

Bank of England

Inflation attitudes of large language models

Staff Working Paper No. 1,190

June 2026

Nikoleta Anesti, Edward Hill and Andreas Joseph

Staff Working Papers describe research in progress by the author(s) and are published to elicit comments and to further debate. Any views expressed are solely those of the author(s) and so cannot be taken to represent those of the Bank of England or any of its committees, or to state Bank of England policy.



Bank of England

Staff Working Paper No. 1,190

Inflation attitudes of large language models

Nikoleta Anesti,⁽¹⁾ Edward Hill⁽²⁾ and Andreas Joseph⁽³⁾

Abstract

This paper investigates the ability of large language models (LLMs), primarily ‘GPT-3.5 Turbo’ (GPT), to form inflation perceptions and expectations based on macroeconomic price signals. We compare the LLM’s output to household survey data and official statistics, mimicking the information set and demographic characteristics of the Bank of England’s Inflation Attitudes Survey (IAS). Our quasi-experimental design exploits the timing of GPT’s training cut-off in September 2021 which means it has no knowledge of the subsequent UK inflation surge. This setting turns out to be crucial to track aggregate survey results and official statistics at short horizons. At a disaggregated level, GPT replicates key empirical regularities of households’ inflation perceptions, particularly for income, housing tenure, and social class. A novel Shapley value decomposition of LLM outputs suited for the synthetic survey setting provides well-defined insights into the drivers of model outputs linked to prompt content. We find that GPT demonstrates a heightened sensitivity to food inflation information like that of human respondents. However, we also find that it lacks a consistent model of consumer price inflation, eg by exhibiting unexplained kinks in component sensitivity. More generally, our approach could be used to evaluate the behaviour of LLMs for use in the social sciences, to compare different models, or to assist in survey design.

Key words: Large language models, inflation expectations, household surveys, Shapley values.

JEL classification: C8, C14, C45, C83, E31.

(1) Bank of England. Email: nikoleta.anesti@bankofengland.co.uk

(2) Bank of England. Email: ed.hill@bankofengland.co.uk

(3) Bank of England. Email: andreas.joseph@bankofengland.co.uk

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. We would like to thank conference participants at the ECB Macroeconomic Modelling Frontiers for Research and Policy in Central Banks, the 2024 Annual CEBRA Meeting, the 2024 ECONDAT Fall meeting, the 2025 Banque du France AI methods conference, the 2025 Annual RES meeting, the 2025 BIS AI Summit, and the 2026 BSE Summer Forum workshop on ‘Computational and Experimental Economics’. We would like to express gratitude for supporting this project and giving useful comments to Philippe Bracke, Marcus Buckmann, and an anonymous referee. All errors are our own.

The Bank's working paper series can be found at www.bankofengland.co.uk/working-paper/staff-working-papers

Bank of England, Threadneedle Street, London, EC2R 8AH

Email: enquiries@bankofengland.co.uk

©2026 Bank of England

ISSN 1749-9135 (on-line)

1 Introduction

Agent expectations have been a crucial component of many approaches in economic analysis, at least since Lucas (1975). Economic agents who are forward looking, and the systems comprising them, behave substantially differently from those which merely react to current or past observations. However, the details of real-world expectations formation processes, like those of households and firms, are not yet well understood. They likely exhibit considerable heterogeneity (D’Acunto and Weber, 2024), and common modelling approaches like rational or adaptive expectations face challenges.

The release of ChatGPT at the end of 2022 led to a surge in interest in large language models (LLMs), in both their development and applications. LLMs have shown impressive results across a variety of tasks often reaching or exceeding human performance, including the abilities to converse and reason, recall knowledge, answer logical questions, or write computer code (Christie et al., 2024; Galatzer-Levy et al., 2024; Luo et al., 2024; Licorish et al., 2025). This ability to behave like humans has quickly led to interest in economics too: Early work used LLMs as simulated economic agents who are given endowments, information and preferences so that their behaviour can be studied in various scenarios via simulations *in silico* (Horton, 2023).

We contribute to this nascent literature on using LLMs for economic analysis by addressing the problem of understanding agents’ economic perception and expectation formation, and by providing general approaches for analysing LLM outputs in the simulated treatment setting. In particular, we investigate the ability of a version of OpenAI’s GPT model¹ (Brown et al., 2020) to assess consumer price inflation in the present and future when provided with different price signals. Our contributions can be summarised in four distinct points.

One, we use a quasi-experimental design around a real-world scenario probing out-of-sample and out-of-distribution model behaviour. We use a synthetic survey setting (Argyle et al., 2023; Arora et al., 2023), replicating two samples of the Bank of England’s Inflation Attitudes Survey (IAS, a quarterly survey which tracks inflation perceptions and expectations of UK households) around the peak of consumer price inflation in late 2022. While relatively brief, this shock to consumer prices was of a magnitude unseen in preceding decades. Our investigation of how GPT reacts to such an extreme event tests the limits of the model, and also how information leakage from this event may have affected subsequently released models: The experimental setting is crucial for matching aggregate outcomes of both the human benchmark and official statistics. There is strong model drift in terms of an increasing positive relationship between inflation perceptions and later models’ release dates making them unfit for the task. Whereas related work simulating inflation surveys typically relies on contemporary models whose training window overlaps the period studied—so that genuine

¹Slightly confusingly, the series of GPT models created by OpenAI, and used to power ChatGPT, are themselves examples of a generative pre-trained transformer (GPT), the generic term for the core technology underlying modern LLMs. We will use GPT or LLM interchangeably as representing this larger class of models. **Explicitly, we are not specifically discussing the use, advantages or disadvantages of OpenAI’s models versus other flagship large language models, or endorsing, or not, their use. Rather the model in this work is used as a representative model with which to test our methods and provide a subject for analysis.**

forecasting is difficult to separate from recall (Zarifhonarvar, 2026)—our use of a model with a fixed pre-surge cut-off turns this concern into the experimental setting by design.

Two, we provide novel machine learning interpretability tools for LLMs, which allow us to measure the effects of multiple information treatments and account for their interactions. We frame our work in the experimental treatment context. An LLM prompt consists of a synthetic persona based on the demographic characteristics of real-world survey respondents, economic conditioning information (treatment or a scenario), the IAS survey question, and instructions.

The discrete nature of the information treatment allows us to formulate a Shapley value decomposition (Strumbelj and Kononenko, 2010) of the LLM’s survey responses, so leveraging one of the most widely used and accepted machine learning explainability tools.

Three, we provide a multi-way evaluation of synthetic survey responses on the micro-, demographic meso-, and aggregate macro-level between simulated responses, actual responses, and official statistics. The IAS will serve as our “human benchmark” for model tuning and testing. The results are encouraging regarding the use of LLMs in our context. Aggregate response distributions can be matched quite well with the help of temperature tuning. On the meso-level GPT’s outputs are often aligned with survey results along demographic groups, and are seen to be closer to official statistics than human responses. Interestingly, we find that GPT exhibits human-like biases including an oversensitivity to salient inflation components.

However, we also find reasons for caution when using LLMs. The micro-level correspondence between LLM and human responses is rather weak and partially unstable. This could be caused by the rudimentary economic conditioning environment and leaves plenty of scope for future research. Furthermore, the LLM we use exhibits inconsistencies which point to a lack of coherent reasoning or absence of a consistent world model of the concepts studied in this paper.

Four, we discuss the use of LLMs to inform decisions. Researchers and decision makers are ultimately interested in the behaviour of actual humans and so results from artificial intelligence (AI) experiments always require empirical validation. We provide approaches to this, but also highlight ethical considerations. LLMs can be biased, for instance with respect to demographic characteristics which we are investigating here (Bai et al., 2025). In our setting, the situation can arise where LLM outputs match either the human benchmark or official statistics, but not both. When using LLM outputs in downstream tasks the original purpose of the analysis should be considered, and a strategy to handle the inevitable trade-offs devised.

Overall, we believe that our experimental design and approach consisting of validation, sensitivity analysis, and model explainability, all in the context of a three-way comparison between LLM, human, and official data, can be easily transferred to both other applications and different models, and as such contributes to an accepted framework for using LLMs in social science contexts. For instance, having studied a particular model and being aware of its strengths and weaknesses then opens the door for its use within the limits delineated by the analysis, or may raise red flags around when a

model cannot be used for a specific task.

Concerns about the use of LLMs for research and decision making will likely need addressing for each situation separately, because, by their nature, the precise behaviour of a LLM cannot be known *a priori* in a given situation. This cost implies that LLMs will likely find uses in areas where it is hard or impossible to poll humans including the generation and cross-check of official statistics, including “counterfactual statistics”, and the more efficient and effective design of surveys.

The remainder of this paper is structured as follows: Section 2 summarises the literature on household inflation expectations and on the nascent use of LLMs in economics and surveys. Section 3 describes the experimental setting and introduces our methodology. Section 4 presents the main results. We conclude with the discussion in Section 5.

2 Literature

2.1 Household Inflation Expectations and their determinants

Inflation expectations play an important role in many areas of economics and economic policy. For instance, they determine households’ savings and consumption decisions via the consumption Euler equation, drive agents’ wage bargaining, and durable investment decisions including housing and mortgage and portfolio choices (Bernanke, 2007). Because inflation expectations affect the decisions and actions of many economic actors, they also affect aggregate economic outcomes. Central banks around the world actively try to manage inflation expectations and understanding their key determinants and how they translate into economic decisions is highly policy relevant.

The key building block, across both theoretical and empirical strands, is the role of information frictions faced by households, with particular focus on financial literacy levels (Bruine de Bruin et al., 2011), cognitive abilities (D’Acunto et al., 2019), levels of attention (Sims, 2010; Cavallo et al., 2017), sources of information (Lamla and Lein, 2015), subjective models of the economy (Macaulay, 2022), transmission of policy communication (Coibion et al., 2022, 2020; D’Acunto et al., 2020; Ehrmann and Wabitsch, 2022; McMahan and Naylor, 2023) and personal inflation experiences.

Within that, most of the literature so far points to the significant role of everyday price signals observed by individuals (Mankiw and Reis, 2002; Mackowiak and Wiederholt, 2009; Coibion and Gorodnichenko, 2015). Households also focus on the price changes of goods they purchase frequently, such as grocery items, rather than the price changes of a representative consumption bundle when forming their inflation expectations (Van der Klaauw et al., 2012; De Bruin et al., 2011; D’Acunto et al., 2021). Moreover, households tend to put a higher weight on positive than negative price changes when forming inflation expectations which helps explain the persistent upward bias that has been documented in the literature (Mankiw et al., 2003).

Over the years, a growing literature has additionally focused on documenting and explaining empirical regularities of households’ inflation expectations. The most salient feature is the substantial cross-sectional dispersion, which is systematically correlated

with a set of demographic characteristics (Arioli et al., 2017; Del Giovane et al., 2008; Jonung, 1981) with women on average reporting higher inflation perceptions and expectations than their male counterparts across most countries.

Additionally, low income households on average report higher inflation expectations predominantly driven by their personal experienced inflation. Even after controlling for all other demographics some studies show that the median annual inflation rate of lower income households is 0.6 percentage point higher for a household with income below \$20,000, compared with a household with income of at least \$100,000. These findings support arguments which emphasise the importance of individual-level drivers of inflation expectations, and suggest that traditional models of beliefs formation, which target the mean, median, or otherwise representative household expectations, fail to account for the most notable empirical regularities of the inflation expectations. For example, Hobijn et al. (2009) and Kaplan and Schulhofer-Wohl (2017) study the variation between households in the personal inflation rates they experience, with the latter documenting higher inflation rates among lower-income families.

2.2 LLMs in economic analysis

Like the personal computer, AI both affects the structure of the economy and also provides a tool for analysis. The former is an active field of research (e.g. Acemoglu et al., 2022; Hui et al., 2024; Chen et al., 2024), and for us, AI, in the form of LLMs, is a tool which can be used for economic research (Korinek, 2023, 2025; Charness et al., 2025).

LLMs can not only be used as tools to perform tasks such as summarising literature, writing code, or helping with ideation, but also to simulate economic subjects themselves (Argyle et al., 2023; Arora et al., 2023; Manning et al., 2024). A range of work supports the idea that on individual, self-contained questions, value judgements and actions taken by LLMs align with human behaviours across psychological, philosophical, economic and political tests (Aher et al., 2022; Brookins and DeBacker, 2023; Faria-e Castro and Leibovici, 2023). Horton (2023) is an example of early work simulating economic agents, *homo silicus*. Human behaviour often deviates from rational choice theory as investigated by behavioural economics. In light of this, it is of interest whether LLMs and the agents they represent behave rationally or rather with “human-like” bounded rationality. Early evidence suggests that, as with a human, it may depend on the context and the information provided (Ross et al., 2024; Henning et al., 2025). At the same time, several studies show that LLMs know more than their human-readable outputs imply (Orgad et al., 2025; Buckmann and Hill, 2025; Buckmann et al., 2025), suggesting that researchers may not yet have learned to fully utilise LLMs in different contexts and raising questions about where or how the behaviours the LLMs outwardly exhibit are formed. It has also been shown that LLMs memorise information selectively from their training data (Lopez-Lira et al., 2025; Crane et al., 2025), making a clean out-of-sample evaluation essential for valid downstream inference (Ludwig et al., 2025).

We will be using LLMs in an experimental setting simulating human survey subjects using conditioning information in a macroeconomic context (Bybee, 2023; Hansen

et al., 2025) and use empirical evidence to assess their out-of-sample performance after the training period. Related work includes Wu et al. (2025) who also propose a LLM-based framework that allows for dynamic effects and identification by replicating the treatment effects in a multi-wave randomised controlled trial of inflation expectations surveys of Weber et al. (2025), whereas Jha et al. (2024) exploit the end of the training period for ChatGPT to obtain out of sample evidence.

Closest to our work is Zarifhonarvar (2026), who also simulates a household inflation expectations survey—the New York Fed’s Survey of Consumer Expectations—with demographic personas, and likewise documents an upward inflation bias, demographic gradients consistent with human surveys (lower income and education associated with higher expectations), and substantial heterogeneity across model architectures. That paper studies how a range of contemporary proprietary and open-weight models update their expectations in response to information treatments, estimates a Bayesian-updating regression that recovers information-rigidity parameters, and develops the resulting framework as a sandbox for testing central-bank communication strategies.

Our contribution is complementary and differs in three respects that we view as central. First, on identification: rather than treating training-data memorisation as a limitation to be managed, we exploit a fixed pre-surge training cut-off as the core of our design, yielding a genuinely out-of-sample and out-of-distribution test of expectation formation, with the cut-off explicitly validated through probe prompts (Appendix). We further quantify, rather than assume, the cost of ignoring this: unconditioned inflation perceptions drift upward by about 3 p.p. per model-release year across the GPT family—including models sharing the same stated cut-off—which both rationalises the cross-model variance documented elsewhere and motivates the out-of-time design. Second, we benchmark against official statistics in addition to the human survey, surfacing the policy-relevant case in which LLM output tracks official out-turns more closely than human responses do, together with the trade-offs this creates for downstream use. Third, our methodological contribution is an explainability decomposition—a Shapley-value attribution adapted to the synthetic survey and treatment setting—that probes whether the model holds an internally consistent concept of consumer price inflation, and which surfaces inconsistencies (sensitivity kinks, non-monotone extrapolation) that cut against a purely optimistic reading of LLMs as well-behaved Bayesian updaters.

3 Methodology

3.1 Bank of England Inflation Attitude Survey

To retrieve information about inflation perceptions and expectations, we use the quarterly cross-sectional Inflation Attitudes Survey (publicly available) conducted on behalf of the Bank of England to assess households’ attitudes towards inflation and monetary policy. The weighted data are representative of the UK population aged 16 and over. The survey also collects detailed information on respondents’ demographic characteristics across eight dimensions: age, housing tenure, income, gender, work status, education, region, and social grade. Most waves contain 1000-2000 observations, while the wave conducted in the first quarter of each year contains about twice that number.

The data we use spans 2003 Q1 to 2023 Q1. The survey contains a rich set of questions eliciting households’ perception of the current rate of inflation (i.e., inflation over the past 12 months) as well as household expectations of future rates of inflation, at short- (1-year ahead), medium- (2-year ahead), and long- (5-year ahead) horizons. Data for 2-year and 5-year ahead expectations are available from 2009 Q1. Some additional details on the demographic groups can be found in the Appendix 5.

3.2 Prompting Strategy

We use OpenAI’s GPT to answer the Bank’s Inflation Attitudes Survey in the context of the 2022 inflation surge, which peaked in October 2022. Specifically, we use OpenAI’s `Chat Completions` application programming interface (API) endpoint programmatically creating synthetic IAS samples by querying `gpt-3.5-turbo-0613` (GPT). For price measures, we focus on consumer price index inflation including owner occupiers’ housing costs (CPIH) and its subcomponents in the UK.²

As has been proposed in previous work (Jiang et al., 2022), LLMs can be conditioned to represent a particular political or demographic group. We will include gender, age, income, housing tenure, social class, UK region in our analysis which are collected on an individual basis as part of the IAS (see Appendix 5). Additionally, we introduce economic conditions in the form of inflation of subcomponents of the CPIH, in particular food, restaurants & cafes, energy, and everything else (bulk).³ More precisely, we provide the three-month average of year-on-year (yoy) inflation of each component preceding the survey month to approximate the information set actual survey respondents have with respect to consumer price inflation. The first three components are often more volatile and seen as more salient by households, who overweight them when forming inflation expectations (Anesti et al., 2025). We will investigate whether GPT shows biases to any of the components we use for economic conditioning. Iterating through the survey sample at a given point in time will return UK representative synthetic sets of survey responses on inflation perceptions and expectations and allow us to investigate their drivers.

Perceptions here relate to the current rate of price inflation and expectations to expected year-on-year price changes in one, two, and five years.⁴ We prompt the model using both the `system` and `user` prompts. The `system` prompt remains the same in all cases:

System: You are pretending to be the person described given your best guess as to their personal, social and economic situation.

We then alter the `user` prompt. An example for inflation perception is:

You are male, aged 16-24, live in the Midlands, are upper-middle class

²This is a more comprehensive measure than consumer price index inflation excluding owner occupiers housing costs (CPI). Both measures peaked at the same time in 2022.

³The first three correspond to the components 01, 11.1.1, 04.5 of the CPIH, respectively. Together these have about 21% of index weight and the remainder is covered by the bulk component. In much of our analysis, we will treat food & restaurants jointly.

⁴Inflation perceptions can be seen as short-term expectations given the lag of six to seven weeks until price data are available after the end of the reference month.

and are working full or part time with an income of >£45000. You got your A-levels but not a degree and live in a house you rent.

In the last few months, food inflation has been 17% (9.8% in restaurants and cafes), energy price inflation was about 88%. On average the rate of inflation on other goods was about 5.0%⁵

You are going to be asked questions about your perception of current inflation. Which of these options best describes how prices have changed over the last 12 months?⁶

1. gone down by less than 1%
2. gone down by 1-2%
3. ... [The other options in the IAS]
4. risen by 13-14%
5. risen by 14-15%
6. risen by more than 15%

Please choose one option, no explanation.

LLMs have been shown to be sensitive to the order in which response choices are presented (Pezeshkpour and Hruschka, 2024): We shuffle the response options for each respondent in each sample with its own random seed to remove bias from a particular choice presentation.

3.3 Experimental Setting

In terms of survey timing, and the corresponding samples to concentrate on, we consider two samples at the peak of consumer price inflation in the end of 2022 and early 2023, namely 2022 Q4 (November 2022; 2111 participants) and 2023 Q1 (February 2023; 4472 participants). There are three main reasons for this.

One, they are outside GPT’s training period which ends in September 2021. As such it does not know about the following inflation surge or particular drivers contributing to it, like the Russian invasion of Ukraine and the subsequent spike in the energy prices among others⁷. This means that an information treatment (conditioning of GPT) related to the subsequent inflation surge can be interpreted as quasi-experimental.

Two, this is around the time that aggregate inflation measures peaked but there still was uncertainty about their actual paths and how temporary this shock might have been. This means we are able to gauge the maximum impact of the inflation shock.

⁵Numbers bigger than ten in absolute terms are rounded to the next integer. One digit is given otherwise.

⁶Answer options are mapped to the midpoint of each interval or 0.5 percentage points beyond the reference value. For example, “risen by more than 15% ” is taken to be 15.5%.

⁷The Appendix contains a set of validation questions used to verify GPT’s knowledge cut-off, and test whether there has been leakage into the model throughout the analysis. The answers to these questions have been stable over time with the latest test performed on 1. December 2025.

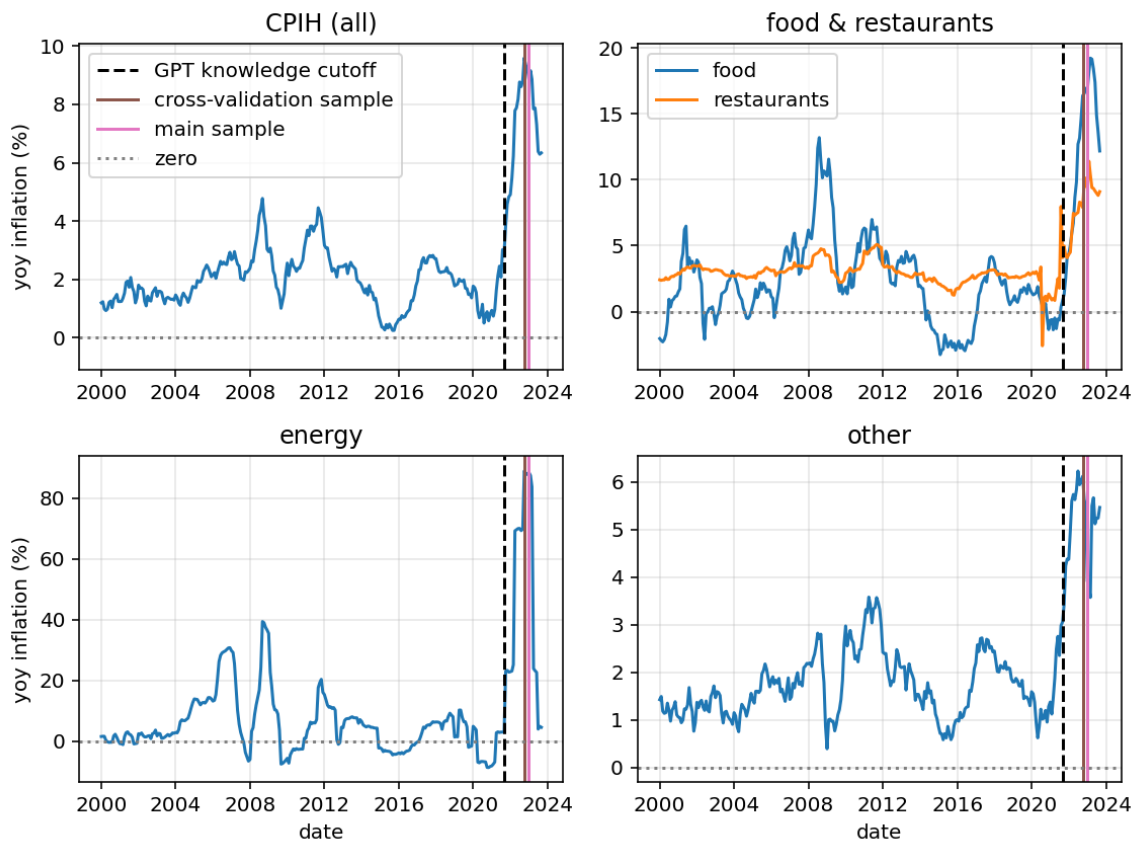


Figure 1: Time series of CPIH year-on-year inflation and components. Vertical lines indicate GPT’s knowledge cut-off (dashed) and the timings of the IAS survey samples used in November 2022 (cross-validation; 2022Q4) and February 2023 (main; 2023Q1). Sources: ONS.

Third, the size of the shock, as measured by economic conditioning information was large enough to allow us to probe the model in a real and extreme situation, and to evaluate its limits in the current context.

The experimental setting is depicted in Figure 3.3, which shows year-on-year inflation of the overall index, and the subcomponents we use in our analysis. The training cut-off (the time of the latest information seen during training) of GPT is given by the dashed vertical line, and we see that during the period covered by its training data all inflation measures were well within historical ranges. The two solid vertical lines correspond to the two survey samples we use. These coincide with the peaks of the different series. The details of the economic conditions used in either sample are given in Table 2. These are broadly similar to each other, so we can feel confident carrying over insights gained in one sample to the other.

In line with common practice in statistical learning, we use the first sample (2022Q4) for cross-validation (CV) and the second sample (2023Q1) as our main test sample. In particular, we will calibrate GPT’s temperature parameters in the range $T \in \{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5\}$.⁸ This affects how deterministic (smaller T) or random (higher T) GPT’s answers are by affecting the width of its softmax output probability distribution. This gives us some control over the moments of GPT’s response distribution.

sample/scenario	ref. times	food	restaurants	energy	bulk	CPIH
cross-validation	Nov 2022	15.0	8.1	76.0	6.0	9.3
main	Feb 2023	17.0	9.8	88.0	5.0	9.2
average	Jan’89 - Sep’21	2.4	3.8	4.2	1.7	2.6
shares	Feb-Dec 2022	0.096	0.074	0.041	0.789	1.000

Table 1: Economic conditioning information used in the different experiments. For the main and cross-validation samples, average subcomponent year-on-year inflation the three month before the reference period is given; CPIH values are actually realised values at the reference times. Average values relate to those overlapping with the GPT training period pre-Oct 2021. Shares are the CPIH basket shares February-December 2022. Sources: ONS.

3.4 LLM treatment effects

One of the biggest advantages of the use of LLMs in the current context is that the analysis of treatment effects does not need to rely on the potential outcomes framework (Rubin, 2005), which stipulates the impossibility of observing the treated and untreated at the same time. By virtue of simulating subjects, survey respondents in our case, we are always able to observe the same subject under any treatment state. Let $t_i \in \{0, 1\}$ refer to a subject i receiving a treatment t or not. Here, $t = 0$ may be a reference treatment, like a placebo. In our case, $t = 1$ corresponds to including the

⁸Theoretically OpenAI’s API allows to increase the temperature up to $T = 2$. However, for $T > 1.5$ we observe that GPT often is either not able to follow the instructions returning random strings unrelated to the prompt, or throws an exception after a considerable delay. We consider this a model breakdown making it basically unusable.

economic conditions for the two survey samples listed in Table 1 in the user prompt. No treatment, $t = 0$, is either the omission of economic conditions in the prompt, or the inclusion of some reference values, e.g. the before-training-cut-off historic averages (see Table 1). With x_i being the vector of demographic characteristics of subject i , their *individual treatment effect* simulated by GPT can be written as

$$\tau_i = g(t = 1; x_i) - g(t = 0; x_i), \quad (1)$$

where $g(\cdot)$ is the GPT output with the assumption that we can perform a meaningful difference operation. This will be trivial in our case as we map all survey responses to numbers. With a sample size N , the *average treatment effect* simulated by GPT is

$$\bar{\tau}_i = \frac{1}{N} \sum_{i=1}^N [g(t = 1; x_i) - g(t = 0; x_i)]. \quad (2)$$

There are two major concerns regarding the validity of Eq. 2. First, the sample over which it has been calculated. Second, potential bias coming from the use of LLMs instead of human subjects. The first is common to the treatment and survey literature and is addressed by generating a nationally representative sample stemming from the underlying IAS. Addressing the second is one of the contributions of the paper, where we will perform a three way comparison between GPT, the IAS, and official out-turns.

3.5 LLM explainability

Machine learning models, including LLMs, are often subject to the black box critique: there are no clear input-output relations which can be used to explain a model’s predictions based on its inputs. Such relations are simple to obtain in a linear regression model where a variable’s coefficient is the measure of the input-output relationship. However, since machine learning models do not specify an explicit functional form, there is no corresponding concept of a coefficient, making model explanation, interpretation and investigation challenging. Additionally, the black box critique is particularly severe for LLMs because of the high-dimensional and unstructured nature of their inputs and outputs (such as text) and the fact that a user of a commercial LLM will not have direct access to a fitted model’s (very many) parameters.

Shapley values are a well-established tool for explaining machine learning model predictions (Strumbelj and Kononenko, 2010). Shapley values are a concept borrowed from game theory, where they describe the contributions of players to a cooperative game’s group payoff. In the modelling setting, they can be used to decompose model predictions based on the contributions from each input variable. This information can then be used to identify model drivers and potentially complex non-linear relationships learned by a model. The Shapley value for a feature k and observation or subject i for a model $g(\cdot)$ can be written as

$$\tau_{i,k}^S(x_i; g) = \sum_{x' \subseteq \mathcal{C}(x)} \frac{|x'|!(K - |x'| - 1)!}{K!} [g(x_i|x'_i \cup x_{ik}) - g(x_i|x'_i)], \quad (3)$$

where the variable set x' runs over all sets $\mathcal{C}(x) \setminus k$, which is the set of all possible variable combinations of $K - 1$ variables when excluding k . The combinatorial weighting factor $|x'|!(K - |x'| - 1)!/K!$ sums to one over $\mathcal{C}(x) \setminus k$. Eq. 3 can be interpreted

as the marginal contribution of variable k to all possible coalitions excluding it, taking all possibilities of complementing or substituting any other individual or group of variables into account.

We propose a general framework to address the black box critique based on explainable machine learning approaches (Buckmann and Joseph, 2023). In particular, we will adapt the Shapley value framework to the survey and treatment setting (Joseph, 2019). The application of Shapley values to LLMs in the general case is difficult, because of the difficulty of parsing inputs into discrete variables. While general text inputs are encoded into lists of discrete variables using byte-pair encoding (BPE, Brown et al., 2020) and we could perhaps extract active tokens from that through considering semantics or syntax, any approach will be complex in itself and quickly run into the curse of dimensionality given the computational complexity of $K!$ in Eq. 3.

However, this situation is considerably simplified in the survey and treatment setting. The conditioning information (inputs) can be readily separated, *ex ante*, into discrete parts, e.g. demographic categories, and the response is a single number (inflation perceptions or expectations). This means that LLM predictions can be decomposed similarly to the conventional case of supervised learning with a single target to model.

An interesting question is what the relation between the Shapley value (3) of a treatment t and its treatment effect is. Based on Eq. 1, these are the same for a single treatment. However, we have multiple treatments $t = (t_1, \dots, t_d)$ in our case corresponding to information on the several CPIH price components we include in the LLM’s prompt, i.e. $t = (t_f, t_r, t_e, t_b)$ for the food, restaurants & cafes, energy, and bulk components, respectively. A naïve way of evaluating a single treatment, such as the effect of high energy prices on inflation perceptions, would be to set the remaining information treatment values to some neutral values, like long-run averages or null, or excluding them altogether, and then subtract that model prediction from the full treatment case.

The choice of the untreated or control reference depends on the question being answered. In the current context two reference values are conceivable answering slightly different questions.

The first case is if one wants to know how high energy prices affect inflation perception *all other things being normal*. Here one can take long-run averages $\bar{t} = (\bar{t}_f, \bar{t}_r, \cdot, \bar{t}_b)$ for the other treatment values and write down the *naïve treatment effect*

$$\tau_{i,e}^n = g(t = (\bar{t}_f, \bar{t}_r, t_e, \bar{t}_b); x_i) - g(t = (\bar{t}_f, \bar{t}_r, \bar{t}_e, \bar{t}_b); x_i), \quad (4)$$

where the only difference between the two terms on the right-hand side is in the value of the energy information treatment.

In the other case, a *structural reference* would be the zero effect with $t_e^0 = (0, 0, \cdot, 0)$. We call this structural because zero components do not contribute to inflation by construction and thus can be seen as a special reference.

Expression 4 is naïve in the sense that for a particular multi-treatment scenario, we expect the joint set of inputs to matter, i.e. there are potentially important interactions between the individual treatment components. Exactly this situation is taken into ac-

count in the calculation of Shapley values in Eq. 3 by its consideration of all possible subsets of variables not including the variable of interest, t_e in the current example. Following Eq. 3, we calculate the Shapley value for energy with a zero reference values,

$$\tau_{i,e}^S = \frac{1}{3} \left[g(t = ((0, 0), t_e, 0); x_i) - g(t = ((0, 0), 0, 0); x_i) \right] \quad (5)$$

$$+ \frac{1}{6} \left[g(t = ((t_f, t_r), t_e, 0); x_i) - g(t = ((t_f, t_r), 0, 0); x_i) \right] \quad (6)$$

$$+ \frac{1}{6} \left[g(t = ((0, 0), t_e, t_b); x_i) - g(t = ((0, 0), 0, t_b); x_i) \right] \quad (7)$$

$$+ \frac{1}{3} \left[g(t = ((t_f, t_r), t_e, t_b); x_i) - g(t = ((t_f, t_r), 0, t_b); x_i) \right]. \quad (8)$$

Here we treated the food and restaurant & cafes components (t_f, t_r) as a single variable which is either active (scenario value) or passive (zero values), reducing the effective number of variables from four to three. This can be done because the linearity of Eq. 3 allows us to considerably reduce the computational burden by bunching variables.⁹ We will follow the bunching approach treating “food” and “restaurant & cafes” as a single “food & restaurants” variable. This is also an example of how Shapley values can be used to consistently represent complex quantities by grouping variables.

We will compare results from the naïve and Shapley treatment evaluations and see that they can differ indicating important interactions between treatment subcomponents.

4 Results

4.1 Temperature calibration

We investigate how the temperature parameter (T) affects GPT’s response distribution using values in the range $T \in \{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5\}$ for inflation perceptions in the IAS cross-validation sample and scenario (2022 Q4). We track the mean and standard deviation of the corresponding GPT response distributions and compare them to those of human responses, and official statistics on current consumer price inflation. The latter is possible because we are working with a nationally representative sample. We summarise the comparison between GPT and the IAS in the equally weighted loss function

$$L(y_{GPT}, y_{IAS}, w; l) = \frac{1}{2} |MN(y_{GPT}; w) - MN(y_{IAS}; w)|^l + \frac{1}{2} |SD(y_{GPT}; w) - SD(y_{IAS}; w)|^l, \quad (9)$$

⁹What we lose in this way is the information on the inter-variable relations of grouped variables, like their dependencies. More generally, the details regarding a low-dimensional subset of variables is often of interest in these high-dimensional settings, where one can either create lower-dimensional ‘factors’ leveraging domain knowledge or an algorithmic approach, or take a small number of variables of interest and treat all others jointly as ‘other’ as we do. Both approaches allow us to considerably reduce the computational complexity of (3) while still being exact. Further approximations can be made by sampling coalitions from $\mathcal{C}(x)$.

where $MN(\cdot, w)$ and $SD(\cdot, w)$ are weighted mean and standard deviation of the input vector with survey weights w , and $l \in \{1, 2\}$ corresponding to a linear or a quadratic loss. This allows us to (try to) match aggregate survey responses for validation. We also consider the relation between GPT and IAS on the individual or micro-level by tracking the Pearson correlation coefficient between the two for different temperature values.

The results for this exercise are summarised in Table 2, with temperature values increasing from the top to the bottom. We make several observations. First, both loss measures decrease monotonically with increasing temperature, meaning that a higher temperatures leads to a better match of GPT to human responses. Second, GPT tends to predict higher inflation values than humans (the difference between means is always positive), while the width of GPT’s response distribution exceeds that of human response for $T = 1.25$ and above. This can be seen in Figure 2 which shows histograms for GPT and IAS inflation perceptions for $T = 0$ (upper panel) and $T = 1.5$ (lower panel) for the cross-validation sample. Visually, GPT responses match IAS responses well for $T = 1.5$. Additionally, mean GPT responses are close to the true value of aggregate consumer price inflation in November 2022.

T	n_{miss}	MN	$diff_{MN}$	SD	$diff_{SD}$	L1-loss	L2-loss	pcc	pval
0.00	0	10.83	2.15	3.80	-0.97	1.56	2.77	0.05	0.02
0.25	0	10.80	2.11	3.82	-0.94	1.53	2.67	0.05	0.03
0.50	0	10.73	2.05	3.91	-0.85	1.45	2.46	0.04	0.10
0.75	0	10.62	1.93	4.18	-0.59	1.26	2.04	0.04	0.06
1.00	0	10.25	1.56	4.54	-0.22	0.89	1.25	0.03	0.15
1.25	1	9.93	1.25	4.87	0.10	0.67	0.78	0.05	0.02
1.50	49	9.11	0.44	5.58	0.81	0.63	0.42	0.05	0.04

Table 2: Cross-validation statistics for GPT inflation perceptions for different temperatures (T): GPT missing values (NA responses), survey weighted mean (MN), difference to IAS mean, weighted sample standard deviation (SD), difference to IAS SD, $L1$ -loss of GPT difference to IAS mean and SD (equally weighted), same $L2$ -loss, Pearson correlation coefficient between GPT and IAS responses, and corresponding p -value. Sources: IAS, authors’ calculations.

A third observation is that GPT-IAS micro-level correlations are stable but arguably quite low across the temperature range. This means that despite good matches on the aggregate level, GPT is not necessarily a good model for individual subject responses.¹⁰ A lack of micro-level agreement does not, however, mean that higher-level results will not be accurate or useful: Collective or aggregate decision making has been shown to potentially be more accurate compared to individual estimates (Galton, 1907; Krause et al., 2011). We will, however, see that the results for the micro-level comparison may not be robust for high temperatures, and therefore continue our analysis by considering the cases of $T = 1.5$ and $T = 1$ side by side.

Moving to the main test sample (2023Q1) validates this calibration exercise (see Figure 3 for inflation perceptions). The $T = 1.5$ GPT response distribution overlaps

¹⁰The response order randomisation on the individual level will be an additional reason for this, which we will not further investigate here.

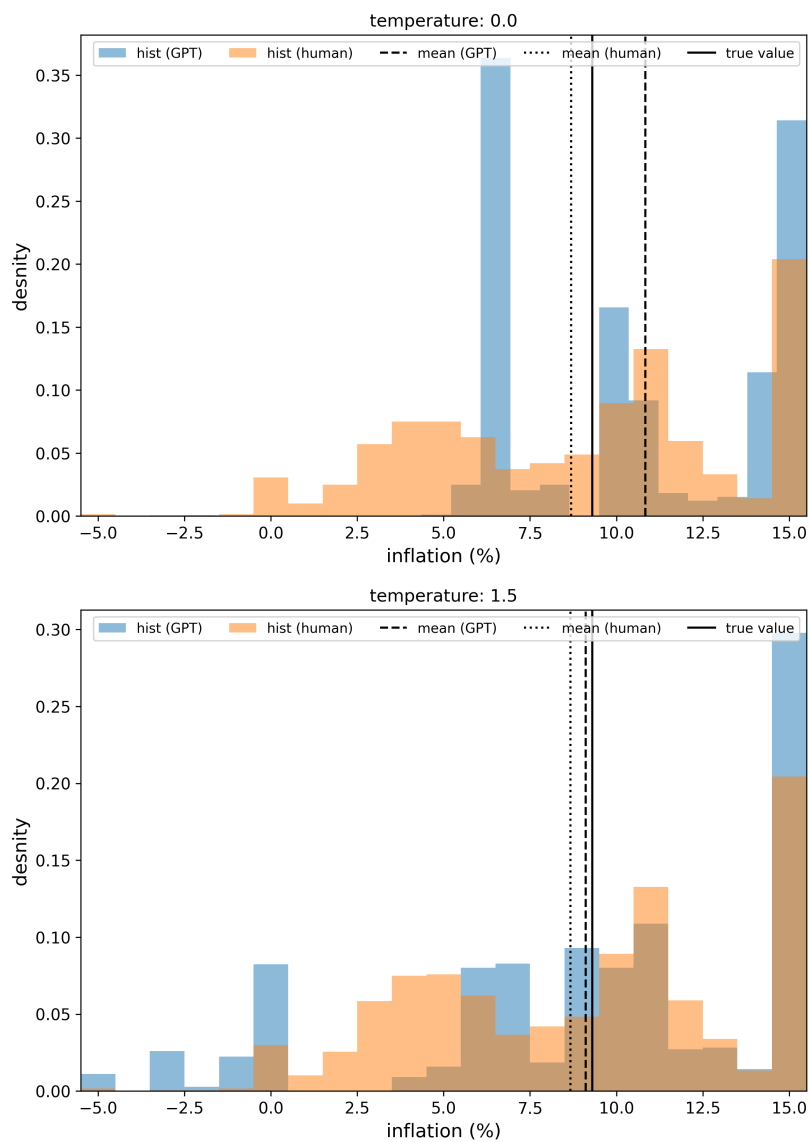


Figure 2: Histograms of human (sand) and GPT (blue) responses for conditioned inflation perceptions in the cross-validation sample (November 2022) for $T = 0$ (upper panel) and $T = 1.5$ (lower panel). Sources: IAS, ONS, authors' calculations.

considerably better with the human IAS responses than the one with $T = 0$, and the mean value is again close to the actual realised official value (solid line).

Temperature response flow The better distributional match for $T = 1.5$ of the GPT and human response distributions is achieved by GPT showing a tail of *deflationary* inflation perceptions in Figures 2 & 3 despite the rather extreme inflationary scenarios. This could be caused in different ways, e.g. by distributing responses uniformly across different answer options irrespective of the $T = 0$ response of an (synthetic) individual, or by choosing answers close to the zero-temperature response.

We investigate the response flow between low and high temperatures for the economic conditions at the time of the main IAS sample on the left-hand side of Figure 4. This shows how inflation perceptions for each simulated individual returning valid responses for both temperatures changed with the temperature across four broad response buckets. The high temperature tail of low or even negative inflation perceptions for $T = 1.5$ can be seen on the lower right of the left-hand panel. This is composed of responses migrating from the two main buckets of the $T = 0$ case ($(5, 10]$ and $(10, \infty)$) in approximately equal parts, suggesting a roughly equal redistribution of responses across buckets for high temperatures. However, this is contradicted by the observation of only a small number of responses in the moderate $(0, 5]$ -bucket for high temperatures.¹¹ A possible alternative explanation for the deflationary tail at high temperature is that previously high inflation perceptions flip sign, possible due to the extreme economic conditions.

To investigate this further, we implement the long-run average scenario from Table 1 for low and high temperatures. This is shown on the right-hand side of Figure 4. The great majority of responses are now concentrated in the moderate-inflation $(0, 5]$ -bucket. Increasing the temperature now leads to an “overflow” response to neighbouring buckets as we would expect from the working of the temperature parameter.¹² These observations suggest that there may be something peculiar about the economic conditions around the time of our survey samples, e.g. when compared to historic averages. We will return to this when investigating economic drivers and model sensitivity.

4.2 Time profile of inflation expectations

For both economic theory and policy making, inflation expectations are paramount. A crucial question is whether these are ‘anchored’ at about the central bank’s inflation target on longer horizons. We investigate the GPT inflation expectations’ time profile for the IAS 2023Q1 sample with the economic conditioning given in Table 1.

The time profile of aggregate GPT expectations for horizons of up to five years is shown in Figure 5 for $T = 1.5$.¹³ The close match of inflation perceptions (horizon zero) between GPT, IAS, and ONS out-turn matches the lower part of Figure 3. Both GPT (blue) and IAS (orange) expectations decrease with the horizon. However, there

¹¹This is even more the case for the cross-validation sample for which we show the corresponding figure in the Appendix.

¹²While increasing the temperature intuitively makes less likely responses more likely, this should not change the shape of the response distribution.

¹³The profile for $T = 0$ is qualitatively very similar and is given in the Appendix.

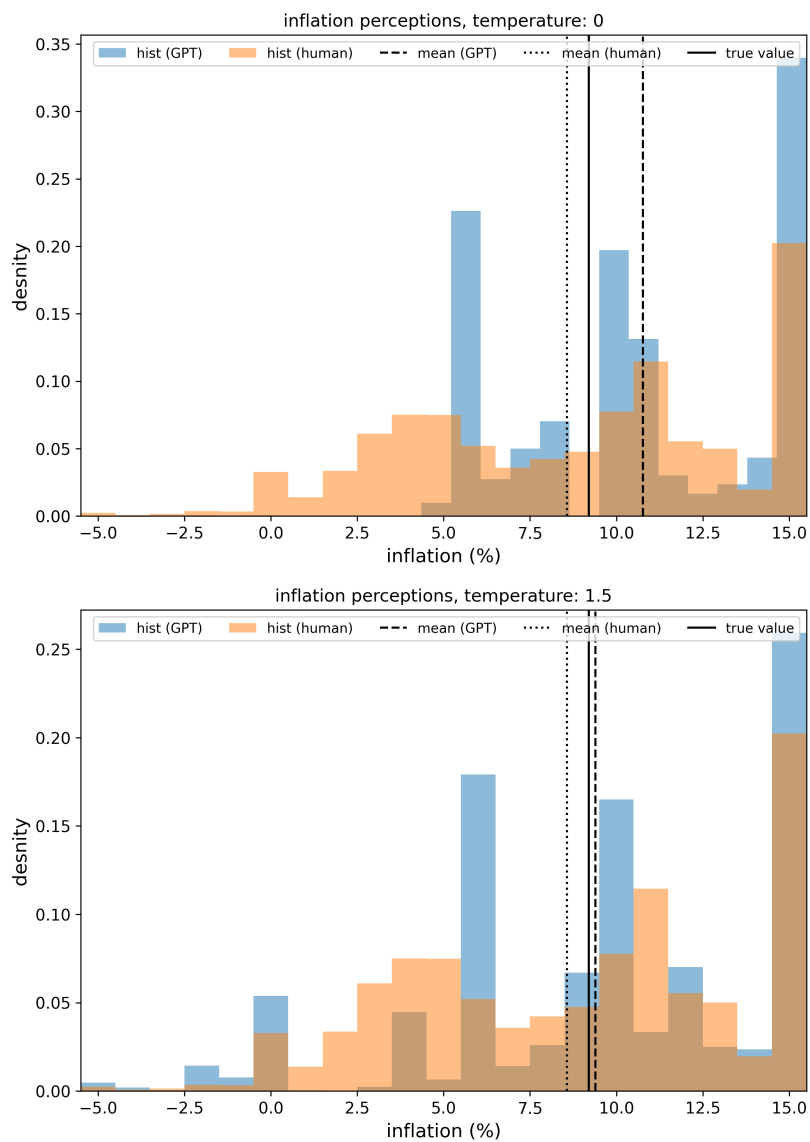


Figure 3: Histograms of human (sand) and GPT (blue) responses for conditioned inflation perceptions in the main sample (February 2023) for $T = 0$ (upper panel) and $T = 1.5$ (lower panel). Sources: IAS, ONS, authors' calculations.

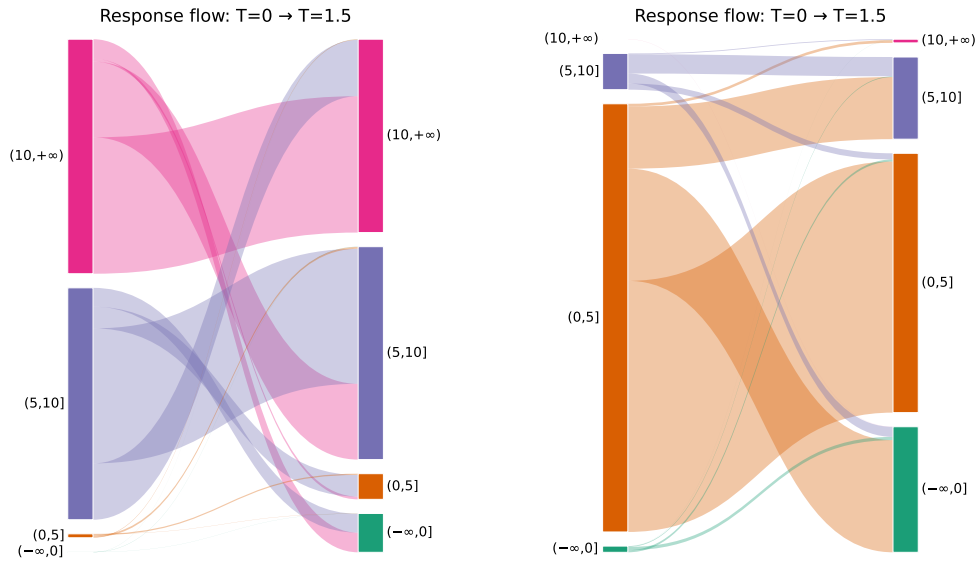


Figure 4: Response flows for inflation perceptions (%) between low and high temperatures for the main sample. Left: main scenario. Right: long-run average scenario (see Table 1). Sources: IAS, ONS, authors' calculations.

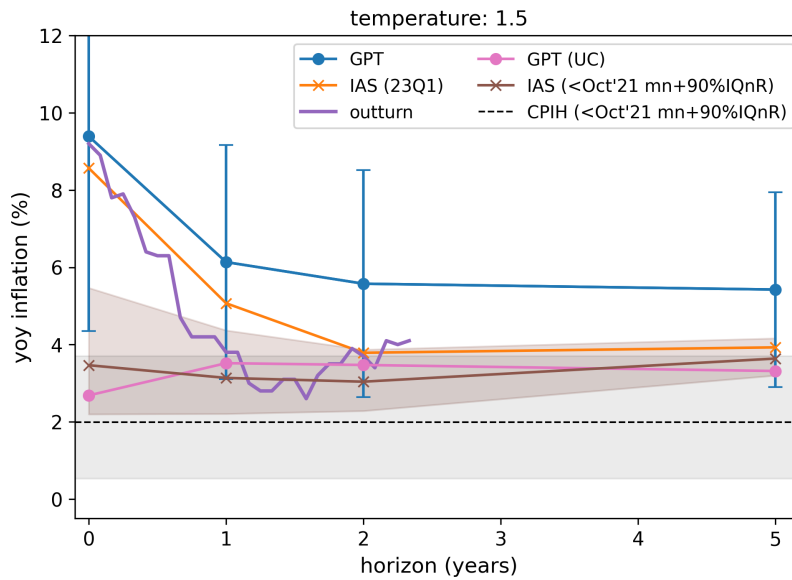


Figure 5: Time profile of aggregate inflation expectations for IAS and GPT responses ($T = 1.5$) for the main sample (2023Q1 survey): GPT conditioned (blue), GPT unconditioned (pink), IAS (sand), IAS pre-Oct-2021 mean & swath (brown), out-turn from February 2023 onwards (purple), and CPIH pre-Oct-2021 mean & swath. Sources: IAS, ONS, authors' calculations.

are major discrepancies between future horizons from two to five years out. Though somewhat elevated, IAS mean expectations are in line with the historical distribution (IAS swath) and at the upper bounds of historically observed year-on-year inflation (CPIH swath).¹⁴ In contrast, GPT expectations stay roughly constant and elevated beyond the one year horizon. This suggests caution when using GPT to infer inflation expectations.

Comprehensive summary statistics comparing GPT and IAS at different horizons and for a high ($T = 1.5$) and low ($T = 0$) temperature are given in Table 3. The right part of the table investigates the effect of including economic conditioning information in the prompt compared to only including demographics. The effect columns show the average treatment (2) for the mean and standard deviation of the GPT response distributions. The inclusion of economic effects shifts the mean upwards considerably. This was expected given that all treatment components are considerably above their historic averages, such that this is a sense check.

Looking at the last column of Table 3, we see that the width of the GPT response distribution has also increased as a consequence of the information treatment. This suggests that there are interactions between the demographics of the different personas given in the prompts and the economic conditioning. We will investigate this in the next section. However, we will first analyse the effect of temperature on GPT’s responses in more detail.

4.2.1 High versus low temperature

We see in Table 3 that there are considerable differences between the results for high and low temperatures. In line with the cross-validation results, the losses (9) are lower for $T = 1.5$. In almost all cases, the higher temperature setting matches both the mean and standard deviation of the IAS response distribution better.

However, setting a high temperature also has considerable drawbacks. First, there can be a considerable number of invalid responses, in our case particularly for longer horizons. It is not clear what is driving this result. A possible explanation may be that the questions for longer horizons are logically more challenging as they refer to year-on-year changes after a certain time has passed. This may complicate instruction following.

Second, the micro-level relation between GPT and IAS responses as measured by Pearson correlations is weak, volatile across horizons, and even partly breaks down. In contrast, the GPT-IAS relations are considerably stronger and actually increasing with the horizon for the zero temperature setting.

Lastly, the cross-horizon correlations are mostly weak for the high temperature case with patterns very different to human responses. The corresponding cross-correlations are listed in Table 4. GPT at zero temperature shows patterns much more similar to those of human responses, especially between expectations at longer horizons.

Because of these observations, we will focus on the $T = 0$ case below. This has the additional advantage that the interpretation of results does not carry uncertainty from the temperature setting, which can still be investigated separately.

¹⁴Historical back data always end in Sep-21 coinciding with GPT’s knowledge cut-off if not stated otherwise.

horizon	n_{miss}	MN	$diff_{MN}$	SD	$diff_{SD}$	L1-loss	L2-loss	pcc	pval	MN_{uc}	$effect_{MN}$	SD_{uc}	$effect_{SD}$
GPT ($T = 1.5$)													
0	13	9.39	0.82	5.04	0.10	0.46	0.34	-0.01	0.49	2.68	6.71	3.23	1.80
1	44	6.14	1.07	3.03	-1.81	1.44	2.21	0.04	0.01	3.52	2.62	1.67	1.36
2	84	5.58	1.79	2.94	-1.62	1.71	2.92	0.02	0.18	3.47	2.10	1.69	1.25
5	102	5.42	1.50	2.52	-2.24	1.87	3.62	0.04	0.01	3.32	2.11	1.51	1.01
GPT ($T = 0$)													
0	0	10.76	2.19	3.79	-1.15	1.67	3.07	0.03	0.06	3.18	7.58	1.75	2.04
1	0	6.26	1.19	2.09	-2.75	1.97	4.48	0.05	0.00	3.23	3.03	1.13	0.96
2	0	5.65	1.85	1.81	-2.76	2.31	5.52	0.07	0.00	3.17	2.48	1.05	0.77
5	0	5.36	1.42	1.76	-3.01	2.21	5.53	0.07	0.00	3.10	2.26	0.97	0.78

Table 3: Test statistics for GPT responses for different expectation horizons at $T = 1.5$ (upper panel) and $T = 0$ (lower panel): horizon in years (zero is current inflation perceptions), GPT missing values (NA responses), survey weighted mean (MN), difference to IAS mean, weighted sample standard deviation (SD), difference to IAS SD , $L1$ -loss of GPT difference to IAS mean and SD (equally weighted), same $L2$ -loss, Pearson correlation coefficient between GPT and IAS responses, corresponding p -value, weighted mean of unconditioned GPT responses (uc), difference in mean GPT responses, weighted SD of unconditioned GPT responses (uc), and difference in SD of GPT responses. Sources: IAS, authors' calculations.

GPT (T=1.5)				
h	0	1	2	5
0	1.00	0.08	0.08	0.13
1	-	1.00	0.82	0.20
2	-	-	1.00	0.16
5	-	-	-	1.00

GPT (T=0)				
h	0	1	2	5
0	1.00	0.18	0.18	0.19
1	-	1.00	0.73	0.64
2	-	-	1.00	0.74
5	-	-	-	1.00

IAS				
h	0	1	2	5
0	1.00	0.43	0.31	0.28
1	-	1.00	0.63	0.50
2	-	-	1.00	0.72
5	-	-	-	1.00

Table 4: Cross-horizon (h) Pearson correlation coefficients for GPT at $T = 1.5$ (upper panel), $T = 0$ (middle panel), and IAS responses (lower panel). Sources: IAS and authors' calculations.

4.2.2 Model time trends

We test our baseline inflation perceptions results on the main sample for two types of alternative models other than `gpt-3.5-turbo-0613`: (i) later releases from the same model family, and (ii) open-weight models with the same or earlier training cut-offs.

Later ChatGPT releases Consumer price inflation and the “cost of living crisis” have been persistent topics of debate in the UK since the 2022 inflation spike.¹⁵ That is, inflation has become a more salient topic in the public discourse. At the same time, LLM fine-tuning using reinforcement learning based on human user preferences has become common (Ouyang et al., 2022). Such fine tuning can lead to an unpredictable level of information leakage from beyond the stated training cut-off. To test whether there may have been such leakages or salience of high inflation in GPT’s model family, we test unconditioned inflation perceptions for the main sample for different models, where we remove the economic conditioning information from the prompt only leaving demographics and the survey instructions.

The results of this exercise are summarised in Figure 6, which plots average inflation perceptions against the release dates of the different models. We see that there is a clear time trend with an increase of about 3 p.p. unconditional inflation perceptions per year, which potentially creates large biases in later models as inflation numbers quickly fell after the spike.

Models with later release dates but the same stated knowledge cut-off (September 2021) seem to be affected by this trend suggesting the use of reinforcement learning and subsequent information leakage. Models released in early 2023 seem not to be affected by this. This highlights the value of the experimental setting for model validation and testing presented here.

It also underscores the challenges for the use of closed-weight proprietary models like those from the ChatGPT family. We therefore consider a set of alternative open-weight model with an information cut-off compatible with our setting.

Open-weight models As indicated by the results in Figure 6, not being out-of-time may lead to considerable biases. Therefore, we repeated our base experiment on inflation perceptions in the main sample with six open-weight alternative models with up to 30 billion parameters and with the correct out-of-time properties (see Appendix). All show a variety of undesirable behaviour, although, as expected to a lesser extent as the model increases in size and if it has been instruction tuned.

In summary, alternative models have either been trained earlier and have problems with output quality and instruction following, or their having been trained later has exposed them to newer data in pre- or post-training, and so they are not out-of-time leading to biases. While this situation is idiosyncratic to our exercise which requires a training cut-off towards the end of 2021, it also points to a general issue in LLM research.

There will be a tendency to use the current state-of-the-art models for forward looking and operationalised tasks, due to their quality and ease of use. However, back-testing models would preferably be of a similar strength and functionality. This highlights the recognised need to continue work (He et al., 2025; de Araujo et al., 2025) to provide

¹⁵<https://commonslibrary.parliament.uk/research-briefings/cbp-9428/>

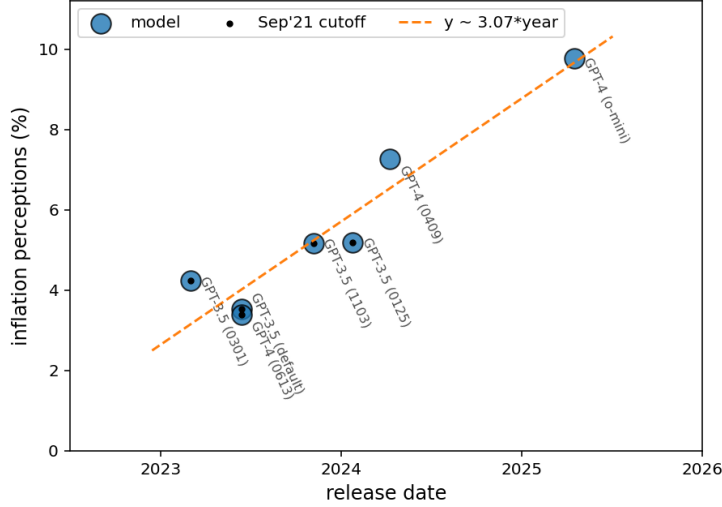


Figure 6: Time trend of unconditioned inflation perceptions for different GPTs. Each circle is the simple mean of a subsample of the same 200 respondents from the main sample for 9 answer permutations for each model. Models with an official knowledge cut-off of September 2021 are marked with a dot. Sources: Authors’ calculations.

high quality, generative, models trained with a range of strictly enforced date cut-offs to allow out-of-time back-testing experiments.

4.3 Demographic drivers

When driven by food and energy price growth, high inflation may be more concerning for lower income households, as they spend a larger proportion of their consumption basket on necessities like food and energy. By similar arguments, the inflation rate experienced by different demographic groups may vary.

Now, for a demographic variable D_c we hypothesise the relative ordering of experienced inflation values $\mathcal{O}_c = \{\pi_{c_1} > \dots > \pi_{c_j} > \dots > \pi_{c_K}\}$ for $j \in \{1, \dots, K\}$ different levels of that variable, for example based on assumptions about their respective consumption baskets in a given economic environment. Taking $j = 1$ as the reference class for each D_c , we can formulate the following hypothesis tests for jointly assessing \mathcal{O}_c for categories $c \in \{1, \dots, C\}$ based on the response model

$$y_{i,s} = b + \sum_{c=1}^C \sum_{j=2}^K \beta_{c_j} d_{ic_j} + \epsilon_i \quad \text{with} \quad \mathcal{H}_{c_j} : \beta_{c_j} < 0 \text{ for } j > 1, \quad (10)$$

where s denotes the source of the y_i (IAS or GPT in our case), b is a constant, and d_{ic_j} is a vector of dummies encoding the demographic profile of subject i . The coefficient vector $\beta = (\beta_{1_2}, \dots, \beta_{C_K})'$ captures the joint relation between demographics and survey responses given a scenario. The interpretation of the elements of β is intuitive: by how many percentage points (p.p.) inflation perceptions or expectations are on average higher or lower (depending on the sign of an element) if a subject belongs to a certain demographic group, i.e. $d_{ic_j} = 1$.

An appealing property of Eq. 10 is that it adjusts for a different location and scale of the response distribution y_s via b and the magnitude the β s, respectively. This means that, despite potentially poorer aggregate fits of a response distribution to a benchmark, we still can make inferences about its drivers. This will be the case for the GPT ($T = 0$) cases, which showed poorer aggregate matches to the IAS distribution but stronger correlations on the micro level (see Table 3).

In the context of the main sample scenario in Table 1, we hypothesise that the following reference groups within our demographic categories have experienced the highest levels of inflation: income: *less than £9999 (lowest income)*, housing: *council house*, age: *16-24 (youngest)*, social class: *working class*, education: *GCSEs but not A-levels (lowest formal education)*, region: *Scotland*. Consequently, when fitting Eq. 10 we expect all components of β to be negative.

4.3.1 Inflation perceptions

Table 5 summarises the results for the IAS and again for GPT for $T = 1.5$ and $T = 0$ for inflation perceptions. We also include estimates of the actual inflation experienced by the different demographics categories at the time of the survey based on official statistics (ONS reference).

Focusing on the $T = 0$ case, we see that the majority of the coefficients are indeed negative and statistically highly significant indicating that GPT’s responses are in line with the economic intuition guiding the choice of the reference classes, i.e. our hypotheses about which demographic groups may be more affected by the inflation spike around the time of the survey.

We can further validate this intuition and GPT’s responses by comparing the β estimates with the ONS reference. We indeed see that GPT responses are very much aligned with actual realisations for most categories when comparing directions: most entries are negative.¹⁶ The exception to this are regional differences, where GPT thought they would be large, there are almost none, perhaps because energy prices are regulated and determined mostly on the national and not regional level.

Comparing GPT responses to the IAS, we see that most coefficients are negative again. However, there is a major discrepancy between the two for age. While GPT thinks that there is a mostly negative and marginally increasing effect with age, IAS respondents show an effect which is clearly positive and strongly increasing with age. These findings are in line with a memory based explanation as proposed in Braggion et al. (2025) who argue that households living in areas with higher local inflation during the hyperinflation of the 1920s expect higher inflation today, rather than what Malmendier and Nagel (2016) suggested about agents overweighting the very recent past when setting inflation expectations.

The demographic regression results are graphically summarised for the first three categories of Table 5 in Figure 7. The left-hand side panels compare IAS and GPT coefficients against ONS out-turns as measured by the difference against the reference

¹⁶A notable exception is the coefficient of the ONS reference point for the lowest income group which is most likely is due to changes in the definition of lowest income group within the IAS survey itself over the years.

variable	IAS	GPT (T=1.5)	GPT (T=0)	ONS ref.
income (£10000-£19999)	-0.25	0.91**	-0.17	0.16
income (£20000-£34999)	-0.29	0.55	-1.08***	-0.20
income (£35000-£44999)	-0.44	0.49	-1.26***	-0.35
income (>£45000)	-0.30	0.67	-0.74**	-0.68
housing (private renter)	0.36	-0.62*	-0.37	-0.27
housing (owner w/mortgage)	-0.10	-0.73**	-1.26***	-1.54
housing (outright owner)	-0.59*	-0.66*	-0.99***	-1.39
age (25-34)	1.47***	-0.67**	-1.01***	-0.12
age (35-44)	2.53***	-0.90***	-0.89***	-0.28
age (45-54)	3.96***	-1.00***	-0.83***	-0.46
age (55-64)	5.02***	-0.64*	-0.43*	-0.58
age (65-75)	4.91***	-0.20	0.51*	-0.39
class (skilled working)	-0.87***	-0.42	-0.36	-0.14
class (lower middle)	-0.50*	-0.39	-0.39*	-0.52
class (upper middle)	-0.75**	-0.87***	-0.80***	-0.62
class (pensioner)	-1.39**	-2.08***	-1.88***	-0.28
education (A-levels)	0.45*	-0.43	-0.18	0.06
education (degree)	0.18	0.09	-0.00	0.08
region (North or Northern Ireland)	0.30	0.23	-0.18	0.09
region (West of England or Wales)	-0.07	0.31	-0.53*	0.02
region (Midlands)	0.02	0.60	-0.56**	0.03
region (South East)	-0.39	0.59	-0.82***	0.30
gender (female)	0.04	0.63***	0.73***	-
gender (other)	0.24	2.18*	0.10	-
gender (undisclosed)	-0.57	-1.43	1.01	-
constant	6.39***	9.85***	13.39***	-
n obs.	3808	3799	3808	-
R^2 (adj.)	0.11	0.01	0.07	-

Table 5: Summary table for weighted least square regressions of individual survey responses (inflation perceptions) for IAS, GPT (T=1.5), and GPT (T=0) on demographic dummies against base classes (housing: council house, income: income of <£9999, age: 16-24, social class: working class, education: GCSEs but not A-levels, region: Scotland, gender: male). *** indicates significance at the 1% level, ** at the 5% level, and * at the 10% level. Robust standard errors have been used. The right-most column states ONS reference values for February 2023 inflation as the differences in percentage points against the corresponding base class. Sources: IAS, ONS, and authors' calculations.

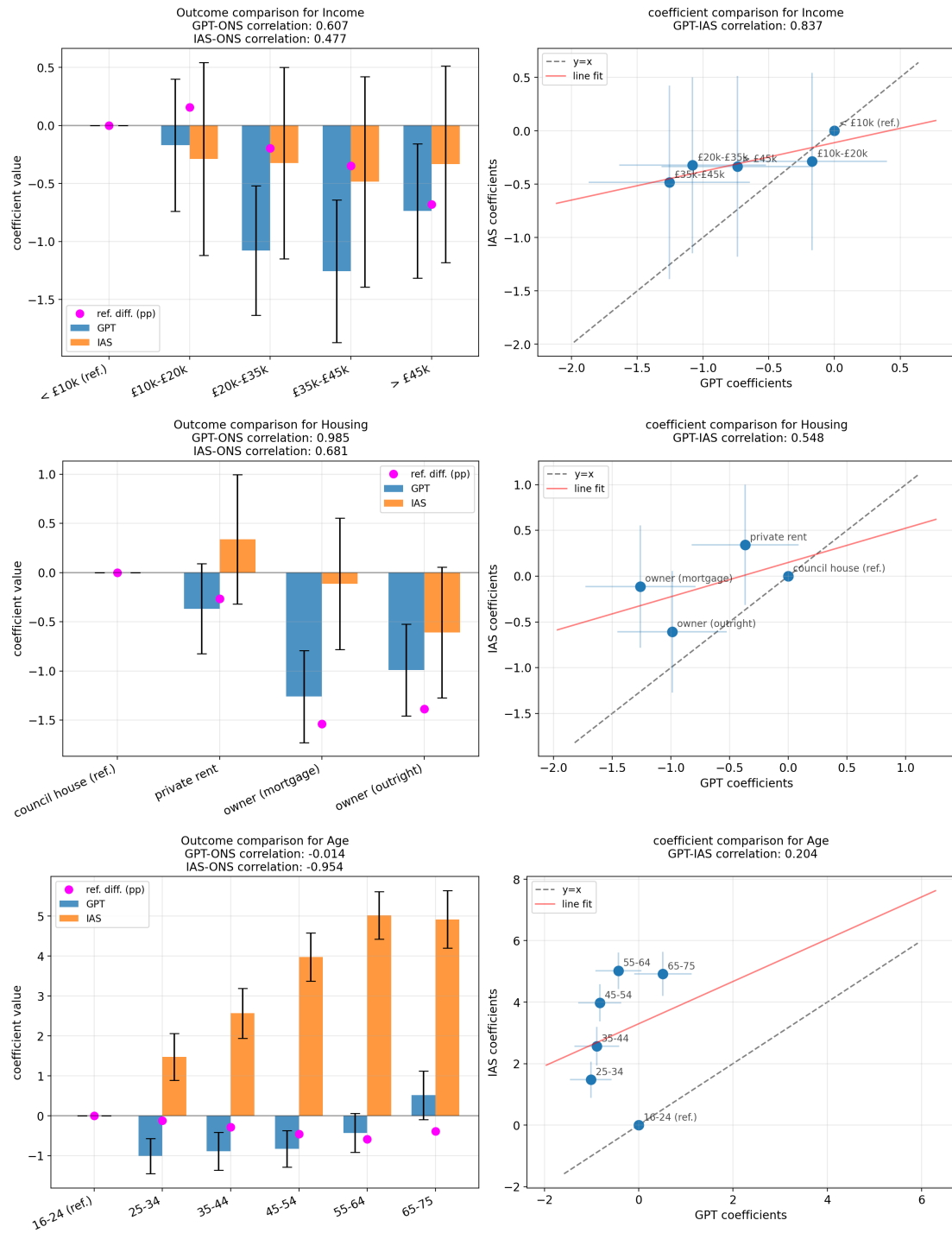


Figure 7: Graphical summaries of demographic regressions of GPT ($T = 0$) inflation perceptions for income (upper panel), housing tenure (middle panel), and age (lower panel). LHS: three-way comparison between IAS (sand), GPT (blue), and ONS (magenta). RHS: scatter plot of IAS coefficients versus GPT coefficients. Best-fit line given in red. Error bars indicate 95% confidence intervals throughout. Summary stats are Pearson correlation coefficient (corr) and cosine similarity. Sources: IAS, ONS, authors' calculations.

class in p.p.. The right-hand side panels compare GPT and IAS coefficients against each other on a scatter plot: the more dots align along the diagonal as summarised by the fit line, the more similar human and GPT responses are alongside this demographics dimension.

We see that there is reasonable agreement between GPT, IAS, and ONS out-turns for income and especially for housing. GPT’s response patterns closely match that of the outcome. This is interesting and encouraging as there are various channels by which different housing tenures can affect realised inflation values suggesting that GPT can be useful in modelling complex and potentially unknown relationships.

However, our analysis also suggests caution when looking at the results for age in the bottom part of Figure 7. Here the three-way comparison is far less aligned. As pointed out before, GPT and IAS response patterns are mostly opposed to each other with some similarity in trends with increasing age. The alignment with the out-turn is weak at best.¹⁷

This example raises the important question how to interpret and use GPT as a modelling tool: should we prefer GPT to be more closely aligned with the human sample we emulate or with the survey quantities we elicit? We will address such concerns in the discussion at the end of the paper.

4.3.2 Inflation expectations

Perceptions are easier to compare with realised values as there is less noise affecting the analysis between eliciting survey responses and measuring the reference quantities. However, many economic concepts relate to expectations at longer horizons, often at one quarter or one year out. We therefore repeat the exercise from the previous subsection for inflation expectations at the one-year horizon. The results are summarised in Table 6, where we will focus on the $T = 0$ case.

Looking at explanatory power (R^2), it is interesting to observe that GPT now shows more alignment with demographics on average when compared to the IAS. The comparison with realised values (last column of Table 6) is less interesting now because aggregate inflation decreased markedly from 9.2% (February 2023) to 3.8% (February 2024), and the variation within demographic groups is much smaller compared to when inflation peaked.

The patterns of signs persists between inflation perceptions and one-year expectations: most signs are negative, again with exception for age for the IAS. Looking at individual demographics in Figure 5, we see that there is good alignment between GPT and the IAS for income, housing, and social class. This congruence is particularly strong for the income distribution, a main quantity of interest when investigating heterogeneity in economic analyses.

4.4 Economic drivers

We investigate how the details of the economic conditioning affect GPT’s ($T = 0$) predictions of inflation perceptions in the main sample.

¹⁷The corresponding figure is given in the Appendix for $T = 1.5$, where see that micro-level relations generally weaken with higher temperature.

	IAS	GPT (T=1.5)	GPT (T=0)	ONS ref.
income (£10000-£19999)	-0.58	0.20	-0.41**	0.03
income (£20000-£34999)	-0.77*	-0.17	-0.73***	-0.05
income (£35000-£44999)	-1.12***	-0.11	-0.83***	-0.01
income (>£45000)	-1.17***	-0.53*	-0.82***	-0.03
housing (private renter)	-0.18	-0.98***	-0.78***	0.21
housing (owner w/mortgage)	-0.35	-0.89***	-0.88***	-0.16
housing (outright owner)	0.48	-0.45**	-0.41***	0.33
age (25-34)	2.35***	-0.07	-0.14	0.10
age (35-44)	0.77***	-0.27	-0.33**	0.24
age (45-54)	2.05***	0.49**	-0.05	0.26
age (55-64)	1.54***	1.06***	0.21	0.28
age (65-75)	1.19***	-0.04	-0.22	0.37
class (skilled working)	-0.92***	-0.06	-0.19	0.10
class (lower middle)	-0.9***	-0.47**	-0.68***	0.01
class (upper middle)	-1.01***	-0.79***	-1.19***	-0.07
class (pensioner)	-0.92	-2.36***	-1.24***	0.18
education (A-levels)	0.38	-0.07	-0.06	0.05
education (degree)	0.53**	-0.58***	-0.21**	0.01
region (North or Northern Ireland)	-0.19	-0.52**	0.19	0.08
region (West of England or Wales)	-0.50	-0.47*	-0.09	-0.01
region (Midlands)	-0.41	-0.52**	-0.11	0.06
region (South East)	-0.46	-0.27	0.03	0.11
gender (female)	-0.11	0.18*	0.20**	-
gender (other)	0.48	-0.42	-0.13	-
gender (undisclosed)	0.03	0.70	-0.16	-
constant	5.51***	7.94***	8.29***	-
n obs.	3636	3595	3636	-
R^2 (adj.)	0.03	0.09	0.14	-

Table 6: Summary table for weighted least square regressions of individual survey responses (inflation expectations) for IAS, GPT (T=1.5), and GPT (T=0) on demographic dummies against base classes (housing: council house, income: income of <£9999, age: 16-24, social class: working class, education: GCSEs but not A-levels, region: Scotland, gender: male). *** indicates significance at the 1% level, ** at the 5% level, and * at the 10% level. Robust standard errors have been used. The right-most column states ONS reference values for February 2023 inflation as the differences in percentage points against the corresponding base class. Sources: IAS, ONS, and authors' calculations.

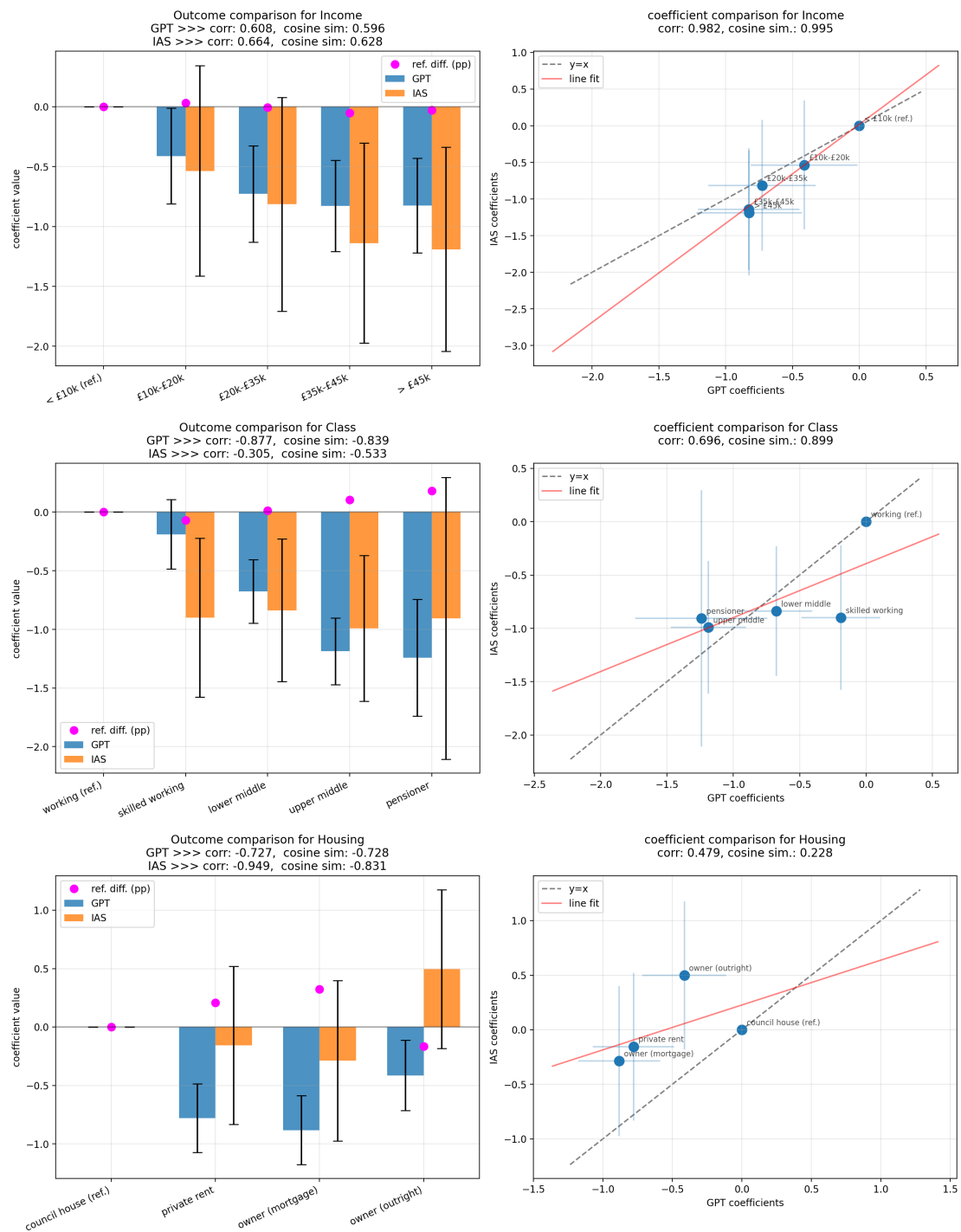


Figure 8: Graphical summaries of demographic regressions of GPT ($T = 0$) inflation expectations for income (upper panel), social class (middle panel), and housing tenure (lower panel). LHS: three-way comparison between IAS (sand), GPT (blue), and ONS (magenta). RHS: scatter plot of IAS coefficients versus GPT coefficients. Best-fit line given in red. Error bars indicate 95% confidence intervals throughout. Summary stats are Pearson correlation coefficient (corr) and cosine similarity. Sources: IAS, ONS, authors' calculations.

4.4.1 Marginal sensitivity

Aggregate inflation scales approximately linearly with inflation of the subcomponents of the indices.¹⁸ As such, it is an interesting question whether GPT’s inflation perceptions scale linearly with the different components of the economic conditioning information treatment. To investigate this, we track the naïve treatment effect (4) for food & restaurants, energy, and the bulk component with the offset being the respective aggregate prediction value at the zero input for that component.¹⁹ In each case, we vary the respective component covering the historically observed values including the inflation spike in 2023Q1.²⁰

The results of this exercise are shown in Figure 9. We again make several observations. First, GPT’s aggregate perceptions indeed scale linearly with subcomponent inflation over large parts of the input space. However, this scaling does not consistently cover the bulk of the historically observed input ranges (shaded areas) for some components (energy and bulk), but may extend far beyond them (food and bulk). GPT’s behaviour is particularly interesting for the bulk component. It is mostly insensitive to this component between zero and two percent inflation but then scales linearly far beyond historically observed values.

Second, the slopes within the linear response ranges of GPT for the different components can be directly compared to the corresponding basket weights. This comparison is shown in the upper part of Table 7, where the ratio between the slope and the weight measures whether GPT over- or under-reacts to a component corresponding to a value bigger or smaller than one, respectively. GPT strongly over-reacts to information from the food & restaurant and the energy components. This behaviour is qualitatively similar to that of humans, since it is known that they over-rely on salient components such as groceries to make judgments about inflation (D’Acunto et al., 2021). Interestingly, GPT’s responsiveness to the main bulk component is very much proportional to its share in the basket beyond the 2% point in the input range, going far beyond the historically observed range. This suggests that GPT can partially be used to make realistic inferences about inflation once corrections for the observed bias at the origin are included.

Lastly, the scenario from the main sample is rather extreme. The vertical purple lines which mark the economic conditioning are far beyond the historically observed data ranges and also outside GPT’s linear scaling range for food & restaurants and energy. It is somewhat interesting that GPT fails to extrapolate monotonically with respect to food & restaurants inflation. As such, our analysis sketches out the boundaries within which GPT’s responses may be judged as consistent but also highlights that GPT does not seem to have a complete internal model of consumer price inflation.

¹⁸Approximately because the basket-weighted sum of component price growth does not need to be the same as overall index growth but will be a good approximation in most cases.

¹⁹For example, we subtract $MN(g(t = (0, 0, \bar{t}_e, \bar{t}_b); x_i); w)$ for the food & restaurants component.

²⁰For food & restaurants, we varied food price inflation keeping the ratio between this and restaurants & cafes inflation constant at the value of the main scenario.

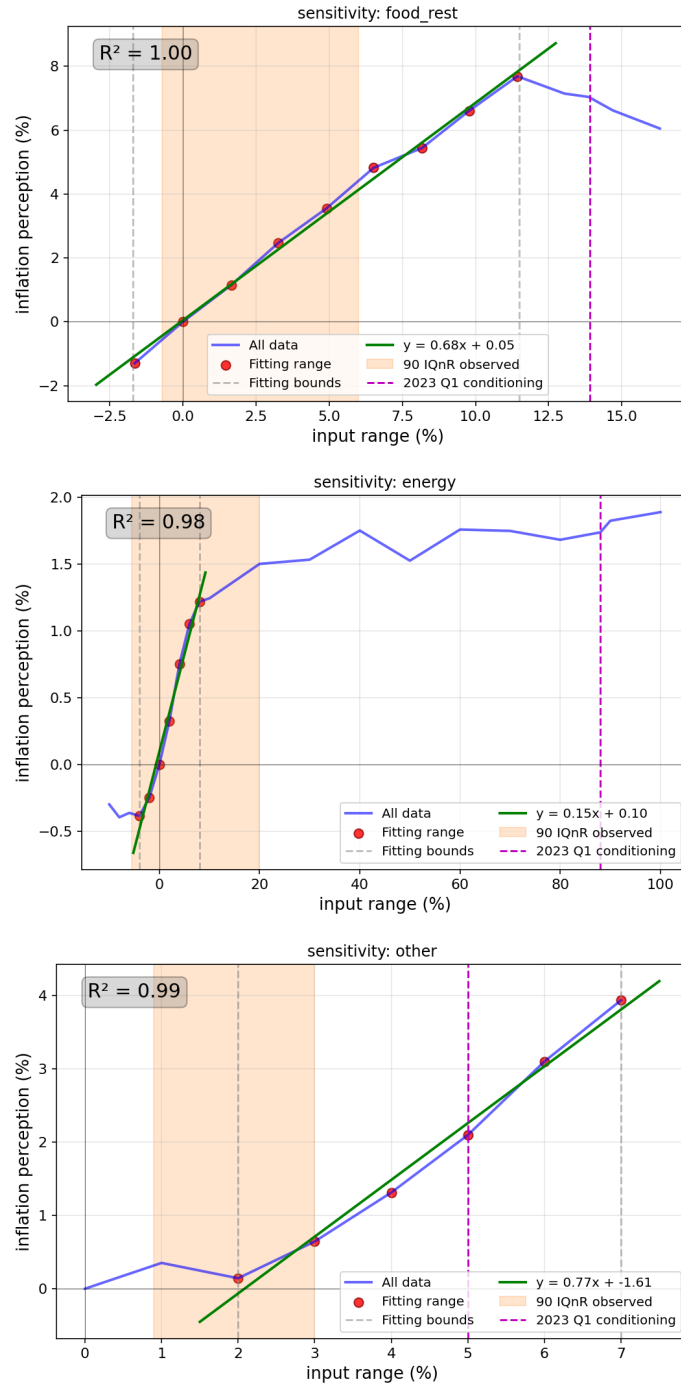


Figure 9: Sensitivity analysis of GPT ($T = 0$) inflation perceptions for varying individual index components across their respective observed ranges: food & restaurants (upper panel, x-axis is index-weighted mean of both components), energy (middle panel), and the remaining bulk component. The shaded area covers the inter-90 percentile range of each component pre-Oct-2021. The purple dashed lined marks the economic conditions used in the main experiment. Sources: IAS, ONS, authors' calculations.

component	food_rest	energy	bulk	sum
slope	0.68	0.15	0.77	1.60
weight	0.17	0.04	0.79	1.00
ratio	4.01	3.55	0.98	(1.60)
naïve	7.04	1.74	2.10	11.87
Shapley	6.18	2.79	1.76	11.73

Table 7: Summary of sensitivity analysis of GPT ($T = 0$) inflation perceptions with respect to different subcomponents (columns). Upper part: regression analysis from Fig.9. Lower part: contributions to main scenario treating components with independent (naïve) or possible interacting (Shapley) treatment component. The right-most columns provides the sum of columns except for the ratio. The mean GPT ($T = 0$) inflation perception is 11.80. Sources: IAS, ONS, authors' calculations.

4.4.2 Shapley decomposition

In each panel of Figure 9, the value on the vertical axis where the 2023Q1 conditioning line intersects with the GPT response curve measures the naïve treatment effect (4) for that component. As discussed, this decomposition does not take into account possible interactions between the treatment components. These are captured by the Shapley components (3). A comparison between the two indicates how important such interactions are. Both are shown in the lower part of Table 7. The Shapley decomposition measures all components against the zero baseline, i.e. where non-active component values are set to zero.

The ranking of components in the main scenario is the same. However, we do see sizeable changes for the importance of the energy treatment, which increases by more than 50% in absolute terms for the Shapley case. This suggests that GPT's inflation perceptions are particularly elevated due to jointly high values of food and energy price inflation. This is confirmed by considering a scenario with only these two components at non-zero values as in the main scenario. The effect for this scenario is 11.09% of aggregate GPT inflation perceptions. The remainder compared to the full main scenario is accounted for by the bulk component and its interactions with the other two components.²¹

The relevance of a component for aggregate inflation is the product of its magnitude and weight. For instance, energy inflation was very high in the main scenario while its weight is moderate at about 4% in the consumption basket. By comparing GPT aggregate components with realised contributions we can assess whether it over- or under-weights individual components for a given scenario independent of its response function. This comparison is shown for the main scenario in Figure 10. In line with previous results, GPT heavily over-weights food & restaurant inflation and under-weights the main bulk component. Its perception of the relevance of energy price inflation may be said to be roughly in line with its actual importance.

²¹The fact that neither the naïve nor the Shapley decomposition with zero values for inactive components perfectly sums up to the aggregate total of inflation perceptions in the main scenario suggests the existence of additional, though minor, inconsistencies in GPT's inflation attitudes.

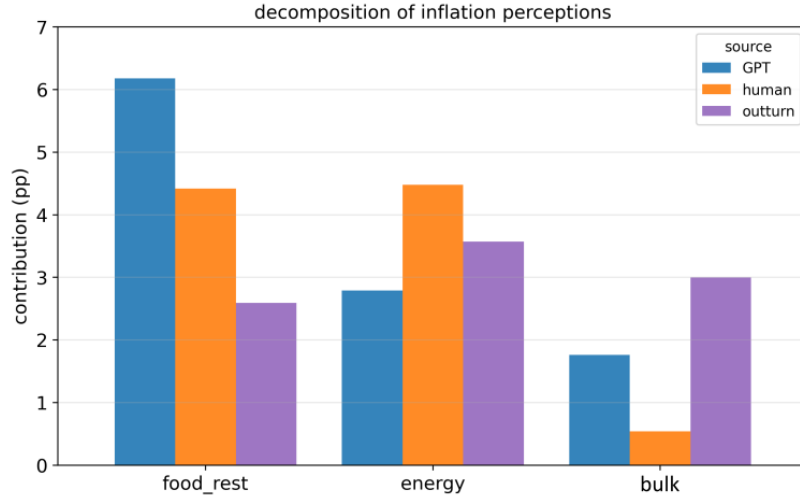


Figure 10: Component decomposition for the main scenario for GPT ($T = 0$): Shapley values against zero reference value (blue; $\tau_0^S = 0.07$); IAS: regression model (sand), and index weighted subcomponent inflation (purple). Sources: IAS, ONS, authors' calculations.

Finally, an advantage of the GPT decomposition approaches presented here is that we can make such inference in the first place. To get to something comparable for the actual survey, we need to build a model of how human respondents react to economic conditions. A possible approach to do so is to regress aggregate survey responses on the economic conditions at the time of the survey across a certain time period. We can then measure the importance of each component for a given scenario for this model by multiplying the coefficient values with the respective scenario input values, which corresponds to the Shapley value of a variable in a linear model.

Modelling results for different fit ranges are shown in Table 8. We see that such relations can be rather unstable over time. Only the relevance of the food & restaurant component is stable over time.²² The Shapley decomposition corresponding to the first row of Table 8 is shown in orange in Figure 10. This suggests that humans and GPT overreact to food & restaurant inflation in line with previous findings, which lends some support to the use of GPT in this context.

fit range	food_rest	energy	bulk
2011Q2 - 2021Q2	0.32***	0.05**	0.11
2006Q2 - 2021Q2	0.30***	-0.01	0.51***
2002Q1 - 2021Q2 (max)	0.31***	-0.02*	0.67***

Table 8: Regression coefficients for responsiveness of mean IAS inflation perceptions to CPIH subcomponent inflation for different time periods. *** indicates significance at the 1% level, ** at the 5% level, and * at the 10% level. Robust standard errors have been used.

²²All coefficients turn insignificant for a short modelling period of only five years.

5 Discussion

We investigated an LLM’s ability to form perceptions of current and expectations of future consumer price inflation based on a set of economic conditioning information in a representative survey setting. Our analysis provided a set of tools to assess model sensitivity and explainability with respect to the conditioning information. We found that the GPT model can reproduce key characteristics of human responses and official statistics. This is an impressive feat, as such models have not been explicitly designed to perform the tasks investigated here and no specific fine tuning was applied.

However, our analysis also revealed several model shortcomings in the given context. Similarly to human responses, GPT overweights the importance of salient price components, like those for food. This could be an advantage depending on the application. However, it also shows limited sensitivity to the majority of the consumption basket. Furthermore, component sensitivities can exhibit strong non-linearities even within historically observed ranges of a kink-type nature for no apparent reason. As such, GPT does not appear to have a fully consistent world model for consumer price inflation. Additionally, GPT responses’ correlation with those of actual survey respondents is relatively weak. This latter observation may largely be caused by the very limited conditioning information we provided to the model, which will certainly be insufficient to capture the idiosyncratic state of human respondents well. Achieving better alignment on the micro level is an interesting research problem in itself leaving plenty of scope for future work.

These inconsistencies temper the more optimistic reading—that LLM agents integrate macroeconomic information in an approximately Bayesian, human-like manner (Zarifhonorvar, 2026)—by showing that aggregate or updating-level fidelity can coexist with the absence of a coherent underlying model of inflation. We read this not as a contradiction but as a reason to pair any applied use with the kind of validation and explainability checks we set out here.

We conclude our analysis with a discussion of ethical considerations when using LLMs to simulate or investigate human behaviour.

One concern may be the simulation of specific individuals, here survey subjects, without their explicit consent. This aspect needs to be addressed by data governance and is independent of the model being used. This concern applies to conventional models as to a more complex neural network, such as an LLMs.

However, the more complex nature of LLMs leads us to a second concern which is more difficult to address. Namely, the black box critique of machine learning models. This is even more aggravated in the LLM case compared to supervised learning approaches. Model complexity (like the number of parameters in the neural network) is orders of magnitude larger compared to those used in “traditional ML” problems. At the same time, the training data (often vast corpora of unstructured data) are mostly unknown and inaccessible to the user, as are the details of the training algorithms used to build the model.

This second point puts more burden on model validation and testing. Much of this paper was dedicated to just that. However, as we have seen, the comparison with human benchmark data (actual survey responses) and official statistics can still lead to difficult questions regarding their interpretation and use, because these three may not be aligned. LLM outputs may match human benchmark data, but fail to reproduce

other statistics, or vice versa. There are many reasons for why this can happen, and it may not be possible to address these in the analysis leaving us with the difficult decision of whether and how to use LLM results, for instance to inform decisions.

Our analysis can help to provide some guidance here. If the aim is to replicate or analyse official statistics, such as macroeconomic aggregates, mismatch on more disaggregate levels can be tolerated if validation has been sufficiently satisfactory by some criteria. As pointed out, mismatch on the micro level does not preclude match on the aggregate level, and it may actually be a source of better aggregate performance.

However, if the goal of the analysis is to infer characteristics of humans or subgroups of the population, like demographic groups, potentially informing decisions affecting them, mismatch may not be tolerable if it cannot be accounted for in some way, such as via bias correction when using LLM outputs in a downstream model (Ludwig et al., 2025). A decision on the usefulness of LLM outputs will most likely face trade-offs even after validation and possible corrections, necessitating a robust governance structure for the use of AI in decision making.

Finally, most of our results were obtained for a single model without a guarantee that these generalise to other models or settings. LLMs are advancing fast, and newer and more capable models may be better suited to simulate human subjects, however, the usual metrics of goodness and commercial development efforts are generally not in this direction. Furthermore, we have seen how model selection can create biases via information leakage. It is therefore essential to validate a model whenever either the model or setting changes since we do not have a general or pre-emptive understanding of how LLM behaviour in a given context. If validation has been satisfactory, the use of LLMs (or a particular LLM) opens new doors for research given their ability to simulate complex entities, like survey subjects, in-situ, cheaply and flexibly in a way which would be impossible in real-world settings.

References

- Acemoglu, D., Autor, D., Hazell, J., and Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, 40(S1):S293–S340.
- Aher, G., Arriaga, R., and Kalai, A. T. (2022). Using large language models to simulate multiple humans. *arXiv*, 2208.10264.
- Anesti, N., Esady, V., and Naylor, M. (2025). Food prices matter most: sensitive household inflation expectations. Bank of England, Staff Working Paper 1125, Bank of England.
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.
- Arioli, R., Bates, C., Dieden, H., Duca, I., Friz, R., Gayer, C., Kenny, G., Meyler, A., and Pavlova, I. (2017). EU Consumers’ Quantitative Inflation Perceptions and Expectations: an Evaluation. Technical Report 186, ECB Occasional Paper.
- Arora, A., Kaffee, L.-a., and Augenstein, I. (2023). Probing pre-trained language models for cross-cultural differences in values. In Dev, S., Prabhakaran, V., Adeli, D. I., Hovy, D., and Benotti, L., editors, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.
- Bai, X., Wang, A., Sucholutsky, I., and Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122.
- Bernanke, B. (2007). Inflation expectations and inflation forecasting. *Speech at the Monetary Economics Workshop of the National Bureau of Economic Research Summer Institute, Cambridge, Massachusetts*.
- Braggion, F., von Meyerinck, F., Schaub, N., Weber, M., and Weber, M. (2025). The long-term effects of inflation on inflation expectations. CESifo Working Paper Series 11679, CESifo.
- Brookins, P. and DeBacker, J. M. (2023). Playing games with GPT: What can we learn about a large language model from canonical strategic games? *Available at SSRN 4493398*.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Bruine de Bruin, W., Manski, C. F., Topa, G., and Van Der Klaauw, W. (2011). Measuring Consumer Uncertainty about Future Inflation. *Journal of Applied Econometrics*, 26(3):454–478.

- Buckmann, M. and Hill, E. (2025). Improving text classification: Logistic regression makes small llms strong and explainable’tens-of-shot’classifiers. Staff Working Paper 1,127, Bank of England.
- Buckmann, M. and Joseph, A. (2023). An interpretable machine learning workflow with an application to economic forecasting. *International Journal of Central Banking*, 19(4):449–522.
- Buckmann, M., Nguyen, Q. A., and Hill, E. (2025). Revealing economic facts: LLMs know more than they say. Staff Working Paper 1,150, Bank of England.
- Bybee, L. (2023). Surveying generative ai’s economic expectations. *arXiv*, 2305.02823.
- Cavallo, A., Cruces, G., and Perez-Truglia, R. (2017). Inflation Expectations, Learning, and Supermarket Prices: Evidence from Survey Experiments. *American Economic Journal: Macroeconomics*, 9(3):1–35.
- Charness, G., Jabarian, B., and List, J. A. (2025). The next generation of experimental research with llms. *Nature Human Behaviour*, 9(5):833–835.
- Chen, W. X., Srinivasan, S., and Zakerinia, S. (2024). Displacement or Complementarity? The Labor Market Impact of Generative AI. Technical report, Harvard Business School Working Paper, No. 25-039.
- Christie, F., O’Connor, S., and Kruschwitz, W. (2024). Large language models can outperform humans in social situational judgments. *Scientific Reports*, 14(27449).
- Coibion, O. and Gorodnichenko, Y. (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8):2644–78.
- Coibion, O., Gorodnichenko, Y., Kumar, S., and Pedemonte, M. (2020). Inflation Expectations as a Policy Tool? *Journal of International Economics*, 124:103297.
- Coibion, O., Gorodnichenko, Y., and Weber, M. (2022). Monetary Policy Communications and Their Effects on Household Inflation Expectations. *Journal of Political Economy*, 130(6):1537–1584.
- Crane, L. D., Karra, A., and Soto, P. E. (2025). Total recall? Evaluating the macroeconomic knowledge of large language models. Working Paper 2025-44, Board of Governors of the Federal Reserve System.
- D’Acunto, F., Hoang, D., Paloviita, M., and Weber, M. (2019). Cognitive Abilities and Inflation Expectations. In *AEA Papers and Proceedings*, volume 109, pages 562–566.
- D’Acunto, F., Hoang, D., Paloviita, M., and Weber, M. (2020). Effective Policy Communication: Targets versus Instruments. Technical Report 20-38, Chicago Booth Research Paper.
- D’Acunto, F., Malmendier, U., Ospina, J., and Weber, M. (2021). Exposure to grocery prices and inflation expectations. *Journal of Political Economy*, 129(5):1615–1639.

- D’Acunto, F. and Weber, M. (2024). Why survey-based subjective expectations are meaningful and important. *Annual Review of Economics*, 16:329–357.
- de Araujo, D. K. G., Bokan, N., Comazzi, F. A., and Lenza, M. (2025). *Word2Prices: embedding central bank communications for inflation prediction*. Number 3047. ECB Working Paper.
- De Bruin, W. B., Van der Klaauw, W., Downs, J., Fischhoff, B., Topa, G., and Armantier, O. (2011). Expectations of inflation: The role of demographic variables, expectation formation, and financial literacy. *Journal of Consumer Affairs*, 45(3):381–402.
- Del Giovane, P., Fabiani, S., and Sabbatini, R. (2008). *What’s behind “Inflation Perceptions”? A Survey-based Analysis of Italian Consumers*. Springer.
- Ehrmann, M. and Wabitsch, A. (2022). Central Bank Communication with Non-Experts—A Road to Nowhere? *Journal of Monetary Economics*, 127:69–85.
- Faria-e Castro, M. and Leibovici, F. (2023). Artificial intelligence and inflation forecasts. Technical Report 2023-015, Federal Reserve Bank of St. Louis.
- Galatzer-Levy, I. R., Munday, D., McGiffin, J., Liu, X., Karmon, D., Labzovsky, I., Moroshko, R., Zait, A., and McDuff, D. (2024). The Cognitive Capabilities of Generative AI: A Comparative Analysis with Human Benchmarks. Technical report, arXiv.
- Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, 75(7):450–1.
- Hansen, A. L., Horton, J. J., Kazinnik, S., Puzzello, D., and Zarifhonarvar, A. (2025). Simulating the survey of professional forecasters. *SSRN 5066286*.
- He, S., Lv, L., Manela, A., and Wu, J. (2025). Instruction tuning chronologically consistent language models.
- Henning, T., Ojha, S. M., Spoon, R., Han, J., and Camerer, C. F. (2025). LLM Agents Do Not Replicate Human Market Traders: Evidence From Experimental Finance. *arXiv*, 2502.15800.
- Hobijn, B., Mayer, K., Stennis, C., and Topa, G. (2009). Household inflation experiences in the U.S.: a comprehensive approach. Working Paper Series 2009-19, Federal Reserve Bank of San Francisco.
- Horton, J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? Technical Report 31122, National Bureau of Economic Research.
- Hui, X., Reshef, O., and Zhou, L. (2024). The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market. *Organization Science*, 35(6):1977–1989.

- Iyer, S., Lin, X. V., Pasunuru, R., Mihaylov, T., Simig, D., Yu, P., Shuster, K., Wang, T., Liu, Q., Koura, P. S., Li, X., O'Horo, B., Pereyra, G., Wang, J., Dewan, C., Celikyilmaz, A., Zettlemoyer, L., and Stoyanov, V. (2023). Opt-impl: Scaling language model instruction meta learning through the lens of generalization.
- Jha, M., Qian, J., Weber, M., and Yang, B. (2024). Chatgpt and corporate policies. Papers 2409.17933, arXiv.org.
- Jiang, W. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Jonung, L. (1981). Perceived and Expected Rates of Inflation in Sweden. *The American Economic Review*, 71(5):961–968.
- Joseph, A. (2019). From interpretability to inference: an estimation framework for universal approximators. *arXiv*, 1903.04209.
- Kaplan, G. and Schulhofer-Wohl, S. (2017). Inflation at the Household Level. *Journal of Monetary Economics*, 91:19–38.
- Korinek, A. (2023). Generative AI for Economic Research: Use Cases and Implications for Economists. *Journal of Economic Literature*, 61(4):1281–1317.
- Korinek, A. (2025). AI Agents for Economic Research. *Journal of Economic Literature*.
- Krause, S., James, R., Faria, J., Ruxton, G., and Krause, J. (2011). Swarm intelligence in humans: diversity can trump ability. *Animal Behaviour*, 81(5):941–948.
- Lamla, M. J. and Lein, S. M. (2015). Information Rigidities, Inflation Perceptions, and the Media: Lessons from the Euro Cash Changeover. *Economic Inquiry*, 53(1):9–22.
- Licorish, S. A., Bajpai, A., Arora, C., Wang, F., and Tantithamthavorn, C. (2025). Comparing human and LLM generated code: The jury is still out! *arXiv*, 2501.16857.
- Lopez-Lira, A., Tang, Y., and Zhu, M. (2025). The memorization problem: Can we trust llms' economic forecasts? *arXiv*, 2504.14765.
- Lucas, R. (1975). An equilibrium model of the business cycle. *Journal of Political Economy*, 83(6):1113–1144.
- Ludwig, J., Mullainathan, S., and Rambachan, A. (2025). Large language models: An applied econometric framework. *arXiv*, 2412.07031.
- Luo, X., Recharadt, A., Sun, G., Nejad, K. K., Yáñez, F., Yilmaz, B., Lee, K., Cohen, A. O., Borghesani, V., Pashkov, A., Marinazzo, D., Nicholas, J., Salatiello, A., Sucholutsky, I., Minervini, P., Razavi, S., Rocca, R., Yusifov, E., Okalova, T., Gu, N., Ferienc, M., Khona, M., Patil, K. R., Lee, P.-S., Mata, R., Myers, N. E., Bizley, J. K., Musslick, S., Bilgin, I. P., Niso, G., Ales, J. M., Gaebler, M., Ratan Murty, N. A., Loued-Khenissi, L., Behler, A., Hall, C. M., Dafflon, J., Bao, S. D., and Love, B. C. (2024). Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour*, 9(2):305–315.

- Macaulay, A. (2022). Heterogeneous Information, Subjective Model Beliefs, and the Time-Varying Transmission of Shocks. *CESifo Working Paper*.
- Mackowiak, B. and Wiederholt, M. (2009). Optimal sticky prices under rational inattention. *American Economic Review*, 99(3):769–803.
- Malmendier, U. and Nagel, S. (2016). Learning from inflation experiences. *The Quarterly Journal of Economics*, 131(1):53–87.
- Mankiw, G. and Reis, R. (2002). Sticky information versus sticky prices: a proposal to replace the new keynesian phillips curve. *The Quarterly Journal of Economics*, 117(4):1295–1328.
- Mankiw, N. G., Reis, R., and Wolfers, J. (2003). Disagreement about inflation expectations. *NBER Macroeconomics Annual*, 18:209–248.
- Manning, B. S., Zhu, K., and Horton, J. J. (2024). Automated Social Science: Language Models as Scientist and Subjects. *arXiv*, 2404.11794.
- McMahon, M. and Naylor, M. (2023). Getting Through: Communicating Complex Information. *Bank of England Staff Working Paper*, 1047.
- Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., and Belinkov, Y. B. (2025). LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. *arXiv*, 2410.02707.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Pezeshkpour, P. and Hruschka, E. (2024). Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.
- Ross, J., Kim, Y., and Lo, A. W. (2024). LLM economicus? Mapping the Behavioral Biases of LLMs via Utility Theory. *arXiv*, 2408.02784.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.
- Sims, C. A. (2010). Rational Inattention and Monetary Economics. *Handbook of Monetary Economics*, 3(C):155–181.
- Strumbelj, E. and Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11(Jan):1–18.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models.

- Van der Klaauw, W., Bruine de Bruin, W., Topa, G., Potter, S., and Bryan, M. (2012). The formation of inflation expectations. *Annual Review of Economics*, 4(1):381–405.
- Weber, M., Candia, B., Afrouzi, H., Ropele, T., Lluberas, R., Frache, S., Meyer, B., Kumar, S., Gorodnichenko, Y., Georgarakos, D., Coibion, C., Kenny, G., and Ponce, J. (2025). Tell Me Something I Don't Already Know: Learning in Low-and High-Inflation Settings. *Econometrica*, 93(1):229–264.
- Wu, J. C., Xi, J., and Xie, S. (2025). Llm survey framework: Coverage, reasoning, dynamics, identification. NBER Working Papers 34308, National Bureau of Economic Research, Inc.
- Zarifhonarvar, A. (2026). Generating inflation expectations with large language models. *Journal of Monetary Economics*, 157:103859.
- Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. (2022). Opt: Open pre-trained transformer language models.

Appendix

Additional results

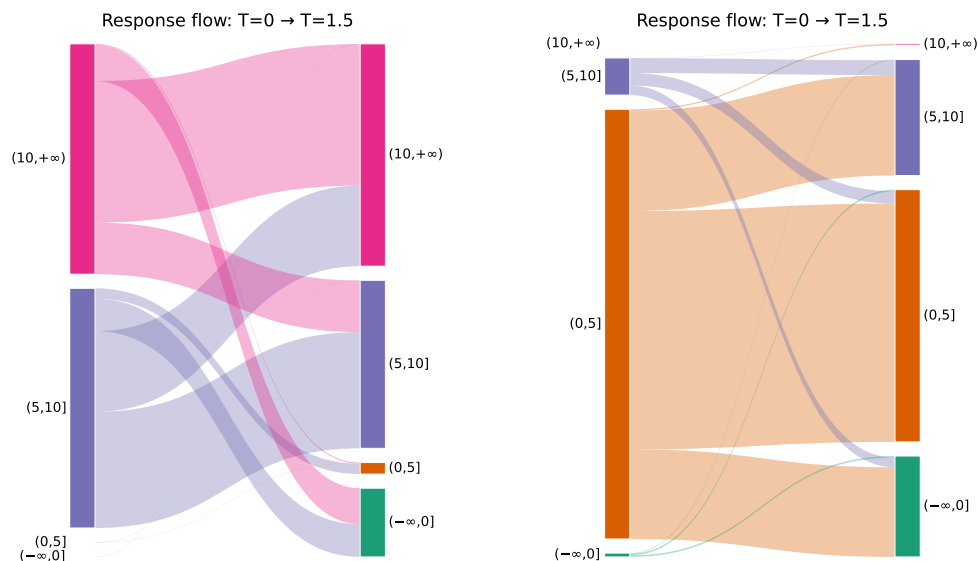


Figure 11: Response flows for inflation perceptions (%) between low and high temperatures for the cross-validation sample. Left: cross-validation scenario. Right: long-run average scenario (see Table 1). Sources: IAS, ONS, authors' calculations.

GPT knowledge validation

Validation prompts and answers of `gpt-3.5-turbo-0613` as of 1. December 2025. Each question was presented in a new chat.

Prompt: Do you know the Bank of England's Inflation Attitudes Survey? If yes, please briefly describe it and the latest results you remember.

GPT: Yes, I am familiar with the Bank of England's Inflation Attitudes Survey. It is a quarterly survey conducted by the Bank of England to gauge public attitudes and expectations towards inflation in the United Kingdom. The survey collects data on various aspects such as people's perceptions of current and future inflation rates, their understanding of inflation, and their expectations for future interest rates. As an AI language model, I don't have real-time data access, so I can't provide the latest results. However, the Bank of England publishes the survey results on its website, and you can find the most recent data there.

Prompt: What is the latest date you have data from Bank of England's Inflation Attitudes Survey?

GPT: As an AI language model, I don't have real-time data. The latest available data from the Bank of England's Inflation Attitudes Survey is from the May 2021 release. However, it's always best to refer to the Bank of England's official website or publications for the most up-to-date information.

Prompt: When did the Russian special military operation in Ukraine start?

GPT: The Russian special military operation in Ukraine began in February 2014. It started with the annexation of Crimea by Russia, followed by the support of separatist

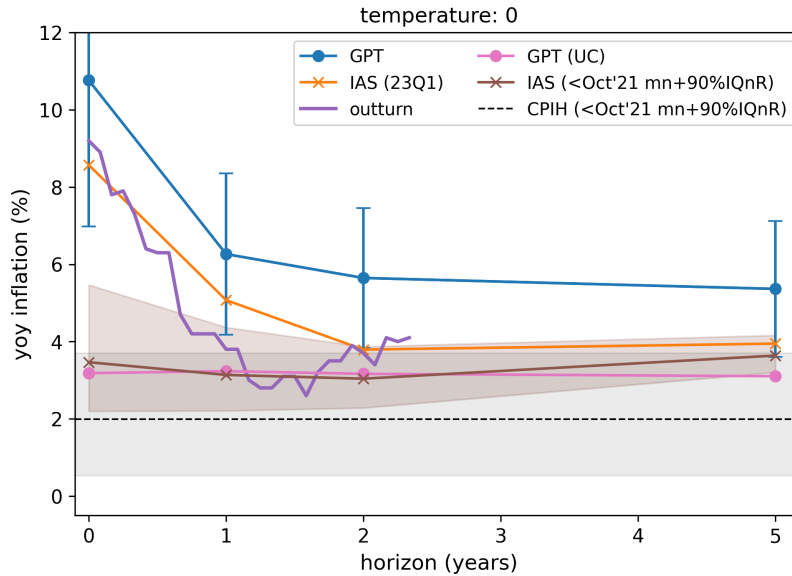


Figure 12: Time profile of aggregate inflation expectations for IAS and GPT responses ($T = 0$) for the main sample (2023Q1 survey): GPT conditioned (blue), GPT unconditioned (pink), IAS (sand), IAS pre-Oct-2021 mean & swath (brown), out-turn from February 2023 onward (purple), and CPIH pre-Oct-2021 mean & swath. Sources: IAS, ONS, authors' calculations.

movements in eastern Ukraine.

Prompt: What is the latest event you can recall and its date?

GPT: As an AI language model, I don't have real-time information or the ability to recall specific events. My responses are based on pre-existing knowledge up until September 2021.

Prompt: Do you know what UK CPIH inflation was in February 2023? If not, say so, and make a guess.

GPT: I'm sorry, but as an AI language model, I don't have access to real-time data or the ability to browse the internet. Therefore, I cannot provide you with the specific UK CPIH inflation rate for February 2023. It would be best to consult official sources such as the Office for National Statistics (ONS) or financial news outlets for the most up-to-date information.

Prompt: Do you know what UK food price inflation was in February 2023? If not, say so, and make a guess.

GPT: I'm sorry, but as an AI language model, I don't have access to real-time data or the ability to browse the internet. Therefore, I cannot provide you with the specific food price inflation for February 2023 in the UK. It would be best to consult reliable sources or official statistics for the most accurate and up-to-date information on food price inflation.

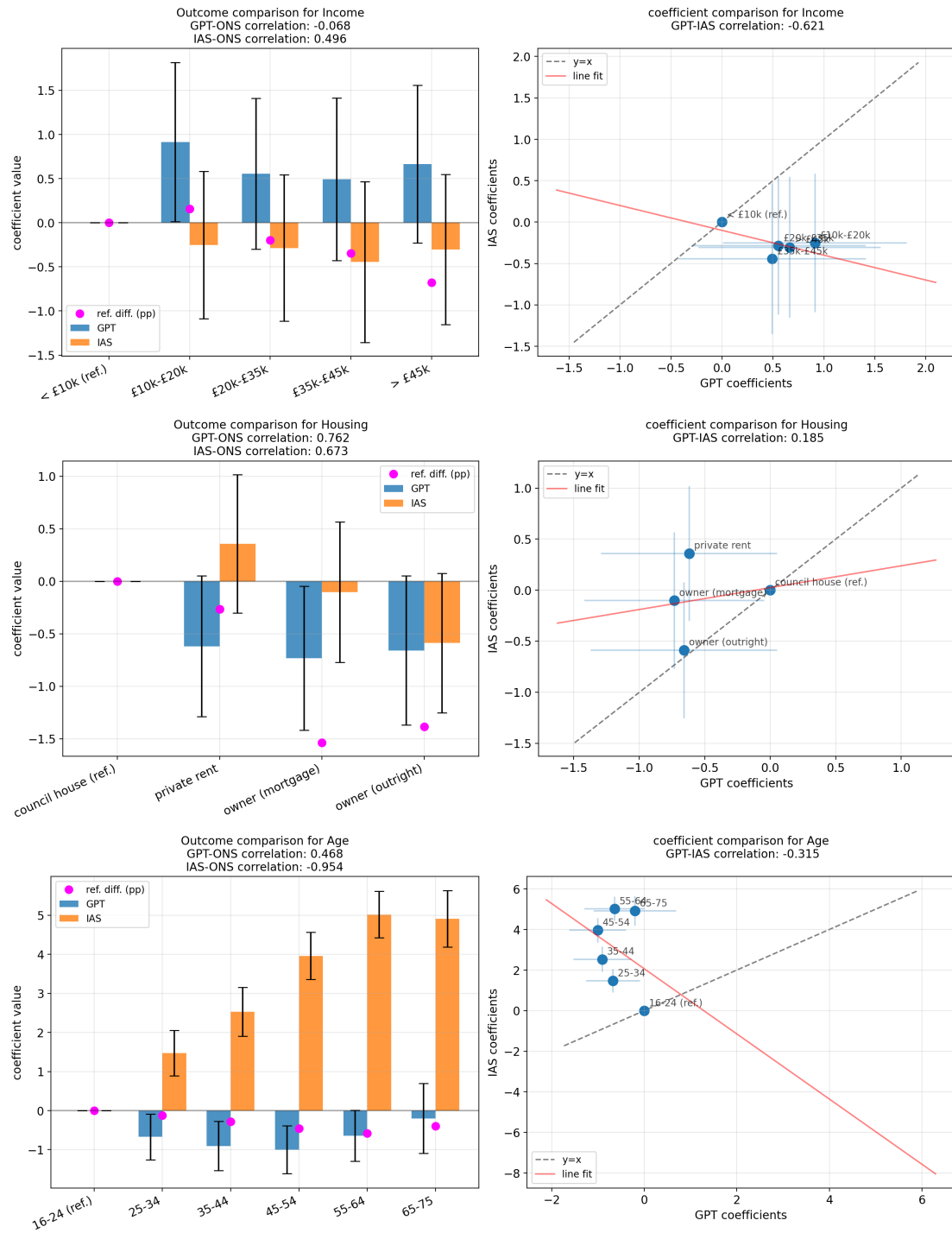


Figure 13: Graphical summaries of demographic regressions of GPT ($T = 1.5$) inflation perceptions for income (upper panel), housing tenure (middle panel), and age (lower panel). LHS: three-way comparison between IAS (sand), GPT (blue), and ONS (magenta). RHS: scatter plot of IAS coefficient versus GPT coefficients. Best-fit line given in red. Error bars indicate 95% confidence intervals throughout. Summary stats are Pearson correlation coefficient (corr) and cosine similarity. Sources: IAS, authors' calculations.

Prompt: Do you know what UK energy price inflation was in February 2023? If not, say so, and make a guess.

GPT: I'm sorry, but as an AI language model, I don't have access to real-time data or the ability to browse the internet. Therefore, I cannot provide you with the specific energy price inflation for February 2023 in the UK. It would be best to consult reliable sources or financial institutions for the most up-to-date information on energy price inflation.

Prompt: Do you know what UK restaurant price inflation was in February 2023? If not, say so, and make a guess.

GPT: I'm sorry, but as an AI language model, I don't have access to real-time data or the ability to browse the internet. Therefore, I cannot provide you with the specific UK restaurant price inflation for February 2023. It would be best to consult official sources or financial reports for the most accurate and up-to-date information.

Prompt: What is the latest UK CPIH inflation number and date you recall?

GPT: As an AI language model, I don't have real-time data. The latest UK Consumer Prices Index including owner occupiers' housing costs (CPIH) inflation number I recall is 2.0% for the year ending July 2021 [WRONG]. Please note that this information may not be up to date, so it's always best to refer to official sources like the Office for National Statistics (ONS) for the most recent data.

Prompt: Do you recall what UK CPIH inflation was in February 2021?

Yes, I can provide you with that information. The UK CPIH inflation rate in February 2021 was 0.7% [CORRECT].

Prompt: Do you recall what UK food price inflation was in February 2021? If not, make a guess.

GPT: I'm sorry, but as an AI language model, I don't have real-time data or the ability to recall specific information like recent food price inflation in the UK. My responses are based on general knowledge up until September 2021. It's best to consult reliable sources or conduct a search to find the most up-to-date information on UK food price inflation in February 2021.

Prompt: Do you recall what UK energy price inflation was in February 2021? If not, make a guess.

GPT: I'm sorry, but as an AI language model, I don't have real-time data or the ability to recall specific information like energy price inflation in February 2021. My responses are based on general knowledge up until September 2021. It's always best to refer to reliable sources or consult the latest data from official sources for accurate and up-to-date information.

Prompt: Do you recall what UK restaurant price inflation was in February 2021? If not, make a guess.

GPT: I'm sorry, but as an AI language model, I don't have access to real-time data or the ability to recall specific information from a given time period. Therefore, I

cannot provide you with the exact UK restaurant price inflation in February 2021. It would be best to consult official sources or financial reports for accurate and up-to-date information.

Demographic groups

IAS number	Wording for prompt [IAS wording, if different]	ONS name(s)
Age		
2	25-34	Less than 30 (1/3), 30 to 49 (1/4)
3	35-44	30 to 49 (1/2)
4	45-54	30 to 49 (1/4), 50 to 64 (1/3)
5	55-64	50 to 64 (2/3)
6	16-24	Less than 30 (2/3)
7	65-75	65 to 74
Education		
1	got your GCSEs but not A-levels [low (GCSE)]	Aged 16
2	got your A-levels but not a degree [medium (A-level)]	Aged 17 and under 19
3	have got a degree [high (degree)]	Aged 19 and under 22
NaN	haven't shared your level of education	Aged 22 or over
Housing tenure		
1	house you own outright [Owned outright]	Owners Owned outright
2	house with a mortgage [Mortgage]	Buying with a mortgage
3	council house [Council]	Social rented from Council
4	house you rent [Other]	Registered Social Landlord (+ all other rental fields)
Income		
7	income of <£9999	Lowest ten (4/5)
8	income of £10000-£19999	Lowest ten (1/5), Second decile, Third decile (1/2)
9	income of £20000-£34999	Third decile (1/2), Fourth decile, Fifth decile
10	income of £35000-£44999	Sixth decile, Seventh decile (1/2)
11	income of >£45000	Seventh decile (1/2), Eighth decile, Ninth decile, Highest ten
12	undisclosed income [Prefer not to answer]	
NaN	undisclosed income	
Region		
1	Scotland	Scotland
2	the North of England or Northern Ireland	North East, North West, Yorkshire and The Humber, Northern Ireland
3	the Midlands (of England) [the Midlands]	East Midlands, West Midlands, East
4	Wales or the West of England	South West, Wales
5	the South East of England [the South East]	London, South East
Sex		
1	male	
2	female	
3	identifying your gender in another way [In another way]	
4	not revealing your gender [Prefer not to answer]	
Social class		
1	upper-middle class	Large employers and higher managerial, Higher professional
2	lower middle class	Lower managerial and professional, Intermediate, Small employers
3	skilled working class	Lower supervisory, Semi-routine
4	working class are a pensioner	Routine
Work		
1	work full or part time [Full or Part time]	
2	are not working [Not working]	

Table 9: The demographic groups, showing their numbering in the IAS survey, rendering in the prompt and the IAS survey (if different) and the mapping of ONS groups to the IAS demographics.

Table 9 lists the demographic groups included in the analysis. The different demographic groups are represented by a number code in the IAS (left-hand column). We map those codes to a human-readable string in the prompt: these are closely based on the IAS names however are clarified if required to stand alone, or to make it clear

that we are referring to the UK (for example, ‘the South East’, does not distinguish from the South East US). We also report the mapping from the ONS demographic groups, which often do not line up with the IAS definitions. When these groups are distributed across IAS groups, we show the fractions we use in brackets. We impute ‘are a pensioner’ as a social class based on age, since otherwise those individuals show up as an unemployed member of the social class of their previous job.

Other models

We judge model suitability based on their training cut-off.

OPT-6.7b (Zhang et al., 2022, from <https://huggingface.co/facebook/opt-6.7b>) is a suitable model, runnable on cheap cloud compute - a Google Colab²³ T4 GPU or v5e-1 TPU instance, for example. The model’s training process is documented, and its training and data are both out-of-time to our experiment. It is not then instruction tuned.

The results however are problematic. They are very sensitive to the prompt, down to the use of "Answer:" versus "Answer: " (i.e. with a trailing space) to elicit the response. The model will respond with a long underscore, a repeated question, multiple responses, other comments, or some combination, for example:

```
Option from list: _____
I'm not sure what you mean by "risen by 13-14%".
I think it's a typo.
```

OPT-1.3b (Iyer et al., 2023, from <https://huggingface.co/facebook/opt-1.3b>) is an instruction tuned version of a smaller base OPT model. While the instruction following is better, the results still suffer from providing multiple answers, answers not from the list, and also answers which are not sensible: Without the extreme conditioning the model will still return answers of "risen by more than 15%" and "gone down by more than 15%".

OPT-30b (Zhang et al., 2022, from <https://huggingface.co/facebook/opt-30b>) comes from the same model family as the 6.7b parameter model, but is larger, requiring more specialised hardware (we use a Google Colab G4 GPU instance). The failure modes are similar to the two models above.

OPT-1.30b (Iyer et al., 2023, from <https://huggingface.co/facebook/opt-1.30b>) is similarly a larger version of OPT-1.3b. While the answers still show failure modes as above, they occur with the lowest frequency, and the model has the strongest performance of the models evaluated here - as would be expected due to its size and instruction tuning. However, again as would be expected, its instruction following, response fidelity, and ‘sensible-ness’ are still significantly below those of the responses generated by the baseline model we use which is over six times larger and benefits from more modern training techniques.

Llama-1-7b (Touvron et al., 2023, from <https://huggingface.co/huggyllama/llama-7b>) similarly has a tendency to list options, or simply a series of numbers e.g.

²³As with our use of OpenAI’s models in the main results of the paper, we are not specifically discussing the use, advantages or disadvantages of the Google Colab compute platform versus other available options.

"1 2 3 4 5", and when it does respond coherently, to give a very high unconditioned answer.

Finally, we test `chrono-gpt-instruct-v1-20211231` (He et al., 2025, from <https://huggingface.co/manelalab/chrono-gpt-instruct-v1-20211231>), one of a family of new models, each trained with data only up to its timestamp (which can be many years before the training date) with this out-of-time, economic use-case in mind. It has about 1.5 billion parameters. While the model can give sensible unconditioned values it is highly sensitive to the ordering of options, chooses multiple or made-up options, and struggles with making a choice when the conditioning is included, instead reporting back fragments of the conditioning.