AI as individualised persona: a useful addition to the economist's toolbox? The case of "Stephen Littlechild AI Agent"

VEPC Working Paper 2511
Bruce Mountain*, Shruti Kant
Initially released in November 2025, revised 8 January 2026

Abstract

This paper examines a retrieval-augmented generation (RAG) customisation of artificial intelligence platform ChatGPT, using papers written by economist Professor Stephen Littlechild from the 1960s to the present, to create "SCL AI Agent" (SCL). SCL seeks to replicate and apply the thinking of Professor Littlechild. Establishing the corpus of Professor Littlechild's papers, uploading it to ChatGPT and then instructing ChatGPT on how to understand that information and apply it, revealed the need for experimentation and learning-by-doing. Careful configuration sought to reduce hallucination and ensure well-informed responses delivered in Professor Littlechild's style. Assessment of SCL by regulatory professionals who have had long interaction with Professor Littlechild rated SCL highly, particularly in respect of "insight", "completeness" and "accuracy". These assessors were less convinced of SCL's ability to replicate Professor Littlechild's written style. However, if users provided SCL with context to their questions and information on the audience for its answers, SCL did deliver responses tailored to those audiences. SCL itself and uncustomised ChatGPT were asked to assess SCL's answers to the assessors' questions. They both agreed on SCL's superiority relative to uncustomised ChatGPT. SCL demonstrated a sophisticated, abstract understanding of Professor Littlechild's scholarship, although its ability to replicate his imagination is less clear and merits further research. Creating AI agents of other economists and setting them to critique each other's work could facilitate the more rapid dissemination of insight and understanding.

Keywords: AI agent, AI economic research, retrieval augmented generation

JEL Classifications: A11, C45, D83, I23, O33

*Corresponding author: bruce.mountain@vu.edu.au

The authors acknowledge, with thanks, Stephen Littlechild's comments, review and invaluable discussion during the preparation of this paper.

1. Introduction

There is increasing interest in how artificial intelligence (AI) can be useful to economists. Korinek (2023, 2025) identifies various applications of AI in economic research, mostly focussing on quantitative methods. Not yet canvassed in the economics literature is the prospect of adapting a generally available AI model to replicate and apply the thinking of a particular economist?[1] That is the purpose of the research described in this article. Chen et al. (2024) explore role-playing language agents, distinguishing between "demographic persona" (such as pilots or accountants), "character persona" (focused on well-established figures[2]) and "individualised persona". The third category describes a form of customised AI that seeks to take account of specific information on that person that is not easily available on the internet. The model described here fits into the third category.

We have developed our individualised persona agent through retrieval-augmented generation (RAG) customisation of ChatGPT.[3] RAG involves customising a large language model (LLM) [4] by providing additional knowledge outside of the LLM's training data, which knowledge is typically not available on the internet, before generating a response to users' prompts. As far as we know, ours is the first application of RAG for the development of an individualised persona agent in the field of economics.

The knowledge base that our RAG application draws on is the (mostly) publicly available set of papers written by Professor Stephen Charles Littlechild from the 1960s to the present. We call it "SCL AI Agent", or simply SCL. Asked to introduce Professor Littlechild, SCL responded: "*Professor Stephen Littlechild is internationally recognised as the architect of the RPI–X price cap model, the central regulatory innovation of the UK's utility privatisations in the 1980s, and as the first Director General of Electricity Supply (1989–1998) he applied this framework to reshape the electricity industry. His scholarship consistently emphasised the limits of regulation, the risks of capture and bureaucratisation, and the superiority of competitive markets and voluntary negotiation in discovering efficient outcomes. Rather than viewing regulation as a permanent substitute for markets, Littlechild saw it as a transitional and facilitating mechanism—an insight that underpins the design of the SCL AI Agent, which seeks to*

[1] There is one very recent exception. On 30 December 2025, Markus'Academy associated with Princeton University hosted a webinar featuring an AI rendering of John Maynard Keynes, grounded in Keynes' collected writings and audio. It was prepared by Professor Markus Brunnemeir and colleagues. Their AI rendering is very similar in intent to the AI agent described in this paper, and they also provide a speaking visual version of the AI agent. However, they have provided no information about the process involved in creating it. See: https://economics.princeton.edu/events/a-conversation-with-john-maynard-ai-keynes-about-ai-generated-by-ai/

[2] Deepai.org has produced publicly available AI agents (unanimated chatbots) of several well-known figures including Margaret Thatcher, George Orwell, Rasputin and Kamala Harris.

[3] ChatGPT is an artificial intelligence chatbot developed by the US company OpenAI and released in November 2022. OpenAI aims to develop "safe and beneficial" artificial general intelligence, which it defines as "highly autonomous systems that outperform humans at most economically valuable work". ChatGPT was designed "to interact in a conversational way. The dialogue format makes it possible for ChatGPT to answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests." (https://openai.com/index/chatgpt/) On November 6, 2023, OpenAI launched GPTs, allowing individuals to create customized versions of ChatGPT for specific purposes. A free version of ChatGPT is presently available, providing access to powerful models for general tasks, writing and learning, including web browsing and project management tools. Paid versions offer faster responses, higher capacity, and more advanced reasoning, priority access during busy periods, and specialized tools and features for intensive work. For further discussion, see https://en.wikipedia.org/wiki/OpenAI

[4] Amazon Web Services defines an LLM as "*very large deep learning models that are pre-trained on vast amounts of data. The underlying transformer is a set of neural networks that consist of an encoder and a decoder with self-attention capabilities. The encoder and decoder extract meanings from a sequence of text and understand the relationships between words and phrases in it*". (https://aws.amazon.com/what-is/large-language-model/#:~:text=help%20with%20LLMs?-,What%20are%20Large%20Language%20Models?,has%20approximately%2057%20million%20pages.)

extend his legacy by applying his reasoning style, sceptical posture, and preference for competition and consumer choice to contemporary regulatory-economic debates."

Many of Professor Littlechild's papers are not generally available, or available at all, on the internet. ChatGPT would not reveal which of Professor Littlechild's papers were included in its training and when we asked it to search for them on the internet we discovered that ChatGPT could only find about a quarter of the corpus of papers that we had assembled. By establishing a large corpus of papers and then instructing our customised/modified version of ChatGPT on how to respond to user prompts we sought to provide AI capability that is better informed, and more insightful in its ability to generate ideas and provide critique consistent with Professor Littlechild's scholarship, than is available through uncustomised ChatGPT. For brevity, future references to ChatGPT in this paper refer to uncustomised ChatGPT while SCL refers to the RAG modified form of ChatGPT that we progressively developed.

The paper proceeds by describing how Professor Littlechild's corpus of writing was established and uploaded and then how ChatGPT was selected and configured to form SCL. This is followed by the assessment of SCL by ourselves and others, and then a discussion of issues that arose in the assessment of SCL. The concluding section draws out the main points and suggests the focus of effort in future. The online Appendices A and B, available here[5], present examples of SCL's and ChatGPT's responses to a user's request (Appendix A); and how SCL's response varies when the user tells SCL of the intended audience for SCL's response to the user's request (Appendix B).

2. Technical background

Various large language models (LLMs) are available publicly and could have been used to develop SCL. OpenAI's "ChatGPT - o3" (subsequently superseded by "ChatGPT 5- Thinking") compares favourably (in terms of its ability to reason and act as agent) with its competitors (Korinek, 2025) and is accessible on the internet as are other comparable AI platforms.

ChatGPT, like other artificial intelligence (AI) chatbots, uses natural language processing to generate human-like conversation and text. It has been created by the AI research company OpenAI. Chat refers to its conversational interface, allowing users to interact with it naturally. The acronym GPT stands for Generative Pre-trained Transformer, which describes the core technology behind the model. Generative means the model can create new content (text, code, images, audio) in response to a prompt, rather than just pulling up existing information from a database like a traditional search engine. Pre-trained refers to the massive dataset of text from the internet, books, and articles to learn language patterns, grammar, and context before being fine-tuned for specific tasks like conversation. Transformer is a type of neural network architecture that allows the model to understand the context and relationships between words in a sentence or across a conversation, making the responses coherent and relevant.

A free version of ChatGPT is publicly available. We used a subscription version ("pro") so that we could customise it to develop SCL. OpenAI's "CustomGPT-o3" allows for the development of customised GPT applications. Its "Builder" chatbox - a software tool integrated into the Developer version of ChatGPT that assists in the creation, configuration, and testing of customised (RAG) GPT-based applications – was used in the configuration of SCL.

RAG uses "semantic search" to find relevant information from the files that have been uploaded to it. Semantic search in this context describes a method of searching through information in a way that understands the intent and context of a user's query, rather than just matching keywords. "Semantic search" differs from "keyword search" which is used in word processor software or document management systems (such as Google Drive) or structured query language (SQL) queries of relational databases. Unlike keyword search and structure query language, semantic search *does* take account of the intent and context of the user's query.

---

[5] https://www.vepc.org.au/_files/ugd/92a2aa_f6e643db7ba04ee39b4b19936176a1ba.pdf

The creation of a RAG application involves “chunking” (breaking the uploaded files into paragraphs or logical blocks), “embedding” (converting text chunks into numerical representations called “vectors”) and storage of the “vectors” in a vector database. In addition to the creation of the vector database, RAG involves the creation of instructions (commonly known as “configuration”) to govern how user prompts are to be understood, and how the additional information is to be used. Configuration is done by the developers, in this case us. Through chunking, embedding, vectorisation and storage SCL draws on the understanding it obtains from Professor Littlechild’s corpus, then uses ChatGPT’s “5-thinking” model to respond to users’ queries.

SCL exists as an application using ChatGPT’s “5-thinking” AI platform, with username and password access through the internet. It is not currently available publicly.

3. Establishing the corpus

Our initial exploration sought to find out what ChatGPT already knew about Professor Littlechild by asking ChatGPT to provide a list of [his works. ChatGPT was not able to do this but did respond to requests for references to Littlechild’s papers when sought over shorter periods, for example from 1975 to 1980. After repeated requests over successive periods, ChatGPT identified 74 documents that it said Professor Littlechild had written, some of which it could access and others of which it could not (many required subscription to journals or paid access to printed or digitised books).

An initial list of papers produced by Professor Littlechild, that he provided to us directly, identified more than 350 papers. ChatGPT was asked to write 150-word abstracts for each of those papers, which it did for almost all. It did this despite being able to identify only 74 documents. For each paper it also returned notes under the headings of “Key concepts”, “Empirical evidence”, “Regulatory stance”, “Influences/references”, “Verbatim gem”, and “Research gaps”. ChatGPT suggested these headings. The “Verbatim gem” we understand was intended to be a single sentence or phrase from each paper that would stand out as a pithy quote drawn from text in the paper. However, in several cases we found that ChatGPT invented these quotes.

Interrogation of the abstracts it had written was also revealing as to ChatGPT’s inventiveness (how could it write abstracts for papers it was not able to access?). The first paper in Professor Littlechild’s list of 350 was a 1966 working paper that he coauthored as a student at Northwestern University with his supervisor and another researcher. ChatGPT wrote a convincing abstract for the paper including that “*an eight-variable thermal-dispatch example solved on an IBM 7090 demonstrates an 80 per cent reduction in solution time relative to contemporary steepest-gradient methods and exhibits monotone convergence of both objective and shadow prices*”. However, when asked to produce the paper on which the abstract was based, it said that the paper was not available, and it could not find it. Asked then to explain how it was able to write the abstract, it said it “*wrote a plausible, inference-based abstract from the title and authors’ known contributions*”. Asked how it knew of the IBM 7090, it said that all such optimisation research from Northwestern University in the 1960s was solved on an IBM 7090 and so it assumed this one was too.

The penultimate sentence of ChatGPT’s abstract for the first paper in Professor Littlechild’s list (the 1966 Northwestern working paper) said “*Although purely methodological, the approach foreshadows Littlechild’s later advocacy of price-guided discovery in regulated utilities.*” Evidently ChatGPT considered that it was was able to form a view not just on the paper itself (which it had not been able to find) but also on the relation of that paper to what it understood of work much later in Professor Littlechild’s life.

ChatGPT[6] was then asked to find and return copies of the papers in the uploaded list of Professor Littlechild’s papers for which it had produced abstracts. It was also asked to identify any missing papers for the years covered by the documents in each batch. This process elicited links to websites, in many

[6] To be clear, we used the “Pro” version of ChatGPT.

cases with access restricted to academic institutions, to enable us to download about 200 pdf-format documents.

Those documents that ChatGPT had not been able to find were then sought through manual Google and Google Scholar searches. In this way about 60 more papers were obtained from academic journals and similar sources subscribed to by our university. The Internet Archive[7] was then used to locate books and edited volumes that contained approximately 25 papers that ChatGPT could not access and that could not be otherwise obtained. We used Optical Character Recognition (OCR) to digitise these papers. Considerable effort was needed to manually correct OCR errors.

Eleven documents that we had scanned from original paper sources (for example Professor Littlechild's lecture notes from the 1970s and 1980s) were turned into PDF documents through OCR, although considerable effort was again needed to correct OCR errors.

All files were stored in PDF format. All of these PDFs were then processed using OCR software so that all text within each PDF was searchable. Once this processing was done, these PDFs were then resaved as searchable PDFs. These searchable PDFs (approximately 300) were then merged into five large PDF files in order to circumvent ChatGPT's document upload limits (20 documents). As noted we were using the paid, "pro" version. Much lower limits apply in the freely available version. The bibliographic guides to the corpus which we compiled were found to be valuable to SCL in its understanding of the corpus. Efficiently extracting information from those files depends on format with text files preferred to PDFs.

The five large PDF files were uploaded and ingested by SCL along with a bibliography of all uploaded files, in CSV format. These PDFs were too big to be used (Custom-GPT said "token limitation reached" after the PDFs were uploaded). A possible alternative of "JSON" format files was explored but found too costly and so not explored further. Instead, two text (".txt") format files were generated using OCR software to process the five large PDFs. These text files contained approximately five million "tokens" (a token is typically a word but may be a punctuation mark or part of a word).

The final corpus consists of 288 documents (several of which are compilations of papers) and 11 of which had been provided to us directly by Professor Littlechild. The corpus can be compared to Professor Littlechild's own list of papers as shown in Table 1.

Table 1. Professor Littlechild's corpus compared to SCL Corpus

List of papers	**Professor Littlechild's list**	**SCL Corpus**
Books	3	2
Monographs	4	4
Major Reports for UK Government & World Bank	3	4
Publications [these are academic articles, conference proceedings, book chapters]	185	157
Working papers (many of these were unpublished, hence do not duplicate the items in Publications)	54	38
Responses to consultations	42	40
Consultancy reports	34	6
Joint submissions from five former regulators	22	1
Magazine and newspaper articles	87	18
Light-hearted pieces	10	2
Overall Customer Satisfaction (OCS) articles	39	9

[7] The Internet Archive is a non-profit organization that provides a digital library of a vast amount of content, including archived web pages, books, music, and software.

Book reviews	31	2
Life story Parts I and II (counted as one publication)	1	1
Correspondence with Coase		1
University economics course lecture series		3
Pre-privatisation advice to Government		4
Total	**515**	**288**

Comparing the corpus with Professor Littlechild's list shows the main shortfalls are in OCS league articles, book reviews, light-hearted pieces, magazine and newspaper articles, joint submissions with other regulators and consultancy reports. However, the included material is thought to cover Professor Littlechild's main academic and policy contributions.

4. Configuring SCL

Surprisingly, perhaps, LLMs have a strong tendency to generate false or nonsensical information, often presented confidently as fact. This is often referred to as "hallucination" (Malmqvist, 2025). Configuring SCL so as to reduce hallucination was a process of trial and error that we found sensitive to small and seemingly innocuous changes.

Initial testing of SCL delivered plausible sounding answers to questions we asked it. Closer inspection found quotes that were made up and references to Professor Littlechild's papers that did not exist. When asked to explain such hallucination, SCL suggested that "next-token prediction[8] favours fluency over truth"; it also referred to "ambiguous prompts"; "retrieval failures" and "spurious patterns learned from noisy training data". To reduce hallucination, SCL advised to retain the PDFs and text files along with the CSV bibliography, but to instruct that greater weight be placed on the text files than on the PDFs. Testing found that configuration changes (i.e. changes to the instructions) to give effect to this recommendation reduced the extent of hallucination and still ensured fast responses to prompts, using the then-available GPT–o3 model.

Malmqvist (2025) suggests various methods for reducing hallucination including improved training data, novel fine-tuning methods and post-deployment control. Efforts that might be classified under these headings were applied to the configuration. Initially we prevented SCL from accessing the internet and so it relied only on the corpus in answering prompts. While this greatly reduced hallucination, SCL returned "not in corpus" to most prompts. Internet access was therefore re-enabled with rules recommended by the Builder software tool on ChatGPT as to the order in which the corpus and internet were examined (specifically, "*first read the text files and only access the internet if the text files are silent on the relevant prompt").*

We found it difficult to ascertain whether SCL had understood the content of the many files that were uploaded to it. Testing SCL's knowledge of one specific fact (Professor Littlechild's subscription to a particular newspaper as a school-boy) returned a negative response until SCL finally agreed that that fact was contained in a document in the corpus (which it could cite and repeat when asked to). When asked to explain why SCL took so many iterations to find out its knowledge of a fact, Builder explained that SCL initially undertook a "broad sweep" across the corpus. When asked how this situation might be improved, Builder suggested the configuration file should be changed to create a "default source filter" which identified all the documents, except the PDFs, to be examined before answering user prompts.

[8] This refers to the method of predicting the next item (token). Large Language Models (LLMs) utilize next-token prediction (NTP) because it serves as an extremely effective self-supervised training signal that requires no manual labeling of data. By attempting to guess the missing next part of a sequence, the model is forced to internalize deep patterns of human knowledge and logic.

A second focus in configuration was to encourage SCL to respond in a way that we considered to be consistent with Professor Littlechild's "voice", i.e. to respond to questions in a written style that we considered Professor Littlechild would use. This was done by iterative development of the configuration file, starting with a configuration suggested by Builder. We assessed the completeness, extent of hallucination, insight and tone of the answers to questions and adapted the configuration file to improve outputs, often after asking Builder's advice for changes that would deliver the improvements we were seeking. OpenAI's "Harmony Response Format"[9] stressed the importance of specificity and clarity in AI conversations. This was reflected in our development of the configuration through plainly worded instructions and concrete "do/don't" instructions.

Testing what SCL recognised in the corpus, after the corpus been ingested, was interesting and challenging. We found at times inconsistent responses to questions on the existence of specific papers in the corpus, unless the question was very specific. For example, a question on the inclusion of a specific book in the corpus ("is Operations Research in Management in the corpus?") returned a negative response. When we asked the same question a little later, SCL responded positively with supporting evidence of the book. We found that if a very specific and precise question was asked, for example "is Operations Research in Management (with Maurice F Shutler), Prentice Hall International (UK) Ltd, 1991" in the corpus, SCL was more likely to provide a consistently correct response.

Considerable effort was directed at attempting to ensure that SCL produced responses that captured Professor Littlechild's "voice". This was adversely affected by the release of substantially new models ("ChatGPT 5-Thinking" replacing "ChatGPTo-3") that resulted in responses that were more pedantic, risk averse and guarded (i.e. bureaucratic) than we associate with Professor Littlechild. As a result, we changed the configuration to instruct SCL to "*always be critical the way Stephen would; don't be a bureaucrat*". This instruction sought to ensure that SCL, using the latest ChatGPT 5-Thinking model, responded to requests in a way that we had found acceptable in its responses using the ChatGPTo-3 model.

We also found that ensuring that responses were in the "voice" that we estimated to be Professor Littlechild's proved to be sensitive to small and seemingly innocuous changes in configuration. Explicit and unambiguous instruction in the configuration file was found to be important. With the assistance of Builder, the final configuration file contained explicit instructions on the hierarchy for the examination of different files in the corpus and then the internet. This included requirements for the verification of any inferences that SCL made and it contained instructions on SCL's tone and voice.

At the end of this development effort, undertaken over several months in parallel with a corpus that was expanding as we found new documents, SCL was able to deliver answers that we considered to be sufficiently complete, insightful and in the style that we considered consistent with Professor Littlechild, to merit independent assessment. However, the configuration file was now long, with many apparent duplications and inconsistencies. To shorten and tidy the configuration file we instructed Builder to edit it. This resulted in a final configuration that was no more than three pages long.

5. Assessment

Was SCL now able to demonstrate a level of reasoning and style of communication that is consistent with what knowledgeable economists would expect of Professor Littlechild? Here we consider relevant literature on assessment, analyse the assessment by ten assessors, and finally examine the assessors' additional comments.

5.1 Literature relevant to our assessment

[9] https://cookbook.openai.com/articles/openai-harmony

Literature relevant to this assessment can be grouped into collections focused on the personalisation of AI and secondly papers focused on AI's ability in economic logic and reasoning. In both areas, the literature is recent and growing.

In the literature on AI personalisation, Jiang et al. (2024) investigate the ability of large language models (LLMs) to express one of five personality traits. They assign personality types to LLMs, ask them to express those personalities and then assess themselves, alongside human assessment. They find that the LLMs were able to express those personalities and each LLM's assessment of itself was reasonable and consistent with human assessment. Salemi et al. (2024) develop a benchmark to test RAG approaches. Dong et al. (2024) develop a supervised fine-tuning method that empowers end-users to control responses during LLM inference and find that it produces outputs that are preferred by human and LLM evaluators. Louie et al. (2024) develop natural language rules to govern LLM-prompted roleplay intended for mental health clinicians to create "AI patients" that can be used to train mental health counsellors. They find that the counsellors and clinicians found it easy to create AI agents that faithfully resembled real patients. Samuel et al. (2025) develop a dynamic evaluation framework to assess the ability of four open-source and three closed-source LLMs to operate persona agents (e.g. accountants, lawyers, pharmacists). They use "state of the art" LLMs to score 200 generic persona responses against human-developed benchmarks. They also use humans to spot test the responses.

The literature seeking to assess AI ability in economic logic and the merits of AI agents in the field of economic analysis is also growing quickly. Guo & Yang (2024) conduct experiments on various open-source and commercial LLMs and find that without supervised fine-tuning on the training data the open-source LLMs perform closely to the random guess and that the commercial LLMs can generate the wrong or hallucinated answers. They conclude that LLMs of both kinds are not sophisticated in economic reasoning. Fish et al. (2025) develop benchmarks and "litmus tests" for assessing LLM economic agents that act in, learn from, and strategise in, unknown environments, the specifications of which the LLM agent must learn over time from deliberate exploration. Such operation in unknown environments is like the tasks SCL was asked to perform. The assessment by Fish et al(2025) is of three specific, reasonably tightly specified tasks: scheduling, task allocation and pricing. In their tests, there is a well-defined notion of an optimal action, and a natural way to measure the relative quality of a non-optimal action. SCL does not operate in such a narrowly defined environment. Quan & Liu (2024) also develop benchmarks, in their case to assess AI agents' ability to navigate sequential complexities inherent in economic contexts. Their data-based tests are interesting but much more specific and narrowly defined than needed to assess SCL.

5.2 Independent assessment

Our literature search found that systematic machine-based objective assessment of applications such as ours do not exist. Our assessment therefore relied on manual, subjective assessment by human assessors. We asked Professor Littlechild to suggest ten people whom he considered would be able to offer well-informed assessments of SCL. All but one are regulatory economists, the tenth also worked in a regulatory context[10]. Six of the ten previously worked for or with Professor Littlechild. The remaining four have researched or worked extensively in his field and have interacted with him for over a decade.

The assessors were invited to ask SCL whatever they wished and then rate SCL's answers with marks, one to five out of five, on six different measures. They were also invited to provide additional comments if they wished. The six measures were:

1. Completeness: Does SCL the cover the ground Stephen would likely cover in answering your question?
2. Fidelity: Are the answers to your questions true to Stephen's approach and his "voice"?
3. Accuracy: Are the facts, quotes, dates and citations correct?
4. Insight: Does SCL present innovative critique that is nonetheless consistent with Stephen's frame ?

[10] These ten assessors were Sonia Brown, Dr Sarah Deasley, Rachel Fletcher, Dr Ahmad Faruqui, Kyran Hanks, Dr Chris Harris, Dr Eileen Marshall, Professor Paul Simshauser, John Stewart and Andrew Walker.

5. Overall usefulness in economic and policy discourse: Would you use this tool (SCL) in your work as an economist?
6. Blind attribution: If you weren't told the source, how likely would you attribute the answers to Stephen Littlechild - based on method, tone and the citations?

The results of their assessment are set out in Table 2.

Table 2. Assessment results

Reviewer	Complete-ness	Fidelity	Accuracy	Insight	Usefulness	Blind attribution	Mean	Mode
A	5	5	4	5	5	3	4.5	5
B	5	3	5	4	5	2	4.0	5
C	5	5	5	5	5	3	4.7	5
D	4	4	4	5	3	4	4.0	4
E	3	3	5	4	4	4	3.8	4
F	4	4	3	3	3	4	3.5	4
G	5	5	4	5	4	4	4.5	5
H	5	4	5	5	5	4	4.7	5
I	4	4	5	5	4	3	4.2	4
J	5	4	4	5	4	3	4.2	4
Mean	4.5	4.1	4.4	4.6	4.2	3.4	4.2	
Mode	5	4	5	5	5	4		5

Table 2 shows that the average score from the review was 4.2 (out of 5) and the mode was 5. The average score for "Insight" was the highest of the six measures (4.6 out of 5) and seven of the ten assessors gave Insight 5 out of 5. The average score for "Blind attribution" was the lowest (3.5) and "Fidelity" (a similar measure to "Blind attribution") the second lowest (4.1). The mode of the scores for four of the six measures ("completeness", "accuracy", "insight" and "usefulness") was 5, and for the remaining two ("Fidelity" and "Blind attribution") the mode was 4. The variance for "usefulness" was higher than for any of the other measures, suggesting diversity of opinion on this measure, which may of course reflect the different situations of the respondents.

It appeared that none of the assessors undertook a systematic comparative assessment of SCL versus standard ChatGPT though they were free to do so in their assessments. However it is notable that the assessor that rated SCL the least favourably (on average) also suggested that ChatGPT would have got one star [i.e. mark] less on all measures.

The relatively low score for "blind attribution" merits particular note. The assessor that gave the lowest score on "blind attribution" also said the "inauthentic voice" … "was entirely unproblematic". Another assessor gave a higher (than the average of all assessors') score for "blind attribution" but noted SCL's failure to pick up "nuance in audience". Another recognised the difficulty in capturing "voice".

The high average scores for "insight", "completeness" and "accuracy" might seem inconsistent with the relatively lower average score for "usefulness". Perhaps the latter might be explained by the suggestion that not all reviewers would find it helpful to consider SCL's views in their own work, even if they found it insightful and complete and accurate in answering the questions they asked of it.

Our own comparative assessment follows at the end of this section.

5.3 Assessors' additional written comments

Nine of the ten assessors also made several written comments in their assessment:

1. “The referencing to general material and Stephen’s material was good. The synthesis of the questions and responses was good, although it felt about halfway between what ChatGPT would say and what Stephen would say. The voice was not at all authentic but that was entirely unproblematic.”
2. “This SCL Avatar is outstanding. Extremely useful and I found the assessment of my own published work by the Avatar to be balanced, highly credible and therefore highly trustworthy.”
3. “The written answers were consistent with Stephen's style of writing. Whilst I did not check the citations, they were presented appropriately and in a way that Stephen would use them. The arguments for a particular approach were balanced and well laid out. The answers were to questions that were general, so they were general in nature too. I could not say how the avatar would operate in response to specific issues. However, [my] question about the Independent Football Regulator did elicit an approach to regulation that many people would think that Stephen would promote.”
4. “There was a definite flavour of Stephen's tone and thinking in the answers to the 3 questions that I asked and, overall, I was impressed by the responses that the SCL gave. Nonetheless, there were several areas where the responses did not seem to fully capture Stephen's modus operandi or the intensity and scope of his critical thinking and evaluation. The [first] response correctly captured Stephen's focus on dynamic rivalry and his preferences for regulatory arrangements that would appropriately support such a process. I guess what was missing was some of Stephen's intellectual curiosity and the 101 questions he would have posed about the background to the consultation, the evidence base and whether the consultation had correctly identified the best options. The second question was whether the UK Government's policy of promoting infrastructure spending will promote economic growth. It was an interesting response, but I would have expected Stephen to be rather more sceptical of the evidence base that was cited, based on the thinking he set out in the Fallacy of the Mixed Economy. The last question was in relation to the CMA's Final Determination in October 2023 of the Heathrow Airport Licence Modification appeals. Overall, this was a useful summary, but I suspect Stephen would have been a bit more careful to qualify his conclusions. I think he would have been saying things like 'based on the reasoning set out, these appear logical conclusions' and 'while I have not probed the detail of the assessment the CMA's approach appears reasonable'.”
5. “I was impressed. The scores I present are against me having a dialogue with the real Stephen Littlechild - you could pretty much add a star on for each if I was reviewing against my expectations of what I would get from an AI version.”
6. “I might use it in future, particularly if I was after a range of insights or just wanted a change in approach from the AI I was using.”
7. “This is a great tool. It captures Stephen's arguments and analysis. It's difficult to capture Stephen's "voice" and engaging style in presentations through any written form but some answers did this well … It's great to have access to an abridged version of Stephen's individual more in-depth pieces.”
8. “Overall, I felt that the economic content was first rate. I did however feel that the answers were perhaps longer than Stephen would offer and more “padded”. I also felt that for an economist audience Stephen would vary his style to pull on more theory at times to explain the concepts versus when he was trying to influence a broader audience. In other words, nuance in audience is something Stephen is very skilful with, that perhaps isn’t being picked up in the AI.”
9. “Very insightful and enjoyable. Felt like talking to Stephen. Even to the point the replies were in some cases, in my view, overly positive. Metering being put with suppliers was a disaster. As was Ofwat failing to stop Thames Water over-levering. The SCL chat seemed positive about both areas.”

Taken together the scores and the specific comments suggest that SCL succeeded in providing insightful and accurate critiques that are consistent with Professor Littlechild’s perspectives, but that it is less successful in capturing the tone of his expression.

5.4 How do SCL and ChatGPT compare?

We considered whether to ask the assessors to assess whether they thought SCL produced better answers than those of ChatGPT. Such comparative assessment would have been very time consuming and so we decided not to ask this of the assessors. However after having developed SCL we ourselves, however, did seek to understand the extent to which SCL had improved upon ChatGPT for our purposes. We noted earlier that we sought to understand what material of Professor Littlechild's writings ChatGPT knew of and found that it knew of less than a quarter of the corpus we had uploaded to create SCL (or one seventh of the publications Professor Littlechild identified). During the development of SCL, and after its completion, we compared SCL's answers and ChatGPT's answers to various prompts, including by asking ChatGPT to compare its responses with those of SCL. Our observation was that SCL produced better informed, more precisely focussed and detailed answers.

ChatGPT seemed to agree. We asked ChatGPT and SCL to respond to the prompt: "*Provide a 1500-word critique of the Independent Water Commission Final Report*[11]*, from the perspective of Professor Stephen Littlechild". We then asked ChatGPT to compare the two responses. In what it called its "bottom-line" of such a pair-wise comparison, it said: "*If you need a high-level caution against the IWC's centralising thrust, C1 (ChatGPT) is the sharper broad-spectrum critique. If you need a credible operating model that embeds Littlechild principles without freezing decision-making, C2 (SCL) is the more practical blueprint—putting negotiated settlements at the heart of the reset and converting regional planning from paperwork into rivalry and consent*".

To do the comparison more rigorously we sought ChatGPT's and SCL's assessments of the differences in their responses. We did this by instructing ChatGPT to answer the same questions that the assessors had asked SCL. We then asked SCL and ChatGPT to rate ChatGPT's responses and SCL's responses using the same six measures that the assessors had used. Both SCL and ChatGPT were also instructed to write a sentence to justify each score. Table 3 below compares the results of this evaluation.

Table 3. SCL and ChatGPT's evaluation of each other's answers

	ChatGPT's answers					
	completeness	accuracy	insight	usefulness	fidelity	blind attribution
SCL judges	3.9	4.0	3.9	4.2	2.1	1.1
ChatGPT judges	4.4	4.8	3.9	4.6	3.8	2.6
	SCL's answers					
	completeness	accuracy	insight	usefulness	fidelity	blind attribution
SCL judges	4.4	4.4	4.6	4.7	4.6	4.6
ChatGPT judges	4.4	4.5	4.4	4.5	4.5	3.7

The first two lines of numbers in this table show that SCL was harsher in its assessment of ChatGPT's answers than ChatGPT was. The third and fourth lines show that ChatGPT was harsher in its assessment of SCL's answers than SCL was (except in respect of accuracy). ChatGPT was much harsher in its assessment of SCL's answers in respect of "blind attribution" This is likely to be largely because ChatGPT would not accept an instruction to impersonate Professor Littlechild (as a result we had to ask

[11] https://assets.publishing.service.gov.uk/media/687dfcc4312ee8a5f0806be6/Independent_Water_Commission_-_Final_Report_-_21_July.pdf

ChatGPT for "a critique from the perspective of Professor Littlechild" to get it to respond). Consequently ChatGPT did not answer questions in the first person. It was SCL's practice of answering questions in the first person (that ChatGPT could not do) that explains ChatGPT's low score for "blind attribution" of SCL's answers. This was evident from ChatGPT's explanation for the scores it gave.

Comparing the average of these six measures, both ChatGPT and SCL gave SCL's answers a similar score (4.4 and 4.5 respectively) which was higher than both gave to ChatGPT's answers (4.0 and 3.2 respectively). If we exclude the scores for "fidelity" and "blind attribution" for the reasons explained above, both ChatGPT and SCL gave SCL's answers the same average score (4.5). However, ChatGPT scored its answers only slightly lower (4.4) while SCL scored ChatGPT's answers much lower (4.0). It is also interesting that the average score that ChatGPT and SCL awarded to SCL is not much different from the average score that the human assessors gave to SCL.

The comments SCL and ChatGPT made about each other's answers to the assessor's prompts were quite revealing. For example, summarising the ChatGPT and SCL answers to an assessor's request for SCL's review of that assessor's published papers, SCL summarised ChatGPT's response as "*competent and broadly reliable overview, but more a good secondary commentary on my themes than an expression of my own style or concerns".* By contrast ChatGPT summarised SCL's answer as "*A rich, well-grounded and practically oriented response that closely matches Professor Littlechild's intellectual stance and would be very useful to economists and policymakers*." In this comment we see that ChatGPT appreciated the additional insight that SCL was able to deliver.

ChatGPT revealed additional insight into how it works when asked to evaluate SCL's answer to a question that ChatGPT could not answer. Specifically, Professor Littlechild's written correspondence with Ronald Coase was included in the SCL corpus but is not available on the internet to ChatGPT. When asked, SCL was able to provide details of that correspondence (which we verified it did correctly). But when asked to evaluate SCL's answer to the question, ChatGPT gave SCL's answers one out of five for "accuracy" on the basis that "*Highly specific letters, dates, and quotations are presented as fact with no verifiable basis; this is almost certainly largely fictional*". This is wrong, the information presented in SCL's answer is correct. ChatGPT did not know of the Coase correspondence, whereas SCL did. This suggests that in this case it dismissed, as likely wrong, answers to questions about information it did not have.

6. **Discussion**

In the development of SCL, effort has been made to ensure that SCL responds to inquiries in a manner that is consistent with the way that we consider Professor Littlechild would. This has been referred to by us and assessors as SCL's "voice".

The "voice" of SCL's communication elicited specific comments from most of the assessors. The most critical said "*It's difficult to capture Stephen's "voice" and engaging style in presentations through any written form but some answers did this well. In other words, nuance in audience is something Stephen is very skilful with, that perhaps isn't being picked up in the AI*". Another reviewer had a similar comment "*The voice was not at all authentic but that was entirely unproblematic*". But other assessors said, "*Felt like talking to Stephen*" and "*There was a definite flavour of Stephen's tone and thinking in the answers*" and "*The written answers were consistent with Stephen's style of writing*".

AI's "voice" is frequently discussed. In April 2025, ChatGPT withdrew an update to its GPT -4o model in response to customer concerns that the updated model had become excessively sycophantic. The later release of GPT 5-Thinking in August 2025 was criticised for its change in tone[12] and was quickly updated to allow users to customise the style and tone of ChatGPT responses. The customisation allowed users to select the default personality ("cheerful and adaptive") or to select an alternative ("cynic", "robot", "listener", "nerd") which could then be customised further with one of 15 further

[12] https://www.thealgorithmicbridge.com/p/after-gpt-5-release-hundreds-begged

styles[13]. The extensive customisation now possible in ChatGPT indicates the challenge in ensuring that AI is able to produce answers in the diverse styles sought by its diverse audience.

The assessors' differing perspectives on SCL's "voice" suggests that our efforts at customising ChatGPT have had mixed success in being able to respond to questions in the way that users had expected of Professor Littlechild. Professor Littlechild himself wondered how the "voice" of SCL would have changed if SCL knew about more of his popular and accessible contributions rather than mainly his academic output. However we did find that, if provided with information about the audience to which its answers we likely to be directed, SCL did vary the style and content of its answers to some degree. Online Appendix B provides examples of how SCL responded to the same question knowing its audience was either a minister, an official or an academic researcher.

Professor Littlechild expressed reservations that SCL "*tends to take my views or approach as given and then asks what they would imply for a given (regulatory) situation. And I can't disagree with the analyses and recommendations. But … I did several times feel that I wouldn't start from here. Or that I wouldn't be spending time on that issue. SCL's response to the Cunliffe report was perhaps a case in point. The analysis and suggestions sounded plausible. But would I have spent time analysing that situation and spelling them out? Or would I have thought: do we want to be here at all? Is there some significant government or regulatory change that could change the situation?*

He went on to suggest how his own thinking had evolved over time.

… So for example … I thought MC (marginal cost) pricing was the way to go for nationalised industries and spent nearly a decade exploring all manner of math programming and game theory methods of representing, analysing and calculating MC in a wide variety of situations. But it eventually became apparent, not that those methods were wrong, but that there was a quite different kind of problem that required a quite different kind of answer, viz privatization and competition, hence I spent another decade exploring how best to do that. But there were still problems of regulation, hence we needed an alternative to the conventional ROR [rate of return] approach, and RPI-X seemed to fit the bill. But in practice that too had problems, hence another decade searching for, appraising and advocating negotiated settlements. So, the avatar is good at deducing, summarising and applying a relatively fixed set of principles. But is it any good at standing back and thinking: surely there must be a better way of doing things? And then finding one?"[14]

By design, SCL seeks to analyse and critique in a way that is consistent with what it understands of Professor Littlechild's writing. Does this mean that SCL would be incapable of original thinking? Yet SCL's assessors gave SCL the highest scores for "insight" which we defined as containing "innovative critique". This might suggest the capacity for original thinking, in some ways at least. Professor Littlechild was initially sceptical of SCL's ability to create new solution concepts. However, in his review of SCL's responses set out in Appendix B, Professor Littlechild wondered whether this was evidence of SCL being innovative: "Faced with situations that I myself have not analysed and *pronounced on, that indeed I would have avoided because of their complexity, SCL has no alternative but to pronounce, and has invented its own criteria and preferred solutions or remedies. Which are ones to which I am sympathetic."*[15]

The assessors' critique of SCL's ability to write as if it were Professor Littlechild, and Professor Littlechild's questioning of the tool's capacity for original thinking, led us to consider how best to describe SCL. It is not an "avatar" in the sense of the original dictionary definition[16]. "*Individualised persona language agent*" as defined in Chen et al (2024) may be more accurate than "avatar" but falls

---

[13] Chatty, Witty, Straight shooting, Encouraging, Gen Z, Traditional, Forward thinking, Poetic, Opinionated, Humble, Silly, Direct, Pragmatic, Corporate, Outside the box and Empathetic.

[14] Personal communication, 26 September 2025

[15] Personal communication, 7 January 2026

[16] The Oxford English Dictionary's original 1986 entry for "avatar" defined it as "*A graphical representation of a person or character in a computer-generated environment*".

short as a usable term since it does not carry with it any precision on the standard for qualification as an "individualised persona". SCL has the functional characteristics of a "*chatbot*" but this term usually refers to uncustomised AI that does not present itself as seeking to emulate the intellectual tradition of a specific individual. Cassell et al. (2000) used the term "Embodied conversational agent" to describe tools that sought to emulate humans, well before the application of large language models. We reject this term here, because we consider that even using the vastly superior tools currently available SCL still falls far short of "embodying" a person. The term "Digital/virtual human" used in Magnenat-Thalmann & Thalmann (2005) also sounds too ambitious when taking account of SCL's limitations in "voice" and perhaps also in its capacity for imagination. Therefore, for want of a better descriptor, we settled on "AI Agent"[17] to describe SCL.

Has this AI customisation effort achieved its objectives? SCL and ChatGPT seem to agree that SCL's responses score more highly than ChatGPT's against the assessment criteria we set. Can our experience be generalised to other economists? In the case of Professor Littlechild, a substantial part (about one-quarter) of his published work is easily accessible on the internet (for example not behind pay-walls) and it may be that this material was included in the training of ChatGPT, although it is impossible to know for sure if that is the case. This means that most of the corpus we have assembled and upload to SCL was not known to ChatGPT and hence retrieval augmentation drawing on the works that ChatGPT was not aware of, has the prospect of being worthwhile and in our case has been so. Perhaps for those economists most of whose works are already easily available on the internet and so are likely to be included in the training of ChatGPT, the gains to be made from retrieval augmentation will be smaller.

Finally, applications like SCL that reference an individual's scholarship point to the importance of confidence in the authenticity of the scholarship that they access. In response to certain prompts, SCL provided output that was particularly sensitive to some of his papers. Trust in the corpus on which the customised application is trained, is important.

7. Conclusions

Can widely used AI models such as ChatGPT be enhanced by providing access to information that such models have not been trained on, so that they can more faithfully replicate and apply the thinking of a particular person? The present research has sought to answer this question, with Professor Littlechild as its focus. The method applied - "retrieval augmented generation" - is now increasingly used but is novel in the way that we have applied it here to an economist. This research revealed the effort in establishing the corpus and in trying to know what parts of this information ChatGPT had not already been trained on. The research also uncovered the challenges involved in instructing AI on how to understand the additional information and how to use that information to process users' prompts.

Are models like SCL a useful addition to the economist's toolbox? The assessment of SCL by ten regulatory professionals (all but one economists) found high scores overall and particular success in "insight", "completeness" and "accuracy". The assessors thought SCL was less successful in "blind attribution" and "fidelity". We found that SCL had difficulty in knowing its audience and responding accordingly, in the way that aware and perceptive humans can. However we did find that telling SCL a little about who was asking the questions resulted in responses that were better tailored to the audience.

When SCL and ChatGPT were asked to assess both SCL's and ChatGPT's answers to the assessors' questions, both rated SCL more highly. Interestingly, the average score that SCL and ChatGPT gave to SCL was not much different from the average score the human assessors gave to SCL.

[17] Microsoft defines AI agents as "systems that enable Large Language Models(LLMs) to perform actions by extending their capabilities by giving LLMs access to tools and knowledge" (https://microsoft.github.io/ai-agents-for-beginners/01-intro-to-ai-agents/)

Reviewers assessed that SCL had high levels of insight. It often provided nuanced and conditional responses that pointed to factors to be weighed in choosing between competing ideas or in assessing existing ideas. We have integrated SCL into our own work so that one of our first actions in reviewing new papers or reports is to ask SCL for its critique of those papers. We, and colleagues, have found this quickly advances our understanding of those papers.

There are many possible extensions of this work. AI agents like SCL, for scholars of different intellectual traditions, can be expected to develop quite different insights consistent with their own intellectual traditions. These different AI agents might then be instructed to critique each other's insights. This might be expected to facilitate the more rapid formation and dissemination of knowledge and understanding.

Bibliography

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. F. (Eds.). (2000). *Embodied Conversational Agents*. The MIT Press. https://doi.org/10.7551/mitpress/2697.001.0001

Chen, J., Wang, X., Xu, R., Yuan, S., Zhang, Y., Shi, W., Xie, J., Li, S., Yang, R., Zhu, T., Chen, A., Li, N., Chen, L., Hu, C., Wu, S., Ren, S., Fu, Z., & Xiao, Y. (2024). *From Persona to Personalization: A Survey on Role-Playing Language Agents*. https://doi.org/https://doi.org/10.48550/arXiv.2404.18231

Dong, Y., Wang, Z., Sreedhar, N., Wu, X., & Kuchaiev, O. (2024). SteerLM: Attribute Conditioned SFT as an (User-Steerable) Alternative to RLHF. *Findings of the Association for Computational Linguistics*, 11275–11288. https://huggingface.co/

Fish, S., Shephard, J., Li, M., Shorrer, R. I., & Gonczarowski, Y. A. (2025). *EconEvals: Benchmarks and Litmus Tests for LLM Agents in Unknown Environments*. https://doi.org/10.48550/arXiv.2503.18825

Guo, Y., & Yang, Y. (2024). *EconNLI: Evaluating Large Language Models on Economics Reasoning*. http://arxiv.org/abs/2407.01212

Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., & Kabbara, J. (2024). PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. *Findings of the Association for Computational Linguistics: NAACL 2024*, 3605–3627. https://replika.ai/

Korinek, A. (2023). Generative AI for Economic Research: Use Cases and Implications for Economists†. *Journal of Economic Literature*, *61*(4), 1281–1317. https://doi.org/10.1257/jel.20231736

Korinek, A. (2025). AI Agents for Economic Research: August 2025 Update to 'Generative AI for Economic Research: Use Cases and Implications for Economists. *Journal of Economic Literature*, *61(4)*.

Louie, R., Nandi, A., Fang, W., Chang, C., Brunskill, E., & Yang, D. (2024). *Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles*. https://roleplay-doh.github.io/

Magnenat-Thalmann, N., & Thalmann, D. (2005). Virtual humans: Thirty years of research, what next? In *Visual Computer* (Vol. 21, Issue 12, pp. 997–1015). https://doi.org/10.1007/s00371-005-0363-6

Malmqvist, L. (2025). Sycophancy in Large Language Models: Causes and Mitigations. In K. Arai (Ed.), *Intelligent Computing* (pp. 61–74). Springer Nature Switzerland.

Quan, Y., & Liu, Z. (2024). *EconLogicQA: A Question-Answering Benchmark for Evaluating Large Language Models in Economic Sequential Reasoning*. https://huggingface.co/

Salemi, A., Mysore, S., Bendersky, M., & Zamani, H. (2024). *LaMP: When Large Language Models Meet Personalization* (Vol. 1). http://lamp-benchmark.github.io/

Samuel, V., Zou, H. P., Zhou, Y., Chaudhari, S., Kalyan, A., Rajpurohit, T., Deshpande, A., Narasimhan, K., & Murahari, V. (2025). *PersonaGym: Evaluating Persona Agents and LLMs*. https://doi.org/10.48550/arXiv.2407.18416