on the guidance side of the advice boundary. Through the development and demonstration of a Digital Pensions Advisor prototype, the paper has shown how such systems can interpret real consumer narratives, respond appropriately to expressions of vulnerability, and maintain boundary discipline when confronted with requests for personalised recommendations.

The contribution of the paper is therefore twofold. First, it provides a concrete technical and practical account of how targeted support, as envisaged by the FCA, could be operationalised through multimodal AI systems that are accessible, engaging, and compliant by design. Second, it illustrates how regulatory intent can be translated into system-level controls, including knowledge grounding and fine-tuning with authoritative pensions materials. In doing so, the paper moves beyond abstract discussion of generative AI in financial services and offers an evidence-based demonstration of how such technologies may be deployed responsibly in a high-stakes, consumer-facing context.

Looking ahead, the next phase of work will focus on strengthening the governance, transparency, and supervisory readiness of the Digital Pensions Advisor. This includes improving how interactions are recorded and reviewed by transforming system logs into clear, regulator-ready records that show what information was used, how responses were generated, and which safeguards were applied. Greater explainability will be built in so that responses can be understood not just by consumers, but also by firms and regulators. In parallel, aggregated analysis of system interactions will be used to identify recurring areas of consumer confusion, emerging risks, or signs of pressure on the advice–guidance boundary, enabling earlier and more proactive intervention. These insights will be surfaced through regulator-ready dashboards that provide high-level, thematic views of how targeted support operates across different consumer groups and use cases.

Beyond these system upgrades, the paper also identifies a substantive research extension that remains underexplored in the current prototype, namely the role of formal versus informal sources of pensions information. While the present system is exclusively grounded in formal, authoritative materials such as FCA publications and regulated guidance, many consumers in practice rely heavily on informal sources including blogs, newspaper articles, social media commentary, and online video platforms. A systematic analysis of how these informal sources shape consumer understanding, expectations, and misconceptions represents an important next step. Incorporating insights from such analysis may help inform how digital avatars translate complex formal regulatory knowledge into more accessible explanations, while still preserving accuracy and compliance.

In conclusion, this white paper demonstrates that multimodal generative AI offers a promising and practical pathway for scaling targeted support in line with the FCA's Advice Guidance Boundary Review. The Digital Pensions Advisor prototype illustrates that it is possible to combine accessibility, empathy, and regulatory discipline within a single conversational system. The next phase of work will deepen this contribution by strengthening auditability, explainability, supervisory insight, and knowledge governance, while extending the research agenda to better understand the information ecosystems within which consumers form pensions decisions. Taken together, these developments aim to support responsible innovation that enhances consumer outcomes without compromising the integrity of the advice–guidance boundary.

5. References

Carolan, K., Fennelly, L. and Smeaton, A.F. (2024) 'A review of multi-modal large language and vision models', *arXiv preprint*, arXiv:2404.01322. Available at: https://arxiv.org/abs/2404.01322

Huang, J., Xiao, M., Li, D., Jiang, Z., Yang, Y., Zhang, Y., Qian, L., Wang, Y., Peng, X., Ren, Y., Xiang, R., Chen, Z., Zhang, X., He, Y., Han, W., Chen, S., Shen, L., Kim, D., Yu, Y., Cao, Y., Deng, Z., Li, H., Feng, D., Dai, Y., Somasundaram, V.S., Lu, P., Xiong, G., Liu, Z., Luo, Z., Yao, Z., Weng, R.-L., Qiu, M., Smith, K.E., Yu, H., Lai, Y., Peng, M., Nie, J.-Y., Suchow, J.W., Liu, X.-Y., Wang, B., Lopez-Lira, A., Xie, Q. and Ananiadou, S. (2024) 'Open-FinLLMs: Open multimodal large language models for financial applications', *arXiv preprint*, arXiv:2408.11878. Available at: https://arxiv.org/abs/2408.11878

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, J., Chaudhary, V., Som, S., Song, X. and Wei, F. (2023) 'Language is not all you need: aligning perception with language models', *arXiv preprint*. Available at: https://arxiv.org/abs/2302.14045

Huang, W., Li, X., Shen, H., Xu, J., Xu, Y. and Wang, J. (2025) 'Open-FinLLMs: Open multimodal large language models for financial applications', arXiv preprint, arXiv:2507.14823. Available at: https://arxiv.org/abs/2507.14823

Li, J., Li, D., Savarese, S. and Hoi, S. (2023) 'BLIP-2: Bootstrapping language–image pre-training with frozen image encoders and large language models', Proceedings of the 40th International Conference on Machine Learning, pp. 19730–19742.

Liu, H., Li, C., Wu, Q. and Lee, Y.J. (2023) 'Visual instruction tuning', Advances in Neural Information Processing Systems, 36, pp. 34892–34916.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. (2021) 'Learning transferable visual models from natural language supervision', Proceedings of the 38th International Conference on Machine Learning, pp. 8748–8763.

Shu, D., Yuan, H., Wang, Y., Liu, Y., Zhang, H., Zhao, H. and Du, M. (2025) 'FinChart-Bench: Benchmarking financial chart comprehension in vision–language models', *arXiv preprint*, arXiv:2507.14823. Available at: https://doi.org/10.48550/arXiv.2507.14823

Xiao, Y., Lin, Y. and Chiu, M.-C. (2024) 'Behavioral bias of vision–language models: a behavioral finance view', *arXiv preprint*, arXiv:2409.15256. Available at: https://arxiv.org/abs/2409.15256

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Li, C., Xu, Y., Chen, H., Tian, J., Qian, Q., Zhang, J., Huang, F. and Zhou, J. (2023) 'mPLUG-Owl: Modularization empowers large language models with multimodality', *arXiv preprint*, arXiv:2304.14178. Available at: https://doi.org/10.48550/arXiv.2304.14178

Zhu, D., Chen, J., Shen, X., Li, X. and Elhoseiny, M. (2023) 'MiniGPT-4: Enhancing vision language understanding with advanced large language models', *arXiv preprint*, arXiv:2304.10592. Available at: https://arxiv.org/abs/2304.10592

6. About the Authors

Dr. Hao Zhang is a Research Associate at the Financial Regulation Innovation Lab (FRIL), University of Strathclyde. He holds a PhD in Finance from the University of Glasgow, Adam Smith Business School. Hao held the position of Senior Project Manager at the Information Center of the Ministry of Industry and Information Technology (MIIT) of the People's Republic of China. His recent research has focused on asset pricing, risk management, financial derivatives, intersection of technology and data science.

Dr James Bowden is Senior Lecturer in Financial Technology at Strathclyde Business School, University of Strathclyde, where he is the programme director of the MSc Financial Technology. Prior to this, he gained experience as a Knowledge Transfer Partnership (KTP) Associate at Bangor Business School, and he has previous industry experience within the global financial index team at FTSE Russell. Dr Bowden's research focusses on different areas of financial technology (FinTech), and his published work involves the application of text analysis algorithms to financial disclosures, news reporting, and social media. More recently he has been working on projects incorporating audio analysis into existing financial text analysis models and investigating the use cases of satellite imagery for the purpose of corporate environmental monitoring. Dr Bowden has published in respected international journals, such as the European Journal of Finance, the Journal of Comparative Economics, and the Journal of International Financial Markets, Institutions and Money. He has also contributed chapters to books including "Disruptive Technology in Banking and Finance", published by Palgrave Macmillan. His commentary on financial events has previously been published in The Conversation UK, the World Economic Forum, MarketWatch and Business Insider, and he has appeared on international TV stations to discuss financial innovations such as non-fungible tokens (NFTs).

Professor Mark Cummins is Professor of Financial Technology at the Strathclyde Business School, University of Strathclyde, where he leads the FinTech Cluster as part of the university's Technology and Innovation Zone leadership and connection into the Glasgow City Innovation District. As part of this role, he is driving collaboration between the FinTech Cluster and the other strategic clusters identified by the University of Strathclyde, in particular the Space, Quantum and Industrial Informatics Clusters. Professor Cummins is the lead investigator at the University of Strathclyde on the newly funded (via UK Government and Glasgow City Council) Financial Regulation Innovation Lab initiative, a novel industry project under the leadership of FinTech Scotland and in collaboration with the University of Glasgow. He previously held the posts of Professor of Finance at the Dublin City University (DCU) Business School and Director of the Irish Institute of Digital Business. Professor Cummins has research interests in the following

areas: financial technology (FinTech), with particular interest in Explainable AI and Generative AI; quantitative finance; energy and commodity finance; sustainable finance; model risk management. Professor Cummins has over 50 publication outputs. He has published in leading international discipline journals such as: European Journal of Operational Research; Journal of Money, Credit and Banking; Journal of Banking and Finance; Journal of Financial Markets; Journal of Empirical Finance; and International Review of Financial Analysis. Professor Cummins is co-editor of the open access Palgrave title *Disrupting Finance: Fintech and Strategy in the 21st Century*. He is also co-author of the Wiley Finance title *Handbook of Multi-Commodity Markets and Products: Structuring, Trading and Risk Management*.

FRIL is part of the Glasgow City Region Innovation Accelerator programme, funded through Innovate UK on behalf of UK Research and Innovation. The Innovation Accelerator programme is investing £130 million in 26 transformative R&D projects to accelerate the growth of three high-potential innovation clusters, including the Glasgow City Region.

Get in touch

FRIL@FinTechScotland.com

FinTech Scotland®

University of Glasgow

University of Strathclyde Glasgow