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
Revised 14 December 2025

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 was needed to reduce hallucination, improve the quality of answers and in pursuit of a representative style. Seemingly minor changes to the configuration settings had large impacts. Assessment of SCL by regulatory professionals who have had long interaction with Professor Littlechild rated it highly particularly in respect of "insight", "completeness" and "accuracy". But assessors were less convinced of SCL's ability to replicate Professor Littlechild's written style. If users provide SCL with context to their questions and information on the audience for its answers, SCL did however deliver tailored responses. SCL and "vanilla" ChatGPT's assessments of SCL's answers to the assessors' questions, agreed on SCL's superiority to vanilla ChatGPT. SCL demonstrated a sophisticated, abstract understanding of Professor Littlechild's scholarship but its ability to replicate his imagination is less clear. Creating AI agents of other economists and setting them to critique each other's critiques may hasten the creation of new knowledge 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 and 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 so as to be able to replicate and apply the thinking of a particular person. 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"[1] and "individualised persona". SCL AI Agent fits into the last category.

We have developed our individualised persona agent through retrieval-augmented generation (RAG) customisation of ChatGPT. RAG involves customising a large language model[2] (LLM) by accessing knowledge outside of the LLM's training data and 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 papers written by Professor Stephen Charles Littlechild from the 1960s to the present. 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 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 we found that ChatGPT could only find about a quarter of the corpus of papers that we had assembled, when it was asked to search for them on the internet. By establishing a large corpus of papers and then instructing ChatGPT on how to respond to user prompts, the resultant RAG application we have developed - "SCL AI Agent" (SCL) - seeks to offer 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 "vanilla" ChatGPT.

The paper proceeds by describing how Professor Littlechild's corpus of writing was established and then uploaded and then how ChatGPT was selected and configured. This is followed by the assessment of SCL by ourselves and others, and then a discussion of the issues. The concluding section draws out the main points and suggests the focus of effort in future.

---

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

[2] 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.)

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 (ability to reason and act as agent) with its competitors (Korinek, 2025) and is accessible on the internet like 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"[3] 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 semantic search, keyword search does not take account of the intent and context of the user's query.

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, in then using ChatGPT's "5-thinking" model to respond to users' queries.

SCL exists as a proprietary application using ChatGPT's "5-thinking" AI platform, with authenticated access through the internet.

3. Establishing the corpus

Initial exploration sought to find out what ChatGPT already knew of Professor Littlechild's works by asking ChatGPT to provide a list of Professor Littlechild's works that it knew of. 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 (many required subscription to journals or paid access to printed or digitised books).

[3] Semantic search is a method that understands the intent and context of a user's query, rather than just matching keywords.

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 was able to do this, despite itself only being able to identify 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 them.

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 the list of 350 was a 1966 working paper that Professor Littlechild 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 paper) said "*Although purely methodological, the approach foreshadows Littlechild's later advocacy of price-guided discovery in regulated utilities.*" Evidently from what it called its abstract, for a paper that it was not able to find and not this was done even before SCL had been developed, ChatGPT was able to form a view not just on the paper itself but also the essential contribution of that paper to what it understood of work much later in Professor Littlechild's life.

ChatGPT was then asked to find and return copies of the papers in the uploaded list of 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 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[4] 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. Optical Character Recognition (OCR) was used to digitise these papers. Considerable effort was needed to 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 improve quality.

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).

[4] 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.

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 AI Corpus

	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 [most 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) league	39	9
Book reviews	31	2
Life story Parts I and II	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

Configuring SCL so as to reduce hallucination[5] and ensure responses are in the "voice" that we estimated to be Professor Littlechild's was a process of trial and error that we found sensitive to small and seemingly innocuous changes.

The five large PDF files described in the previous sub-section were uploaded and ingested by "CustomGPT" 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).

Initial testing delivered plausible sounding answers to questions we asked it. Closer inspection found quotes that were made up and references to papers that did not exist. When asked to explain such

[5] The tendency of LLMs to generate false or nonsensical information often presented confidently as fact is often referred to as "hallucination"(Malmqvist, 2025).

hallucination, ChatGPT suggested that "next-token prediction[6] favours fluency over truth"; it also referred to "ambiguous prompts"; "retrieval failures" and "spurious patterns learned from noisy training data". To reduce hallucination, it 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 SCL was prevented from accessing the internet and so relying only on the corpus in answering prompts. While this greatly reduced hallucination, it returned "not in corpus" to most prompts. Internet access was therefore re-enabled with rules recommended by the Builder software tool on ChatGPT] on 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 was difficult to ascertain whether ChatGPT had understood the content of the many files that were uploaded to create SCL. Testing SCL's knowledge of a 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 it took so many iterations to find out its knowledge of a fact, Builder explained that it 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.

The configuration file was developed through an iterative process, 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"[7] 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" directives.

Testing what ChatGPT 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. Asking the same question a little later was responded to 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*".

[6] This refers to the method of predicting the next item (token).
[7] https://cookbook.openai.com/articles/openai-harmony

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

Is 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 10 assessors and finally examine the assessors' additional comments.

5.1 Literature relevant to our assessment

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 the 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 the clinicians that created the AI agents 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 economic agents 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. However, Fish et al's assessment is of 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 also much more specific and narrowly defined than needed to assess SCL.

5.2 Independent assessment

The literature search finds that systematic machine-based objective assessment of applications such as ours do not exist. Our assessment therefore relies 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[8]. Six of the ten have 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 are:

1. Completeness: Does it 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, citations correct?
4. Insight: Does it present innovative critique that is nonetheless consistent with Stephen's frame?
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, with a minimum of 3.5. Two assessors' average score was 4.7 out of 5. The average score for "Insight" was the highest (4.6 out of 5) and seven of the ten assessors gave it 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 of the "insight", "blind attribution" and "fidelity" measures was equal, and lower than for other three measures. 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.

[8] 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.

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. It is notable that the assessor that rated SCL the least favourably (on average) also suggested that they expected that vanilla ChatGPT would have got one star 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 work, even if they found it insightful and complete and accurate in answering the questions they asked of it.

Reviewing the questions that assessors asked and SCL's answers to those questions, it is evident 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. Our own comparative assessment follows at 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 [SLC] 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 AI Avatar 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 where 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, over 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 does SCL and vanilla ChatGPT compare?

We considered whether to ask the assessors to assess whether they thought SCL produced better answers than those of unmodified (vanilla) ChatGPT (i.e. standard 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 which SCL had improved upon vanilla ChatGPT (“Vanilla”). 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 a 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 Vanilla to compare its response with that of SCL’s. Our observation was that SCL produced better informed, more precisely focussed and detailed answers.

Vanilla seemed to agree. We asked Vanilla and SCL to respond to the prompt: “*Provide a 1500-word critique of the Independent Water Commission Final Report, 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 Vanilla and SCL’s assessment of the differences in their responses. We did this by instructing Vanilla to answer the same questions that the assessors had asked SCL. We then asked SCL and Vanilla to rate Vanilla’s responses and SCL’s responses using the same six measures that the assessors had used. Both SCL and Vanilla were also instructed to write a sentence to justify each score. Table 3 below compares the results of this evaluation:

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

	Vanilla's answers					
	completeness	accuracy	insight	usefulness	fidelity	blind attribution
SCL judges	3.9	4.0	3.9	4.2	2.1	1.1
Vanilla 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
Vanilla judges	4.4	4.5	4.4	4.5	4.5	3.7

The first two lines of numbers in this table shows that SCL was harsher in its assessment of Vanilla's answers than Vanilla was. The third and fourth lines of numbers shows that SCL judged its own answers to be equal or better than Vanilla judged SCL's answers, except for "accuracy" which it said was slightly worse. SCL's judgement of "fidelity" and "blind attribution" of Vanilla's answers, were particularly low. This is likely to be largely because Vanilla would not accept an instruction to impersonate Professor Littlechild (as a result we had to ask Vanilla for a "a critique from the perspective of Professor Littlechild" to get it to respond). Consequently, Vanilla did not answer questions in the first person. It was this failure to answer in the first person that SCL said it judged harshly in its assessment of "blind attribution".

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

The comments SCL and Vanilla made about their evaluations were also very revealing. For example, summarising the Vanilla and SCL answers to an assessor's request for SCL's review of that assessor's published papers, SCL summarised Vanilla'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 Vanilla 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*."

But Vanilla revealed additional insight into how it works, when asked to evaluate SCL's answer to a question that it (Vanilla) 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 Vanilla. 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, Vanilla gave SCL's answers 1 out of 5 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*". But this is wrong, the information presented in SCL's answer is correct. Vanilla did not know of the Coase correspondence, whereas SCL did. Vanilla revealed a bias to dismiss, as necessarily wrong, answers to questions about information it did not have.

6. Discussion

AI in general, and large reasoning models in particular, present the possibility of "AI agents" that impersonate individuals by capturing their way of thinking, as established in their life's work. However, in the case of Professor Littlechild, we found that ChatGPT's most advanced LLM knew of only a small part of his publicly available research and writing. ChatGPT's response to requests to formulate a critique in the style of Professor Littlechild often produced responses that did not reflect a deep understanding of his scholarship.

This motivated exploration of the potential for improvement, by establishing Professor Littlechild's corpus and then providing that to ChatGPT to create a retrieval-augmented generation (RAG) agent. While our approach is a recognised application of RAG, as far as we know this paper is the first contribution to the literature on the application of RAG to an "individualised persona" in the taxonomy developed by Chen et al (2024).

Establishing the corpus took considerable effort and required access to academic repositories, the Internet Archive and digitising documents not available electronically. Uploading this material (around 4 million words in around 300 documents) required manipulation to circumvent upload limits (typically 20 documents). The bibliographic guides to the corpus which we compiled were found to be valuable to ChatGPT in its understanding of the corpus. The format of files was found to be important (text files preferred to PDFs). Instructing ChatGPT on how to understand the material and how to respond to questions on it in the "configuration" file was a process of trial and error, with limited available guidance.[9]

Reducing hallucination and ensuring 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. Interacting with ChatGPT often felt like a very human interaction. At times it would seem agreeable, at other times obsequious almost cunning, at yet other times pedantic and often inconsistent. Explicit and unambiguous instruction in the configuration file was found to be important. With the assistance of ChatGPT's "Builder" tool, the final configuration file contained explicit instructions on the hierarchy for the examination of different files in the corpus and then the internet. It also set requirements for the verification of any inferences that ChatGPT makes, and it contained instructions on the agent's tone and voice. The development of the configuration file was successful in drastically reducing hallucination and ensuring a more consistent response.

The "voice" of SLC'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*".

While this research was underway "voice" has been publicly debated in ChatGPT's development of its models. For example, 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[10] and was quickly updated to allow users to customise the style and tone of ChatGPT responses. The customisation allowed users to select

[9] OpenAI provides high-level guidance at https://cookbook.openai.com/articles/openai-harmony . The sort of specific understanding that would have been more helpful we discovered through trial and error as explained in Section 2.

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

the default personality ("cheerful and adaptive") or to select "cynic", "robot", "listener", "nerd" which could then be customised further with one of 15 further styles[11].

The assessors' differing perspective on "voice" suggests that SCL has had mixed success in being able to respond to requests in the way that users had expected of Professor Littlechild. 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. Evidently, even RAG applications such as SCL have difficulty in knowing their diverse audiences and anticipating their different preferences and responding accordingly, in the way that aware and perceptive humans do or can. Professor Littlechild was reluctant to formally assess the agent, but he wondered how the "voice" would have changed if SCL knew about more of his popular and accessible contributions, rather than mainly his academic output.

Professor Littlechild also expressed reservations that SLC "*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. SLC'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?*

… 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?"[12]

By design, SCL seeks to analyses and critique in a way that consistent with what it understands of Professor Littlechild's known, written outputs. This might suggest SCL would be incapable of independent, innovative thought. Yet SCL's assessors gave SCL the highest scores for "insight". There may be different perspectives on what constitutes innovation, novelty and insight, but perhaps it is not reasonable to imagine that SCL, which is designed to channel Professor Littlechild's demonstrated way of thinking, could also be expected to demonstrate his imagination.

The assessors' critique of SCL's ability to write as if Professor Littlechild, and Professor Littlechild's questioning of the tool's capacity for original thinking led us to consideration of what SCL might be most accurately claimed to be. It is not an "avatar" in the sense of its original dictionary definition[13]. "*Individualised persona language agent*" as defined in Chen et al (2024) may be more accurate than "avatar" but falls short as a usable term and 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 fails to convey its specific construction, purpose and limits. "*Embodied conversational agent*" - see Cassell et al. (2000) - a term that long predates the rise of large language models fails to convey SCL's existence as an AI tool. "*Digital/virtual human*" - see Magnenat-Thalmann & Thalmann (2005) - a term which also long predates the rise of LLMs - might be helpful in indicating its existence in AI but

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

[12] Personal communication, 26 September 2025

[13] 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*".

falls short of communicating its limitations in "voice" and in its capacity for original thought. Ultimately, for want of a better descriptor, we settled on the generic "AI Agent"[14].

Has all this effort to develop SCL been able to improve on Vanilla's ability to answer questions as if it were Professor Littlechild? SCL's and Vanilla's evaluations of each other's answers to the assessors' questions found that both Vanilla and SCL rated SCL's answers more highly. But Vanilla thought its answers was only slightly worse than SCL's whereas SCL thought Vanilla's answers were much worse than its own? While Vanilla also showed remarkable insight despite much more limited access to Professor Littlechild's scholarship, it dismissed the credibility of SCL's answers when asked to evaluate SCL's answers to questions to which SCL, but not Vanilla, knew the answers.

The case for RAG is that, if there is a substantial body of information that ChatGPT has not been trained-on or which is not easily accessible on the internet, a RAG application may produce better responses than Vanilla. In the case of Professor Littlechild, while a substantial part of his work is easily accessible on the internet and so may have been included in Vanilla's training, most of his work is not easily accessible on the internet. While in this case both Vanilla and SCL rated SCL's answers more highly than Vanillas (albeit by different margins), this finding can't be generalised to all RAG since it depends on the extent of the additional information only available to the RAG application.

Finally, applications like SCL that reference an individual's scholarship point to the importance of confidence in the record of 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 which RAG applications have access to is important if those RAG applications come to be widely relied upon.

7. Conclusions

Can AI be enhanced by providing access to information that it has not been trained on, so that AI can faithfully replicate and apply the thinking of a particular person, through the creation of a dedicated AI "agent"? This has been the question this research has sought to answer, with Professor Littlechild as its focus. The method applied here - "retrieval augmented generation" - is now increasingly used but is novel in the way that we have applied it here, at least in the field of economics. This research revealed the difficulties in establishing the additional information (and what parts of this information ChatGPT was not already aware of). The research also uncovered the nuances 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 AI Agent 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 success in the measures for "insight", "completeness" and "accuracy". But the assessors thought SCL was less successful in "blind attribution" and "fidelity". We found that even retrieval-augmented AI such as SCL has 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 tailored responses.

Both SCL and "vanilla" ChatGPT rated SCL more highly than vanilla ChatGPT when they were asked to assess SCL and vanilla ChatGPT's answers to the human assessors' questions. Interestingly, the average score that SCL and vanilla gave to SCL, was not much different to the average score the human assessors gave SCL.

SCL has proved capable of 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.

[14] 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/)

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 envisage that AI agents like SCL, for scholars of different intellectual traditions, might then be set up to debate their critiques and ideas. This might speed up the creation 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