Synthetic Firms: Employing LLMs as Organisational Simulacra

Preliminary draft – Please do not circulate*

Johannes Dahlke[1,2,3], Flavio Calvino[4], and Anna Kostiuchenko[1]

[1]University of Twente, BMS-ETM, Enschede, Netherlands
[2]ETH Zurich, KOF-MTEC, Zurich, Switzerland
[3]ISTARI.AI, Mannheim, Germany
[4]OECD Directorate for Science, Technology and Innovation, Paris, France

March 26, 2026

Abstract

This study examines whether large language models (LLMs), grounded in theory and calibrated against empirical data, can approximate population-level firm behaviour. The analysis focuses on innovative activity, as measured by the Community Innovation Survey (CIS). Using structured profiles of German firms, we compare uncalibrated and theory-calibrated LLM predictions enriched with web-derived information. The uncalibrated baseline largely over-predicts innovation. Theory-informed data-enriched calibration substantially improves accuracy and aligns predictions with observed cross-industry patterns, in a validation that is both temporally displaced and where theories are weighted using cross-source information. Optimised weights specific to technological regimes relevantly align with established characterisations. This articulates a firm-level framework for LLM-based social simulation, extending the approach beyond individual agents and highlighting both its analytical potential and its limitations for synthetic experimentation and counterfactual analysis on innovation and firm behaviour.

Keywords: large language models; firm innovation; simulation; WebAI; synthetic data
JEL Classification: O31; O33; C63; C53; L25

*The views expressed here are those of the authors and cannot be attributed to the OECD or its member countries.

1 Introduction

Firms are the primary locus at which technological knowledge is created, accumulated, and translated into economic activity. They are repositories of routines, capabilities, and organisational knowledge, whose heterogeneity produces persistent differences in innovation and productivity (Nelson and Winter, 1982; Winter, 1987; Dosi, 1988; Teece et al., 1997). Despite the richness of this theoretical tradition, gaps remain salient in their empirical modelling. First, firms often appear as black boxes mapping inputs to outputs. Although agent-based models relax this assumption, they often rely on narrow heuristics that struggle to capture the richness of organizational reasoning. Second, central mechanisms – expectations, narratives, strategic intent – resist quantification because they operate through language and meaning. Even along numerical dimensions, timely and representative firm-level data remain difficult to obtain. Overall, existing approaches in business economics offer limited leverage for experimental or counterfactual assessment of how policy instruments or shocks would alter firm behaviour.

Recent advances in large language models (LLMs) suggest methodological extensions that address these constraints. LLM-based agents can generate outputs from natural-language representations of environments and constraints, allowing flexible adaptation across heterogeneous contexts. In adjacent disciplines, such agents have demonstrated the capacity to simulate structured decision environments at scale, producing behavioural patterns that align closely with observed human responses (Park et al., 2023; Manning and Horton, 2025; Horton et al., 2026), and can reproduce emergent economic regularities that traditional rule-based approaches fail to generate (Li et al., 2023). However, these studies also show that LLM-based simulation requires theoretical grounding and calibration against empirical data in order to mitigate biases while simultaneously guarding against overfitting. Such calibration of general firm agents, in the innovation domain, demands firm-level indicators that are both scalable and available independently of the specific outcomes one aims to approximate. Here, a related stream of literature becomes relevant. The emergence of what Dahlke, Schmidt, et al. (2025) term the "WebAI paradigm" – the application of AI to large-scale organisational web data – has made it possible to derive general firm-level, notably innovation-relevant, indicators from firm websites at scale and in near real time. Kinne and Lenz (2021) demonstrate that supervised deep learning applied to such data can classify product innovators with substantial accuracy. These two strands – rich web-derived firm data and LLM-based behavioural simulation – have not been brought together, and the application of LLM agents to simulate organisational innovation behaviour remains unexplored.

This study addresses that gap by examining whether LLM-based firm agents can reproduce observable patterns of firm behaviour. The focus is on innovative activity, as measured in the Community Innovation Survey (CIS). Each firm is encoded as a structured natural-language profile designed to capture the information available to a firm representative responding to the CIS question. We compare three configurations: a baseline using only structural firm characteristics; an enriched variant incorporating web-derived technological indicators and patent data together with calibration grounded in six established innovation theories; and a regime-differentiated model allowing theory weights to vary across eight technological regimes (Castellacci, 2008). Theory weights are optimised against WebAI-derived innovation metrics and vali-

dated out of sample against CIS 2024 aggregate rates for Germany. Our findings indicate that theory-grounded, and empirically calibrated LLMs exhibit substantially improved predictive performance. The uncalibrated baseline exhibits systematic optimism bias, whereas calibrated variants largely reduce prediction error and achieve statistically significant cross-industry correlations.

The study makes three contributions. First, it extends LLM-based social simulation from individual to organisational agents, demonstrating both the feasibility and the boundary conditions of this extension. Second, it documents a large and systematic optimism bias in uncalibrated LLM predictions of firm innovation and identifies context and theory-grounding as effective correction mechanisms. Relevantly, established innovation theories provide the LLM with mechanistic reasoning structures that enable prediction patterns its default reasoning cannot produce, substantially improving aggregate accuracy and generalising across time and data source. Third, the regime-specific theory weights that emerge from calibration align with established innovation economics, suggesting that different theories enable distinct LLM reasoning patterns that correspond to how innovation actually operates across sectoral contexts.

The proposed method may produce implications that extend beyond survey replication. Establishing the extent to which theory-grounded context-enriched LLM agents can reproduce population-level information provides a foundation for future research aimed at analysing further potential market scenarios, policy interventions, and firm behaviour dynamics that are challenging to study using conventional empirical methods.

2 Theory and related work

2.1 The firm simulation problem

Agent-based models (ABMs) have long provided a foundation for studying innovation by modelling systems as collections of interacting, autonomous agents whose local rules generate aggregate outcomes (Macal and North, 2009). The existing literature on this modelling approach highlights its relevance for studying complex adaptive systems characterised by heterogeneity and local interaction (An et al., 2021) . In standard ABM implementations, agents are constrained through explicit behavioural rules, heuristics, and predefined interaction structures (Macy and Willer, 2001). While such a structure allows for analytical clarity and control, it also requires researchers to determine decision processes ex ante, reducing complex reasoning to simplified rule sets. This creates a challenge in representing path-dependent processes, contextual interpretation, or flexible strategic reasoning without significantly increasing model complexity.

Large language models offer a different way of specifying agents. By generating decisions through internal language-based reasoning processes, LLM-based agents draw on rich semantic representations and prompted instructions (Gao et al., 2023). This allows for context-dependent interpretation and adaptation to novel inputs without reprogramming behavioural rules. A further and more fundamental limitation of conventional modelling approaches is that central mechanisms of organisational behaviour – expectations, strategic narratives, interpretive routines – are constitutively linguistic, relying on language and meaning in ways that resist reduction to numerical parameters. In addition, technological change makes timely firm-level data increas-

ingly difficult to obtain through conventional survey methods, and even when firm behaviour can be described empirically, existing approaches offer limited leverage for counterfactual assessment of economic interventions.

2.2 Organisational web text as a firm observatory

A precondition for LLM-based firm simulation is that sufficiently rich and structured textual information about firms is available to serve as agent context. A growing body of work has established that organisational web data can meet this condition. Kinne and Lenz (2021) demonstrate that firm website text alone can train deep neural network classifiers to predict product innovator status in the German Community Innovation Survey (CIS) with high accuracy, showing that the publicly observable textual surface of firms encodes information that correlates strongly with survey-measured innovation outcomes – including sectoral and regional patterns that align with patent-based indicators. Hajikhani and Suominen (2022) extend this by connecting web-scraped firm content to hierarchical scientific topic models, demonstrating that textual content can be systematically mapped to bodies of knowledge and used for detailed sectoral analysis of innovation complexity. Bergeaud et al. (2025) and Dahlke, Schmidt, et al. (2025) have shown that WebAI methods – the application of AI models to large-scale organisational web data – allow innovation scholars to observe what organisations offer, adopt, and integrate at scale, capturing late-stage diffusion and implementation rather than upstream knowledge production alone. Patents and scientific publications provide complementary, more established text-based traces of knowledge production and diffusion (Pakes and Griliches, 1984). Crucially, all three sources are fundamentally textual, which means they can be analysed with similar NLP toolkits and directly compared to LLM-generated textual artefacts.

The significance of this literature for the present study is twofold. First, in a conceptual sense, it establishes that firm-level web text is not merely anecdotal self-description but a structured, predictively valid signal of firm innovation behaviour. The organisational web content that WebAI methods systematically harvest – firm websites, press releases, product descriptions, strategic communications – is also the kind of content that constitutes a substantial portion of the pretraining corpora of large foundation models (Rae et al., 2021). An LLM has, in a general and non-specific sense, already been exposed to the type of language that firms use to describe their activities, capabilities, and strategies at scale. This prior exposure does not substitute for firm-specific context, but it means that the model's reasoning about organisational behaviour is not constructed from scratch when a firm description is supplied.

Second, in practical sense, the same kind of textual data used to train classifiers in the WebAI tradition – web-scraped firm descriptions, structured financial characteristics or patent records – constitutes precisely the information that can be supplied to an LLM agent as organisational context. Structured WebAI-derived context activates and orients a set of generally trained representations and associations that the model has already formed.

WebAI metrics thus serve a dual function in this framework – as an empirical training and validation signal in the conventional sense, and as a structured interface between the specific characteristics of a given firm and the broad organisational knowledge already encoded in the foundation model. The specificity that individual firm context adds to the model's generally

trained reasoning abilities is what transforms a generic language model into a firm-specific simulacrum.

2.3 LLMs as simulated economic agents

Recently, an emerging stream of literature in experimental economics and computational social science has argued that LLMs can be used as both "scientists and subjects" in automated social-science workflows, where structural causal models and theory-grounded prompts define experiments and agent behaviours (Manning, Zhu, et al., 2024). The related notion of "general social agents" shows that prompts encoding behavioural theory can produce LLM agents whose decisions generalise across diverse situations and tasks, often outperforming standard heuristic behavioural models (Manning and Horton, 2025).

The idea that LLMs can function as stand-ins for human decision-makers in economic settings was introduced systematically by (Horton et al., 2026) under the framing of *homo silicus* – an AI agent that can be queried about hypothetical economic decisions much as a survey respondent or experimental subject would be. Horton et al. (2026) showed that GPT-based models produce responses to classic economic scenarios that are qualitatively coherent with known patterns of human behaviour, following from LLMs having implicitly absorbed the regularities of human social and economic decision-making during training on human-generated text at scale. Horton et al. (2026) argue that LLMs acquire social information through two distinct channels. First, by encoding findings and theories from the published social science literature. Second, by implicitly ingesting non-scholarly human-generated text that captures economic reasoning and decision-making in practice. Subsequent work has confirmed that across a large set of experimental replications, LLM agents respond with notable similarity to human subjects, including in studies that did not appear in the models' training corpora (Binz et al., 2024; Tranchero et al., 2024). In parallel, Park et al. (2023) demonstrated that generative agents supplied with biographical memory and natural language instructions can simulate coherent human-like social behaviour over extended interaction sequences, providing a computational proof of concept for the simulacra idea.

Unlike traditional rule-based simulation approaches, LLM-based agents rely on natural-language descriptions of environments, incentives, and actor characteristics to generate behaviour, allowing agent decisions to be more flexible and adaptive. Recent attention in research has focused on this flexibility in simulation settings (Korinek, 2025; Horton et al., 2026). By embedding theoretical logics and contextual information – such as industry and company type – into prompts, researchers can construct heterogeneous agents whose behaviour varies across scenarios (Manning and Horton, 2025). Such a prompt-based approach allows for iteration and systematic scenario testing, as modifying inputs and theoretical assumptions does not require rewriting the underlying model structure. Combined with increasing accessibility and low implementation costs (Bommasani et al., 2021; Korinek, 2025), these features make LLM-based simulations a promising methodological tool for exploring technological change.

The scope of proposed applications has expanded rapidly beyond individual economic choices, from simulating professional economic forecasters (Hansen et al., 2024) and voting behaviour (Cerina and Duch, 2025) to generating realistic social networks (Chang et al., 2025) and au-

tomating the scientific work from hypothesis generation to data analysis (Manning, Zhu, et al., 2024). Ludwig, Mullainathan, and Rambachan (2025) have offered an applied econometric framing for this strand of work, situating LLMs within the broader toolkit of prediction and hypothesis generation in empirical economics (Ludwig and Mullainathan, 2024), treating them as flexible, steerable models whose outputs can be evaluated against human data using standard econometric loss functions. Kozlowski and Evans (2025) and Charness et al. (2025) both identify AI-augmented simulation as potentially transformative for the social sciences, offering the equivalent of a controlled laboratory for testing hypotheses before expensive human-subject or survey-based data collection. A considerable amount of literature has also been published on the use of LLMs for synthetic data generation, suggesting they can mitigate data scarcity by producing structured datasets that resemble real-world distributions (Long et al., 2024).

2.4 The generalisation problem and the role of theory

Substantial challenges nonetheless accompany the use of LLMs as research instruments. A considerable amount of literature has discussed general and simulation-specific limitations (Anthis et al., 2025; Bommasani et al., 2021; Naveed et al., 2025). A primary concern is hallucination: models optimise next-token prediction rather than truthfulness, and may therefore generate factually incorrect or inconsistent outputs with great confidence (Ji et al., 2023; Naveed et al., 2025). The use of LLMs in simulations is further complicated by sensitivity to prompt formulation, where small changes can lead to substantially different outputs (Wei et al., 2022). A number of studies have begun to examine challenges specific to LLM-based simulation approaches, including limited diversity in agent outputs, poor generalisation to out-of-distribution scenarios, and the reproduction of structural biases (Anthis et al., 2025; Horton et al., 2026). Previous research findings also highlight that synthetic data generation approaches carry risks of domain inaccuracies and distorted output distributions (Gupta et al., 2023; Nadăș et al., 2025), which makes it necessary to simulate innovation-relevant signals and systematically compare them against empirical observations. Without careful grounding in established theory and systematic validation against empirical data, LLM-based simulations risk producing unreliable results, which necessitates addressing these limitations in a systematic manner.

The core methodological challenge is one of *generalisation*: an agent calibrated to match responses in one setting may do so by exploiting distributional features specific to that setting, and will fail when transferred. This is the classical overfitting problem, transposed into the space of natural language prompts. Because the space of possible prompt formulations is effectively unbounded, iterating through prompts until one reproduces the desired output distribution is not methodologically meaningful – it overfits the entire data-generating process rather than merely a training set (Manning and Horton, 2025).

Manning and Horton (2025) propose a solution structured around two essential components. The first is *theory-grounding*: the candidate set of prompts is restricted to formulations motivated by established social or economic theories, which encode stable causal relationships rather than surface distributional features of any particular dataset. Economic and behavioural theories are well-suited to this role because good theory rests on the choices and actions of purposive agents and is designed precisely to capture generalisable causal mechanisms (Friedman et al.,

1953). The second component is *cross-setting validation*: optimised agents are evaluated not on the data used to calibrate them, but on responses from related but structurally distinct settings. If calibrated agents achieve good predictive fit on this held-out data, it constitutes evidence that the embedded theory reflects genuinely stable behavioural regularities rather than overfitted artefacts. Manning and Horton (2025) demonstrate this framework empirically in the domain of strategic reasoning, showing that theory-grounded agents substantially outperform both game-theoretic equilibria and unoptimised baseline LLMs in predicting initial human play across a large population of novel games. Without theoretical grounding, they show, the optimisation procedure may fail even to improve in-sample fit. Horton et al. (2026) further show that constructing calibrated mixtures of theory-grounded agent types – for instance, combining efficiency-oriented, inequity-averse, and self-interested personas in optimised proportions – substantially improves out-of-sample prediction of human choices in structurally distinct games relative to uncalibrated baselines.

2.5 From individual agents to organisational simulacra

The extension of the LLM-agent framework from individual decision-makers to firms involves a conceptual shift that requires explicit justification. Firms are not individuals: their decisions emerge from the interaction of multiple actors operating under formal governance, competitive pressure, and accumulated organisational routines. The behavioural coherence that makes an individual-level *homo silicus* plausible – a stable preference structure, consistent reasoning patterns – has no straightforward organisational analogue.

The justification for the extension draws on two sources. The first is the economics tradition that treats firms as behaving entities in the sense that their observable decisions, however internally complex, generate regularities at the aggregate level systematically related to the capabilities, knowledge stocks, and competitive environments encoded in their organisational traces (see, for instance, Cyert and March, 1963; Nelson and Winter, 1982; Dosi, 1988; Teece et al., 1997). If these regularities are stable enough to be detected empirically, there is no a priori reason they should be opaque to an LLM trained extensively on organisational, managerial, and economic text describing how firms of different types respond to technological change.

The second source is the methodological precedent from the WebAI literature: the fact that Kinne and Lenz (2021) can train a classifier to predict product innovation outcomes that match CIS survey patterns from firm web text implies that those texts carry sufficient signal about firm innovation propensities to support inference. Although, in this study, we do not supply LLMs with entire web texts, we can assume that their patterns are encoded in their training data and can be evoked through providing respective WebAI metrics derived from such texts.

3 Methodology

The empirical strategy of this study rests on a deliberately simple premise: a large language model is asked to assume the role of a firm representative and answer the core product innovation question from the Community Innovation Survey. Each firm in the sample is presented to the model as a natural-language profile constructed from representative data, and the model returns

a binary yes/no response. Aggregating these responses across firms yields predicted innovation rates that can be compared directly to the official CIS statistics.

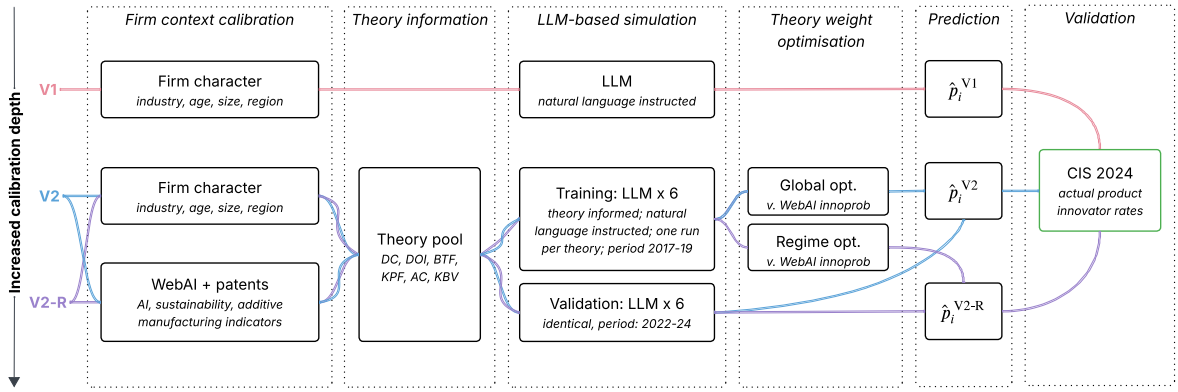


Figure 1: Overview of simulation procedure.

Note. In all three variants, a large language model assumes the role of a firm representative and answers the CIS product innovation question. The variants differ in the richness of the firm context provided to the model and in whether theory-grounded calibration is applied. V1 receives only basic structural characteristics. V2 and V2-R additionally receive web-derived innovation indicators, patent data, and theory-specific instructions; their predictions are combined using optimised weights estimated against WebAI ground truth. Out-of-sample validation is performed against CIS 2024 aggregate rates, which are never used during training or optimisation.

Figure 1 summarises the approach. We investigate three model configurations that represent increasing levels of calibration depth. The uncalibrated baseline (V1) provides the LLM with only basic firm characteristics – industry, size, age, and region – and no further guidance. This establishes the default prediction behaviour of the model. The globally calibrated variant (V2) enriches the firm profile with web-derived innovation indicators and patent data, and instructs the model to reason through established innovation theories before responding. A selection method optimises the relative weights of these theoretical framings against WebAI ground truth. The regime-specific variant (V2-R) extends this calibration by allowing theory weights to vary across technological regimes, accommodating the possibility that different sectors innovate through different mechanisms. All three variants are validated against CIS 2024 aggregate rates that were never used during training or optimisation.

3.1 Data sources

Four complementary data sources were utilised in this study, each serving a distinct role in the simulation and validation pipeline. Table 1 provides an overview.

3.1.1 Firm sample and structural data

Our baseline population is a sample of 50,000 firms that is designed to be representative of the German firm population across manufacturing and service sectors (OECD and Eurostat, 2018). Due to the computational cost of running multiple LLM simulation batches per firm, the current version of this study draws on a stratified subsample of approximately 5,400 firms (2,600 for some configurations), maintaining representation across all NACE-2 industry codes and size

Table 1: Data sources, temporal coverage, and role in the simulation pipeline

Source	Variables	Period	Role in pipeline
Web data (ISTARI.AI)	Industry, size, age, region	2016–2024	Firm identity (all variants)
WebAI indicators (ISTARI.AI)	Innovation prob., AI intensity, sustainability, additive mfg.	2016–2024	Enriched context (V2, V2-R); training ground truth
Patent records	Patent counts	2016–2024	Enriched context (V2, V2-R)
CIS 2024 (Eurostat)	Product innovation rates	2022–2024	Out-of-sample validation only

classes. Structural firm-level characteristics – four-digit NACE Rev. 2 industry classification, size class (small, medium, or large enterprise), firm age, and NUTS-3 regional location – were obtained from the ISTARI.AI database, which extracts all publicly available information from firm websites. These variables constitute the core of each simulated firm's identity and are available across all model variants, including the uncalibrated baseline.

A key methodological feature is the deliberate separation between training and validation along two dimensions. The training period (2017–2019) and the validation period (2022–2024) are temporally disjoint. The training ground truth (WebAI indicators) and the validation benchmark (CIS aggregate rates) originate from entirely independent data collection processes. This dual separation guards against overfitting and provides a stringent test of the model's generalisability.

3.1.2 Web-based innovation indicators

The calibrated model variants (V2 and V2-R) draw on firm-level indicators derived from the WebAI paradigm, which applies machine learning to large-scale organisational web data to extract innovation-relevant signals (Dahlke, Schmidt, et al., 2025; Kinne and Axenbeck, 2020). Organisational websites serve as comprehensive, self-curated representations of firm identity, and their textual content can be systematically processed to infer innovation activities at scale with high timeliness and broad coverage. Historical values of WebAI indicators are constructed from CommonCrawl web archives, enabling consistent measurement across the 2016–2020 observation window used in this study.

Four WebAI indicators were utilised. The *innovation probability* score is a continuous measure between 0 and 1 reflecting the likelihood that a firm has introduced a new or significantly improved product (Kinne and Lenz, 2021). When binarised at a threshold of 0.5, this score served as the firm-level training ground truth. As further discussed below, this indicator is only used for the optimisation of theory weights.

The *AI technology intensity* indicator captures the depth of a firm's engagement with AI, based on paragraph-level transformer classification that distinguishes substantive AI integration from superficial mentions (Dahlke, Beck, et al., 2024). The *sustainability engagement* indicator reflects the extent to which a firm's web presence documents environmental or sustainability-related practices (Schmidt et al., 2022). The *additive manufacturing* indicator captures en-

gagement with 3D printing technologies (Schwierzy et al., 2026). All three technology adoption indicators were mapped from continuous scores to categorical levels (none, very low, low, moderate, high, very high) for inclusion in the enriched firm context as natural-language descriptions.

The choice of WebAI metrics as calibration data rather than the CIS survey responses themselves is deliberate. Because the ultimate aim is to develop general-purpose firm agents capable of simulating a range of innovation-related outcomes, the calibration data must be broadly applicable rather than tied to a single survey instrument. WebAI indicators are suited to this purpose: they are derived from publicly available textual information that can be retrieved at scale, updated in near real-time via repeated web scraping, and customised to capture different dimensions of firm behaviour depending on the research question. By calibrating against web-derived metrics and validating against an entirely independent survey, the present study guards against overfitting to any single data source while establishing that the resulting agents capture generalisable patterns of firm innovation behaviour.

3.1.3 Patent records

Patent counts were derived from matched patent records where available, providing a measure of formalised inventive activity complementary to the downstream, web-based innovation signals. For each firm, the maximum patent count observed during the relevant period was used. Patent coverage is uneven across sectors (firms in services and those engaged primarily in non-patentable innovations may hold zero patents despite active innovation), which is a limitation with consequences for specific industries as discussed in the results.

3.1.4 Community Innovation Survey

Aggregate industry-level product innovation rates from the German Community Innovation Survey (CIS) were used exclusively for out-of-sample validation. The CIS reports the share of firms that introduced new or significantly improved goods or services during a reference period, disaggregated by NACE-2 industry group and firm size class. The 2024 wave, covering 2022–2024, served as the primary validation target. The CIS data were never used during training or optimisation.

3.2 Firm context construction

Each firm was represented as a structured natural-language description, designed to convey the information an informed firm representative would draw upon when responding to a survey question about innovation activity. The level of contextual detail provided to the LLM varies across model variants and constitutes a key dimension of calibration depth. Table 2 summarises the information available to each variant.

The baseline variant (V1) received only the structural variables: industry, size, age, and region. The calibrated variants (V2 and V2-R) additionally received an enriched context incorporating patent holdings and three WebAI-derived technology adoption indicators (AI technology intensity, sustainability engagement and additive manufacturing indicators), each mapped from categorical levels to natural-language descriptions. AI intensity, for example, was expressed as

Table 2: Firm context provided to each model variant

Context component	Source	V1	V2	V2-R
NACE industry code and label	Web/ISTARI	✓	✓	✓
Size classification	Web/ISTARI	✓	✓	✓
Firm age	Web/ISTARI	✓	✓	✓
NUTS-3 region	Web/ISTARI	✓	✓	✓
Patent holdings	Patents	–	✓	✓
AI technology intensity (categorical)	WebAI/ISTARI	–	✓	✓
Sustainability engagement (categorical)	WebAI/ISTARI	–	✓	✓
Additive manufacturing usage (categorical)	WebAI/ISTARI	–	✓	✓
Theory-grounded system prompt	Theory pool	–	✓	✓
Global theory weights	Optimisation	–	✓	–
Regime-specific theory weights	Optimisation	–	–	✓

"shows moderate use of AI technologies" rather than as a numeric score, ensuring that the information was presented in a form amenable to LLM reasoning. For the training phase, firm characteristics were aggregated to the most recent observation within the 2017–2019 window, with patent counts taken as the maximum across the period. For the validation phase, the latest available data point (typically 2020) was used, ensuring that the LLM's predictions for the 2022–2024 period were based on information available prior to the reference period.

3.3 Baseline simulation (V1)

The first phase established an uncalibrated baseline. All 5,417 firms were submitted to OpenAI's GPT-5.2 model via the Batch API, with each firm's basic context (industry, size, age, and region) provided alongside a standardised survey question modelled on the CIS instrument:

> *"It is the year 2024. During the years 2022 to 2024, did your enterprise introduce any new or significantly improved good or service onto the market?"*

The model was instructed to respond in the role of a firm representative, returning a binary answer ("yes" or "no") together with a confidence score. No theoretical framing was provided and no web-based innovation indicators were included in the firm context. The reasoning effort parameter was set to "medium" to balance response quality and computational cost. This baseline serves as the equivalent of the off-the-shelf LLM prediction ($\hat{P}_0$) in the framework of Manning and Horton (2025), capturing the model's prior beliefs about firm innovation when given only minimal structural identifiers.

3.4 Theory-grounded simulation and calibration

3.4.1 Theory pool

Drawing on the evolutionary and innovation economics literature, six established theories of firm innovation behaviour were selected for the theory pool. Each theory was encoded as a natural-language instruction of approximately 150 words, designed to guide the LLM's reasoning process when assessing a given firm's likelihood of innovating. The theories were deliberately formulated as reasoning directives rather than mere definitions, following the finding of Manning

and Horton (2025) that theory-grounded prompts must meaningfully shape the model's decision process to produce distinct behavioural regimes. Table 3 summarises the six theories and their core mechanisms.

Table 3: Theory pool: innovation theories used as prompt variants

Code	Theory	Core mechanism encoded in prompt
DC	Dynamic Capabilities (Teece et al., 1997)	Firms innovate by sensing opportunities, seizing them through investment, and reconfiguring resources. Larger firms have stronger routines but may suffer rigidities.
DOI	Diffusion of Innovations (Rogers, 2003)	Adoption follows predictable patterns from early adopters to laggards, shaped by firm resources, industry dynamism, and proximity to innovation hubs.
BTF	Behavioural Theory of the Firm (Cyert and March, 1963)	Boundedly rational firms innovate primarily when performance falls below aspiration levels; satisfactory performance reduces the drive to search.
KPF	Knowledge Production Function (Griliches, 1979)	Innovation is output from knowledge inputs (R&D, human capital, external knowledge flows), with sector-specific technological opportunity.
AC	Absorptive Capacity (Cohen and Levinthal, 1990)	Innovation depends on prior related knowledge determining the capacity to recognise, assimilate, and exploit new external information; creates path dependence.
KBV	Knowledge-Based View (Grant, 1996)	Knowledge is the key strategic resource; innovation emerges from creation, integration, and application of organisational knowledge.

3.4.2 Training simulations

For the training phase, a stratified subsample of 2,704 firms was drawn from the full panel, maintaining representation across all NACE-2 industry codes and size classes. Stratified sampling ensured that small industry–size cells (fewer than six firms) were retained in full, while larger cells were sampled at 50%. Six independent simulation batches were conducted – one per theory – each posing the training-period question:

> *"It is the year 2019. During the years 2017 to 2019, did your enterprise introduce any new or significantly improved good or service onto the market?"*

Each simulation provided the LLM with (i) the firm-specific enriched context, including WebAI-derived indicators and patent data, (ii) the theoretical framing as a system-level instruction, and (iii) the survey question. The model returned a binary prediction and a confidence score. A single run per firm was conducted for each theory, yielding approximately 16,200 predictions across six theory batches. All simulations were processed via the OpenAI Batch API using sequential chunked submission to respect token-limit constraints.

3.4.3 Global weight optimisation (V2)

Theory weights were optimised using the selection method proposed by Manning and Horton (2025). Let $\Theta = \{\theta_1, \ldots, \theta_6\}$ denote the set of theory-grounded prompts, and let $\hat{P}_\theta$ denote the distribution of LLM predictions under prompt θ. The objective is to find a convex mixture of prompts that minimises the distance to the observed ground-truth distribution P:

$$\min_{w} \; d\left(P, \sum_{\theta \in \Theta} w_\theta \, \hat{P}_\theta\right) \quad \text{subject to} \quad \sum_{\theta} w_\theta = 1, \quad w_\theta \geq 0, \tag{1}$$

where $d(\cdot, \cdot)$ denotes the mean absolute distance (MAD) between the actual and weighted-predicted innovation rates across aggregation groups. Predictions and ground-truth rates were aggregated into cells defined by two-digit NACE industry codes and firm size classes. Only industry codes represented by at least ten firms were retained, yielding 59 aggregation groups. The ground truth P was computed as the share of firms with WebAI innovation probability exceeding 0.5 in any year of the 2017–2019 training window.

The constrained optimisation problem was solved using sequential least-squares programming (SLSQP) as implemented in SciPy, with equal weights ($w_\theta = 1/6$) as the starting point. The resulting optimal weight vector assigns heterogeneous importance to each theory, reflecting the degree to which its induced behavioural regime aligns with empirical innovation patterns across industries and firm sizes.

3.4.4 Regime-specific weight optimisation (V2-R)

To account for heterogeneity in the relevance of innovation theories across different types of industries, a regime-specific variant of the optimisation was implemented. Industries were classified into eight technological regimes following the extended taxonomy proposed by Castellacci (2008), which augments the original manufacturing-focused classification (Pavitt, 1984) with service-sector categories. Table 4 presents the mapping of NACE-2 codes to technological regimes.

Table 4: Castellacci technological regimes and NACE-2 mapping

Regime	**NACE-2 codes**
AKP – Knowledge-Intensive Business Services	62, 63, 69, 70, 71, 72, 73, 74, 85, 86, 87, 88
AKP – Specialised Suppliers	26, 28
MPG – Science-Based	20, 21, 27
MPG – Scale-Intensive	19, 22, 23, 24, 25, 29, 30
SIS – Network Infrastructure	35, 36, 37, 58, 59, 60, 61, 64, 65, 66, 77, 84
SIS – Physical Infrastructure	41, 42, 43, 46, 49, 50, 51, 52, 53
PGS – Supplier-Dominated Goods	1, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 31, 32, 38, 39
PGS – Supplier-Dominated Services	33, 45, 47, 55, 56, 68, 78, 79, 80, 81, 82, 90–96

Note. AKP = Advanced Knowledge Providers; MPG = Mass Production Goods; SIS = Supporting Infrastructure Services; PGS = Personal Goods and Services.

For each regime r, a separate weight vector $w^{(r)}$ was optimised by restricting the objective in Equation (1) to the subset of aggregation groups belonging to that regime. This procedure yields eight regime-specific weight vectors, each reflecting the theoretical mixture that best explains innovation patterns within a given technological context. The rationale for it is that the relative importance of, say, absorptive capacity versus diffusion pressures should differ between science-based manufacturing and supplier-dominated services. Regimes with fewer than two NACE groups defaulted to equal weights.

3.5 Validation

3.5.1 Validation simulations

Six additional theory-specific simulation batches were conducted for the validation period, using the same 2,704 firms but posing the 2022–2024 question. The enriched firm contexts, theoretical system prompts, and model configuration were identical to the training phase, with only the temporal reference in the survey question changed. This design ensures that any difference between training and validation performance reflects the model's ability to generalise across time rather than artefacts of prompt construction. Approximately 16,200 theory-specific predictions were generated for the validation period.

3.5.2 Weighted prediction

For each firm i and theory θ, let $\hat{y}_{i,\theta} \in \{0, 1\}$ denote the binary prediction from the validation simulation. The weighted prediction under variant $k \in \{\text{V2}, \text{V2-R}\}$ is computed as:

$$\hat{p}_i^{(k)} = \sum_{\theta \in \Theta} w_\theta^{(k)} \cdot \hat{y}_{i,\theta}, \tag{2}$$

where $w_\theta^{(\text{V2})}$ are the globally optimised weights from Equation (1) and $w_\theta^{(\text{V2-R})}$ are the regime-specific weights corresponding to firm i's technological regime. The firm-level predicted innovation status is obtained by thresholding at 0.5: $\hat{Y}_i^{(k)} = \mathbf{1}[\hat{p}_i^{(k)} \geq 0.5]$.

3.5.3 Evaluation metrics

Predicted innovation rates were aggregated to the CIS industry groupings (25 NACE-2 aggregates) and firm size classes (SME vs. large enterprise) and compared against the CIS 2024 reported rates. Four families of metrics were employed. The *mean gap* (in percentage points) captures the average signed difference between predicted and CIS actual innovation rates across industry groups, indicating whether the model systematically over- or underpredicts. The *mean absolute distance* (MAD, in percentage points) measures overall prediction accuracy irrespective of direction. The *Pearson correlation* (r) tests whether the model correctly captures the magnitude of cross-industry variation, while the *Spearman rank correlation* (ρ) assesses whether the model correctly ranks industries by innovation intensity. Statistical significance for both correlation measures is assessed at the 1% (***), 5% (**), and 10% (*) levels.

Comparisons were conducted across three model variants – V1 (uncalibrated baseline), V2 (globally calibrated), and V2-R (regime-specific calibration) – enabling assessment of the

marginal contribution of each calibration layer. Additionally, regime-level comparisons assessed whether the Castellacci taxonomy captures meaningful heterogeneity in the relative explanatory power of different innovation theories.

3.6 Computational implementation

All simulations were conducted using OpenAI's GPT-5.2 model accessed via the Batch API, with the reasoning effort parameter set to "medium" and the `store` parameter disabled to prevent data retention. Batch files were generated in JSONL format and submitted in sequential chunks of approximately 800 requests to respect the organisation-level token enqueuing limit of 900,000 tokens. Each chunk was monitored until completion before the next was submitted.

All data processing, optimisation, and visualisation were implemented in Python using pandas, NumPy, SciPy, and Matplotlib, executed in a Google Colab environment. Results were persisted to Google Drive to ensure reproducibility across sessions.

4 Results

This section presents the results of the simulation and validation pipeline. We first report the baseline prediction bias, then the calibration outcomes including optimised theory weights at both global and regime-specific levels. We then turn to the out-of-sample validation against CIS 2024, comparing predicted and actual innovation rates across industries, firm size classes, and technological regimes. Finally, we examine the per-theory prediction structure and the correlation between predicted and actual cross-industry variation.

4.1 Baseline prediction

The uncalibrated baseline (V1), which provided the LLM with only basic firm identifiers (industry, size, age, region) and no theoretical framing, predicted a product innovation rate of 95.3% across all 5,417 firms. This constitutes a systematic overestimation of approximately 63 percentage points relative to the CIS 2024 weighted average, and the bias was pervasive: V1 predicted innovation rates above 80% in 23 of 25 industry groups. The single most striking observation to emerge from this baseline is that the off-the-shelf LLM, when asked to simulate firm responses to the CIS, exhibits a strong prior that virtually all firms innovate, which is a finding that is consistent with the optimism bias documented in LLM simulations of individual decision-making (Manning and Horton, 2025). This result underscores the necessity of calibration: without theoretical grounding and enriched contextual data, the LLM's parametric knowledge alone is insufficient to reproduce realistic innovation patterns.

4.2 Calibration: optimised theory weights

4.2.1 Global weights (V2)

The selection method applied to the training data (2017–2019) yielded a global weight vector that concentrates entirely on two of the six candidate theories. Absorptive Capacity (AC) received a weight of 0.60 and the Knowledge-Based View (KBV) received 0.40, while the remaining four

theories (Dynamic Capabilities, Diffusion of Innovations, Behavioural Theory of the Firm, and the Knowledge Production Function) were assigned zero weight. This concentration suggests that, at the aggregate level, the two knowledge-oriented theories most accurately capture cross-industry variation in innovation rates when mediated through the LLM's reasoning process. The resulting V2 variant predicted an overall innovation rate of 35.4%, which is a substantial correction from the V1 baseline.

4.2.2 Regime-specific weights (V2-R)

Regime-specific optimisation revealed substantial heterogeneity in the theoretical mixtures that best explain innovation patterns within each technological regime. Figure 2 visualises the full weight matrix. What stands out is the diversity of dominant theories across regimes, in contrast to the global solution's exclusive reliance on AC and KBV.

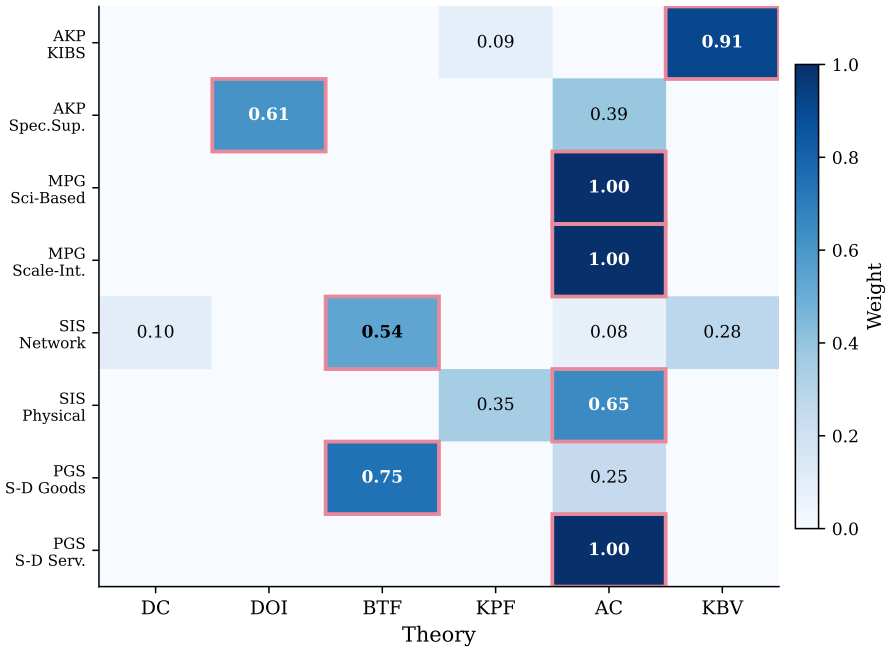


Figure 2: Regime-specific theory weights (V2-R). Each row represents a technological regime; columns represent the six candidate theories. Cell values indicate optimised weights. Red borders mark the dominant theory per regime.

Note. AKP = Advanced Knowledge Providers; MPG = Mass Production Goods; SIS = Supporting Infrastructure Services; PGS = Personal Goods and Services.

Several patterns merit attention. Absorptive Capacity emerges as the single dominant theory in both manufacturing regimes (science-based and scale-intensive) as well as in supplier-dominated services, consistent with the central role that prior knowledge accumulation plays in these sectors' innovation processes. In contrast, the Knowledge-Based View dominates in knowledge-intensive business services (KIBS), where innovation depends more on the integration of diverse professional expertise than on sequential knowledge absorption. The Behavioural Theory of the Firm receives its highest weights in network infrastructure services and supplier-dominated goods – sectors characterised by more incremental innovation patterns where sat-

isficing behaviour and performance-gap-driven search are theoretically salient. Perhaps the most interesting result concerns specialised suppliers (electronics, machinery), where the Diffusion of Innovations framework receives 61% of the weight, suggesting that innovation in these technology-producing sectors is best understood through the lens of adoption dynamics and market positioning rather than internal knowledge processes alone. Overall, the V2-R variant predicted an aggregate innovation rate of 30.4%, a further reduction from V2 and closer to the CIS population average.

4.2.3 Theory-level prediction patterns

To contextualise the weight optimisation results, Figure 3 presents each theory's predicted innovation rate across technological regimes. These raw theory-level rates represent the "ingredients" from which the weighted mixtures are assembled.

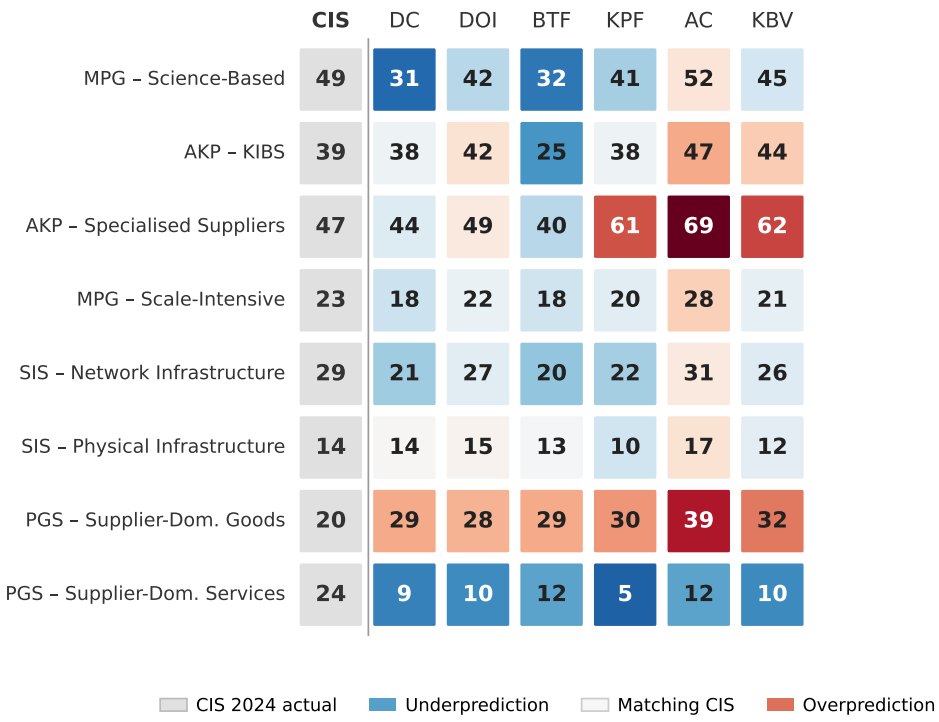


Figure 3: Per-theory predicted innovation rates (%) by technological regime (validation period, 2022–2024).

Note. AKP = Advanced Knowledge Providers; MPG = Mass Production Goods; SIS = Supporting Infrastructure Services; PGS = Personal Goods and Services.

What is interesting about the data in Figure 3 is two-fold. First, there are notable differences between prediction outcomes produced under consideration of different theories even within technological regimes. This highlights the contribution of including theory-informed instructions next to firm-level empirical context to calibrate the LLM. If different theories would have produced similar innovation rates within each technological regime, the increased prediction accuracy of the calibrated LLMs would have been primarily driven by the additional empirical context provided in V2.

Second, the consistency of the inter-theory ordering across regimes is notable. Absorptive Capacity produces the highest innovation rate in every regime without exception, while the Behavioural Theory of the Firm and Dynamic Capabilities generally yield the lowest rates. This ordering is preserved even as the absolute level of predicted innovation varies substantially across regimes – from 5–12% in supplier-dominated services to 40–69% in specialised suppliers. The consistency suggests that the theoretical framings induce stable behavioural shifts in the LLM's reasoning, rather than producing regime-contingent reversals, which lends support to the interpretability of the weight optimisation procedure as the weights are selecting among consistently differentiated "behavioural types" rather than among theories whose relative effects are unpredictable.

4.3 Out-of-sample validation against CIS 2024

4.3.1 Aggregate fit statistics

Table 5 presents the fit statistics for each model variant when predicted innovation rates are compared against CIS 2024 actual rates across 25 industry groups. All three metrics – mean gap, mean absolute distance (MAD), and correlation – show progressive improvement from V1 through V2 to V2-R.

Table 5: Fit statistics: predicted vs. CIS 2024 actual innovation rates by industry ($n = 25$)

	Mean gap	**MAD**	**Pearson**		**Spearman**	
Variant	(pp)	(pp)	r	sig.	ρ	sig.
V1 (baseline)	+62.7	62.7	0.362	$*$	0.564	$***$
V2 (global)	+5.8	13.6	0.504	$**$	0.569	$***$
V2-R (regime)	+1.0	10.5	0.578	$***$	0.586	$***$

Note. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. Mean gap is the average signed difference (predicted − actual); positive values indicate overprediction. MAD is the mean absolute distance.

The most striking result to emerge from the data is the magnitude of the calibration effect. V1 overpredicts innovation rates by an average of 62.7 percentage points and exhibits only a weak linear correlation with actual cross-industry patterns ($r = 0.36$, $p < 0.10$). The globally calibrated V2 reduces the mean gap to 5.8 pp and the MAD from 62.7 to 13.6 pp – a 78% reduction in absolute prediction error – while improving the Pearson correlation to $r = 0.50$ ($p < 0.05$). The regime-specific V2-R achieves further gains, reducing the mean gap to just 1.0 pp and the MAD to 10.5 pp, with a Pearson correlation of $r = 0.58$ ($p < 0.01$). The Spearman rank correlations are statistically significant at the 1% level for all calibrated variants, indicating that the models correctly rank industries by innovation intensity even where point estimates deviate.

Figure 4 visualises the relationship between predicted and actual innovation rates for each variant. The progressive tightening of the scatter around the 45-degree line from V1 to V2-R is clearly apparent.

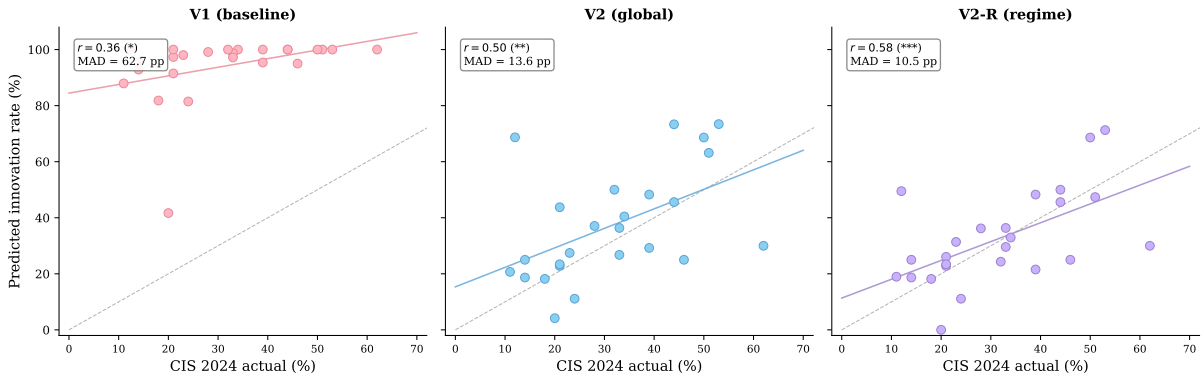


Figure 4: Predicted vs. CIS 2024 actual innovation rates by industry group. Each panel shows one model variant. The dashed line indicates perfect prediction; the solid line is the OLS fit. Annotations report the Pearson correlation and mean absolute distance.

4.3.2 Size class results

Figure 5 compares predicted and actual innovation rates by firm size class. The baseline V1 model massively overpredicts both size classes, assigning innovation rates of 91% (SME) and 99% (large firms) against CIS benchmarks of 24% and 52%, respectively. The theory-informed and empirically calibrated V2 and V2-R variants substantially correct this bias and successfully reproduce the direction and approximate magnitude of the size–innovation gradient. For large firms, V2-R predicts 43%, more closely matching the CIS figure of 52% (9 pp deviation), while SME innovation is predicted at 16% versus 24% actual – an underprediction of 8 pp. However, the theory-informed variants, whose weights were globally (and at regime-level) calibrated rather than for size-classes, both slightly undershoot for large firms and SME innovation. Notably, all three variants correctly reproduce the qualitative pattern that large firms innovate at higher rates than SMEs, though theory-informed and calibrated variants achieves the closest quantitative match overall and also in relative terms – halfing innovation rates between large firms and SME.

4.3.3 Industry-level results

Figure 6 presents the full comparison across 25 CIS industry groups, ordered by descending CIS actual innovation rate. Table 7 provides the corresponding numerical values in the Appendix.

Closer inspection of the table reveals several noteworthy patterns. In the lower half of the innovation rate distribution – industries with CIS actual rates below 25% – the calibrated models perform remarkably well. For metal production (CIS: 18%), V2-R predicts 18.2%, yielding an error of just 0.2 pp. Similarly, wood/paper/printing (CIS: 21%, V2-R: 22.8%), glass/ceramics (CIS: 21%, V2-R: 23.4%), and engineering services (CIS: 34%, V2-R: 32.9%) are all predicted within 3 pp of the actual rate.

The bottom chart in Figure 6 shows the absolute prediction error for each industry across the three model variants, illustrating how calibration reduces errors. The largest improvements from V1 to V2 occur in mid-to-low innovation industries where the baseline's near-universal prediction was most distorted: wholesale trade improves from 78.9 pp error to 4.7 pp, and metal production from 63.8 pp to 0.2 pp.

At the same time, the data reveal systematic residual biases. In the most upper and lower

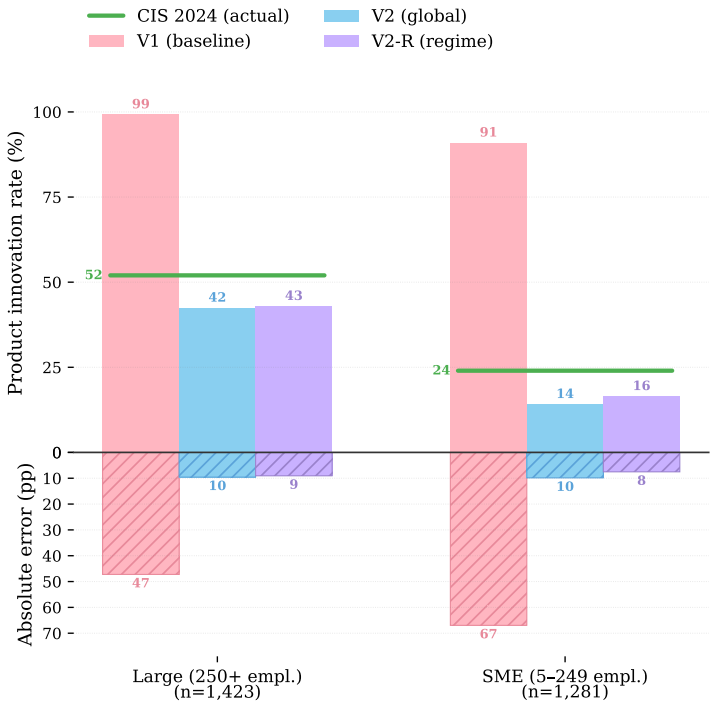


Figure 5: Predicted vs. actual product innovation rates by firm size class. Value labels indicate the predicted or actual rate in percent.

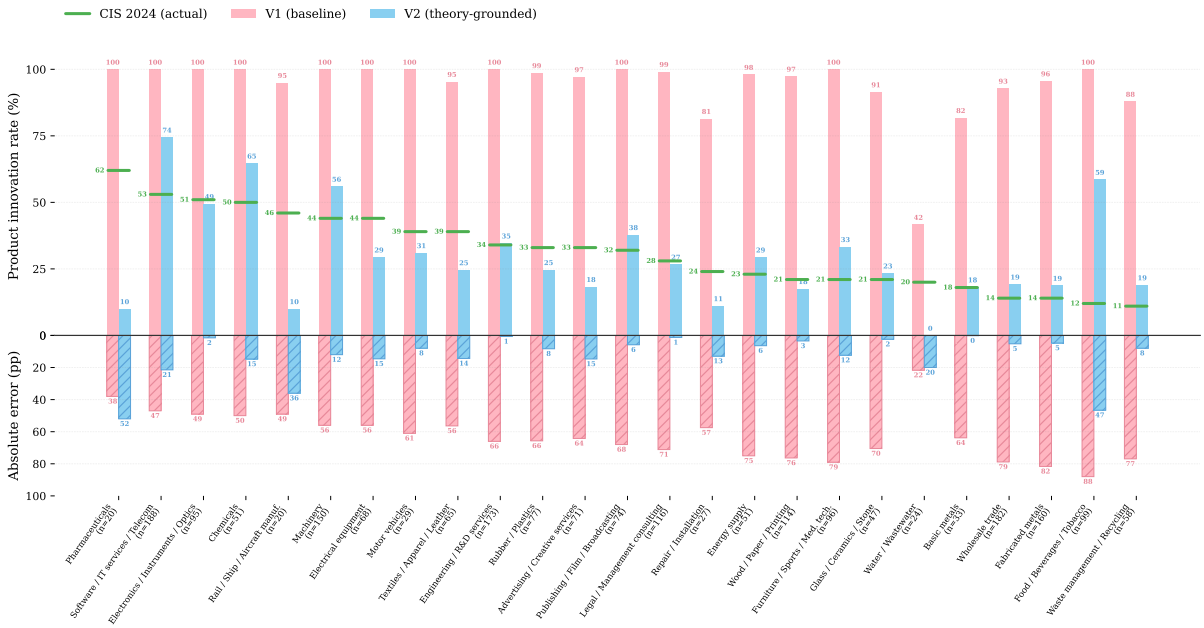


Figure 6: Predicted vs. actual product innovation rates by CIS industry group. Industries are sorted by CIS 2024 actual rate (descending). For V1, the near-uniform prediction at approximately 95% is visible across all industries.

portion of the actual CIS distribution, the calibrated model tends to overpredict: software/IT (CIS: 53%, V2: 74%), chemicals (CIS: 50%, V2: 65%), and food/beverages (CIS: 12%, V2-R: 59%) all exhibit positive residuals exceeding 15 pp. One notable exception to the general overprediction pattern is pharmaceuticals (CIS: 62%, V2: 10%), where the model substantially underpredicts. This anomaly is attributable to data limitations: the pharmaceutical firms in the reduced sample are sparse (20 firms) and uniformly lack patent records and exhibit very low WebAI technology indicators, depriving the LLM of the signals most relevant to pharmaceutical innovation (clinical development pipelines, regulatory knowledge) and leading all theories to predict non-innovation.

4.3.4 Regime-level results

Table 6 presents the corresponding fit statistics at the regime level. At the regime level, the calibration effect is again pronounced. The V1 baseline shows no statistically significant correlation with actual regime-level innovation rates ($r = 0.45$, $p = 0.27$), whereas V2 achieves a strong and significant correlation ($r = 0.77$, $p < 0.05$). V2-R reduces the MAD further from 10.4 to 7.6 pp and achieves near-zero mean bias (-0.7 pp), although the Pearson correlation is marginally lower than V2 ($r = 0.74$ vs. $r = 0.77$). This pattern suggests that regime-specific weights improve absolute calibration (lower MAD, lower bias) but do not necessarily improve the rank ordering across the eight regimes relative to the global solution.

Table 6: Fit statistics: predicted vs. CIS 2024 actual innovation rates by regime ($n = 8$)

	Mean gap	**MAD**	**Pearson**		**Spearman**	
Variant	(pp)	(pp)	r	sig.	ρ	sig.
V1 (baseline)	+63.2	63.2	0.445	ns	0.539	ns
V2 (global)	+4.8	10.4	0.771	**	0.690	*
V2-R (regime)	−0.7	7.6	0.744	**	0.643	*

Note. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$; ns = not significant.

Figure 7 presents the comparison at the level of technological regimes, and Table 8 provides the numerical values in the Appendix. As can be seen, the most accurate prediction is for science-based manufacturing (CIS: 48.8%, V2-R: 51.8%, error: 3.0 pp), followed by physical infrastructure services (CIS: 14.0%, V2-R: 17.2%, error: 3.2 pp). The largest regime-level improvement from V2 to V2-R occurs for specialised suppliers, where the regime-specific emphasis on diffusion dynamics reduces the prediction from 69.4% to 49.0%, closely approaching the CIS actual of 46.7%. The most problematic regime is network infrastructure services, where V2-R underpredicts by 21.2 pp (CIS: 41.5%, V2-R: 20.3%), likely reflecting the heterogeneity of this aggregate – which spans energy, telecommunications, finance, and public administration – and the dominance of the Behavioural Theory of the Firm in its weight vector, which tends to suppress innovation predictions.

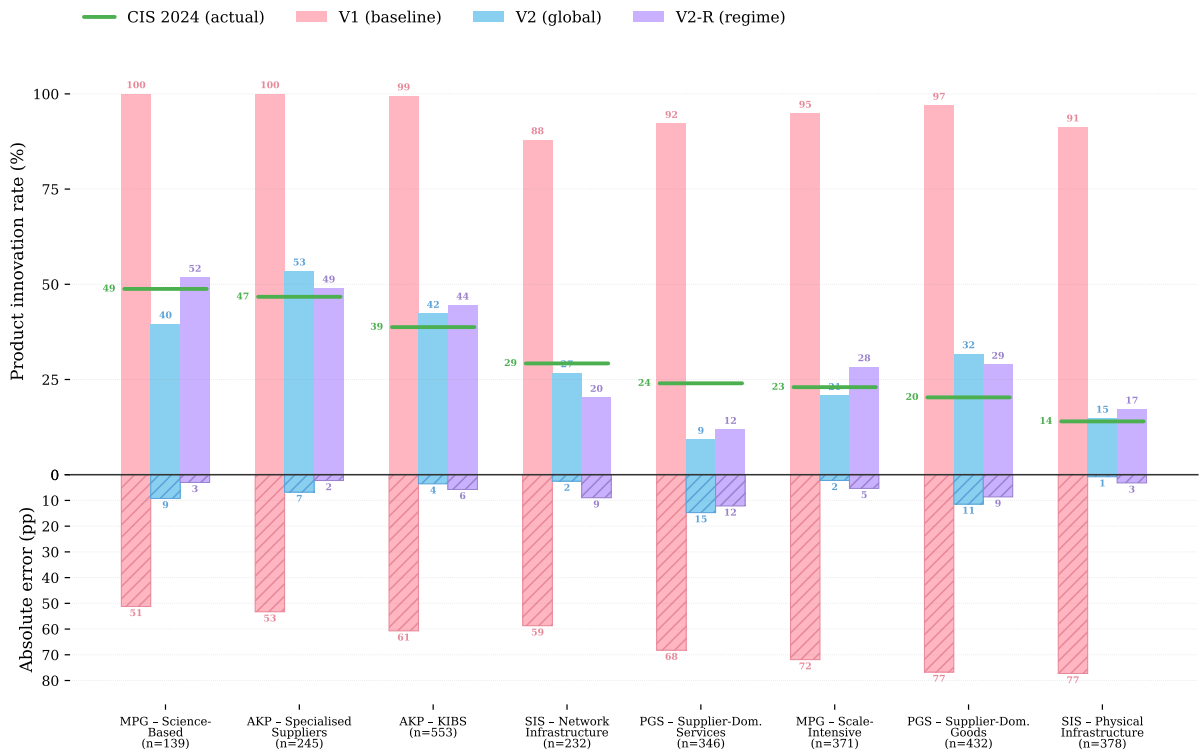


Figure 7: Predicted vs. actual product innovation rates by technological regime. Regimes are sorted by CIS 2024 actual rate (descending). Value labels indicate rates in percent. Firm counts are shown below regime labels.

Note. AKP = Advanced Knowledge Providers; MPG = Mass Production Goods; SIS = Supporting Infrastructure Services; PGS = Personal Goods and Services.

4.4 Summary of results

Taken together, these results indicate that theory-grounded calibration substantially improves the ability of LLM-based firm agents to reproduce aggregate innovation patterns. The uncalibrated baseline (V1) exhibits a pervasive optimism bias, predicting near-universal innovation. The globally calibrated variant (V2) reduces average absolute prediction error by 78% and achieves statistically significant cross-industry correlations. The regime-specific variant (V2-R) yields further improvements in absolute fit, reducing the mean industry-level bias to 1.0 pp and the MAD to 10.5 pp, with a Pearson correlation of $r = 0.58$ ($p < 0.01$). At the regime level, V2-R achieves a MAD of 7.6 pp with near-zero aggregate bias. The optimised theory weights also reveal interpretable patterns: Absorptive Capacity dominates in manufacturing regimes, the Knowledge-Based View in knowledge-intensive services, the Behavioural Theory of the Firm in low-innovation sectors, and Diffusion of Innovations in specialised supplier industries. Residual prediction errors are concentrated in industries where the firm context is impoverished relative to the true innovation drivers (pharmaceuticals) or where sample composition diverges from the CIS population (food/beverages).

5 Discussion

This study set out to examine whether theory-grounded LLM agents can approximate population-level firm behaviour, focusing on innovative activities as measured in the Community Innovation Survey. The results demonstrate that calibrated LLM-based firm simulations can reproduce aggregate innovation distributions with reasonable accuracy, reducing mean absolute prediction error by 83% relative to an uncalibrated baseline and achieving statistically significant cross-industry correlations with CIS 2024 actual rates. In this section, we discuss the substantive and methodological implications of these findings, identify boundary conditions, and address the conceptual status of LLM-based organisational agents.

5.1 Uncalibrated optimism bias

A striking finding of the analysis is the magnitude of failure without calibration. As a matter of fact, the uncalibrated baseline predicted that 95.3% of firms innovate – an overestimation of approximately 63 percentage points relative to CIS population averages. This result is consistent with the optimism bias documented in LLM simulations of individual decision-making (Manning and Horton, 2025), but its severity in the organisational context warrants separate consideration.

A possible explanation for this pronounced bias lies in the composition of LLM training corpora. Firm websites – which constitute a substantial portion of the organisational text on which foundation models are trained – are inherently self-promotional: firms emphasise novelty, capability, and forward-looking strategy while rarely documenting stasis or the absence of innovation. Dahlke, Schmidt, et al. (2025) characterise this as the "selective self-presentation" inherent in organisational web data, noting that it introduces systematic biases into analyses derived from web-based methods. Business press coverage exhibits a similar selection effect, disproportionately featuring firms that launch new products while treating non-innovation as

a non-event. The LLM has, in effect, been trained on a corpus in which firms overwhelmingly appear to innovate, and this skewed prior is faithfully reflected in its uncalibrated predictions.

A complementary explanation draws on the broader LLM simulation literature. Anthis et al. (2025) identify *sycophancy* – the tendency of instruction-tuned LLMs to generate unrealistically positive or agreeable outputs – as one of five tractable challenges for LLM-based social simulation. Because foundation models are fine-tuned to be "helpful" assistants, they may systematically overstate positive outcomes even when asked to simulate rather than advise. In the innovation context, this could manifest as a default tendency to predict that firms *do* innovate, since affirming innovation activity is the more agreeable and encouraging response. Acerbi and Stubbersfield (2023) provide further evidence that LLMs are not neutral transmitters of information but instead mirror and amplify biases present in their training data, including preferences for positive and socially desirable content. The optimism bias we document may therefore reflect a confluence of training corpus skew and instruction-tuning artefacts, rather than either alone.

This finding carries implications beyond the present study. It suggests that uncalibrated LLMs are not neutral instruments for organisational simulation – they carry the systematic biases of their training data and fine-tuning in ways that may not be immediately apparent. Researchers employing LLM-based agents to simulate firm behaviour without theoretical grounding or empirical calibration risk reproducing these biases rather than capturing genuine behavioural patterns. The optimism bias documented here may also extend to other domains of LLM-based social simulation where training data are similarly skewed, such as simulations of entrepreneurial decision-making, technology adoption, or strategic planning, although further research would be needed to confirm this.

5.2 The functional roles of theory and empirical context in LLM-based simulation

An important methodological insight that emerges from these results concerns the role of empirical context, theory and behavioural considerations when embedded in LLM-based organisational simulation. In particular, Manning and Horton (2025) frame theory-grounding as a means of restricting the candidate prompt space to formulations motivated by established mechanisms, thereby guarding against overfitting. Our findings support the concern that motivates their framework, but suggest a more differentiated account of how theory functions in the organisational setting. We identify distinct functions that theory plays in conjunction with parametrisation in order to improve simulation outcomes.

A critical function of theory in conjunction with empirical calibration in our framework, is to correct a systematically wrong prior. The uncalibrated LLM predicts that 95.3% of firms innovate. That is, not because it has overfit to any training signal, but because its parametric prior, likely shaped by training corpus composition and instruction-tuning, is biased toward affirming innovation. Theory-grounded and context-enriched prompts appear to largely correct this bias. We conjecture that this can be for two reasons. First, the empirical calibration on distinct yet related data (captured by the web-derived technology indicators encoded in the firm-specific prompts) can help establish a more realistic parametric baseline. Second, the role of theory is to provide the LLM with mechanistic reasoning structures under which non-innovation

is a legitimate and justified outcome. Without these mechanistic structures – where theory helps provide mechanisms and empirical calibration provides behavioural indications – the LLM may lack the conceptual mechanisms to justify predicting non-innovation. Here, theory appears to enable rather than constrain reasoning patterns that the model's default reasoning cannot produce. Although the current setting is not able to disentagle the role of theory grounding from the role of additional contextual information derived from WebAI indicators, the variation among different theory-level prediction patters suggests that both reasons are salient, as discussed below.

Based on our results, it is evident that all six theories shift the LLM's prediction distributions substantially downward relative to V1, reducing the mean predicted innovation rate from 95.3% to a range between approximately 30% and 75% depending on the theory and technological regime. Across theoretical prompt specifications, the predicted innovation rate is substantially lowered into a more realistic range. This suggests that the empirical calibration against web-derived information is re-anchoring the models prior assumptions. However, recalling Figure 3, we see that different theories still under- while other overpredict innovation within the same technological regime, while the same theories sometimes under- and overpredict across regimes (based on the exact same empirical context provided). Thus, we may interpret that theory directs the model's attention to different aspects of the firm's information envelope. This structured heterogeneity is what gives the optimisation procedure meaningful variation to work with. The six theory-conditioned prediction distributions represent mechanistically distinct ways of processing the same firm characteristics into innovation predictions. The optimiser selects among these distributions – or mixtures of them – to best approximate the target. The quality of the resulting calibration depends entirely on the quality of this upstream structuring. If the theories produced indistinguishable prediction distributions, the optimiser would have nothing to select among. If they produced distributions that varied along dimensions unrelated to actual innovation patterns, the optimiser might achieve good in-sample fit but poor generalisation.

This account resonates with the broader argument advanced by Horton et al. (2026) that economic theory maps naturally into LLM instructions because it specifies state variables, constraints, and decision rules that the model can execute across settings without retraining. Thus, a potential function of theory concerns precisely the LLM's capacity to emulate firm agents that generalise across contexts (Manning and Horton, 2025). If the optimiser were free to search across arbitrary prompt formulations – or if theories were selected purely for in-sample fit without regard to their mechanistic content – the resulting calibration might exploit idiosyncratic features of the training distribution (in our case, WebAI-derived technology metrics) that do not generalise to the validation target (CIS survey data). By constraining the optimiser to select among theory-grounded distributions, theory ensures that fit is achieved along dimensions that have been identified as stable across contexts by the relevant economics literature. Because these mechanisms describe general properties of organisational innovation – knowledge absorption, knowledge integration, adoption dynamics, search – there are reasons to expect that calibrations based on them will transfer across data sources and time periods.

The cross-source generalisation we observe – calibration against WebAI metrics from 2017–2019 producing predictions that correlate with CIS survey data from 2022–2024 – is consistent with this interpretation. The temporal displacement and the fundamental difference between

web-derived indicators and survey self-reports mean that the training and validation distributions differ in ways that would penalise overfitting to source-specific features. That the calibrated models nonetheless achieve a correlation of $r = 0.58$ and a mean absolute distance of 10.5 percentage points suggests that the optimiser has identified something more stable than artefacts of the WebAI data.

5.3 Interpretability and plausibility of regime-specific weights

The regime-specific weight patterns (Figure 2) provide a further source of evidence for the substantive role of theory in this framework. The dominant theories in each regime are broadly consistent with notions from the innovation economics literature, notably when considering the meso level (Dopfer et al., 2004).

Absorptive capacity dominates in both manufacturing regimes (science-based and scale-intensive) as well as in supplier-dominated services, consistent with the central role that prior knowledge accumulation and R&D investment play in these sectors' innovation processes (Cohen and Levinthal, 1990). The knowledge-based view dominates in knowledge-intensive business services, where innovation depends primarily on the integration and application of diverse professional expertise (Grant, 1996), which is a mechanism that is conceptually distinct from sequential knowledge absorption. The behavioural theory of the firm receives its highest weights in network infrastructure services and supplier-dominated goods, sectors characterised by more incremental innovation patterns where satisficing behaviour and performance-gap-driven search may be theoretically salient (Cyert and March, 1963). The prominence of diffusion of innovations in specialised suppliers (electronics, machinery) suggests that innovation in these technology-producing sectors is best understood through the lens of adoption dynamics and market positioning rather than internal knowledge processes alone.

These patterns are noteworthy because they emerge from an optimisation procedure that has no *a priori* knowledge of innovation economics. The algorithm simply minimises aggregate prediction error. The fact that it arrives at theory weights consistent with established sectoral characterisations suggests that the theoretical framings are capturing something real about how the LLM processes different types of firm contexts.

However, these results must be interpreted with caution. It remains an open question whether the regime-specific weights reflect genuine differences in the causal mechanisms driving innovation across sectors, or whether they reflect the LLM's internalised associations between sector descriptions and textbook theoretical narratives. In other words, the model may be "reasoning through" absorptive capacity when prompted to do so for a manufacturing firm not because that theory accurately captures the firm's innovation process, but because the model has learned that absorptive capacity is typically discussed in connection with manufacturing innovation.

5.4 The simulacrum question: reasoning or retrieval?

A fundamental conceptual question raised by these results concerns the nature of the LLM's decision process when simulating firm behaviour. Are these agents performing a form of organisational reasoning – drawing on theory-grounded prompts to evaluate firm characteristics and arrive at contextually appropriate innovation judgements? Or are they functioning as sophis-

ticated lookup tables, mapping firm characteristics to base rates via compressed associations from training data, with theoretical prompts merely selecting among pre-encoded distributional patterns?

The evidence does not permit a definitive answer, but several observations may favour an interpretation possibly beyond mere retrieval. First, the model is asked to predict innovation behaviour for a temporal period (2022–2024) that postdates most of its training data, using firm characteristics from an earlier period. Simple retrieval of memorised base rates would not account for the temporal displacement inherent in the task. Second, the theory-grounded prompts produce consistently differentiated behavioural regimes across all technological contexts (Figure 7), and these regimes interact meaningfully with firm-level characteristics: the same theory produces different predictions for firms with different innovation indicator profiles. If the model were merely retrieving sector-level base rates, the enriched firm context would have limited additional effect.

Third, and perhaps most tellingly, the uncalibrated baseline demonstrates that the model's "retrieval" of organisational innovation patterns is systematically wrong in the absence of theoretical guidance. If LLMs could simply look up how frequently firms innovate, V1 would approximate CIS rates without calibration. The fact that theory-grounding and empirical calibration is necessary to correct the model's prior suggests that the enriched prompts are doing more than activating pre-stored associations – they are reshaping the model's reasoning process in ways that override its default distributional assumptions. This observation suggests that LLMs can adapt the internal representations deriving from its training data rather than simply memorising it (Manning and Horton, 2025; Lindsey, 2026; Ameisen et al., 2025). If LLMs do construct genuine internal models of the domains they process, it becomes plausible that theory-grounded prompts are steering these internal models toward more empirically grounded reasoning patterns.

Li et al. (2023) provide a suggestive parallel from macroeconomic simulation, showing that LLM-based agents can reproduce emergent economic regularities – including the Phillips curve – that traditional rule-based and reinforcement-learning agents fail to generate. This suggests that LLMs may possess a capacity for generating plausible economic behaviour that extends beyond pattern matching to something resembling structured economic reasoning, at least at the aggregate level.

At the same time, the extent to which this constitutes genuine organisational simulation, e.g. in the sense intended by the evolutionary economics tradition (Nelson and Winter, 1982), remains limited. Arthur (2021) argues that firms are complex adaptive systems – network structures where value arises from connections between elements, not from the aggregation of elements alone (see also Foster, 2006). The firm, in this framing, is a "governance structure" (Williamson, 1985) whose productive process cannot be adequately captured by input–output mappings. Our LLM agents do not model internal firm processes, resource allocation conflicts, or the accumulation of organisational routines over time (although one could imagine a council of LLMs to do so). They process a static snapshot of firm characteristics through a theory-informed reasoning template and produce a single binary output.

What the framework achieves may be better characterised as *structured inference about*

organisational behaviour – inference that is informed by both empirical firm context and established theoretical mechanisms, but that does not simulate the dynamic processes from which innovation decisions emerge. The goal is not to replicate the internal decision-making of firms but to produce population-level distributions that are useful for prediction and, ultimately, for counterfactual policy analysis. For this purpose, the relevant question is not whether the model "truly understands" firm innovation, but whether its outputs are empirically valid, theoretically interpretable, and responsive to changes in the conditions that theory identifies as causally relevant. The results presented here suggest that, within the documented boundary conditions, this standard is met.

5.5 Limitations

Several limitations of the present study should be acknowledged. First, the empirical validation is restricted to a single country (Germany), a single survey instrument (CIS), and a single firm behaviour outcome (product innovation). Although the theoretical framework is general, the optimised theory weights are specific to the German innovation landscape and may not transfer to other national contexts where the relative importance of absorptive capacity, knowledge integration, or diffusion dynamics differs, or to other areas of firm behaviour. Cross-country validation, ideally using harmonised CIS data from multiple European economies or focusing on alternative indicators, would be needed to establish the portability of the approach. Anthis et al. (2025) identify generalisation as one of five key challenges for LLM social simulations more broadly, noting that models validated in one setting may fail in another due to distributional shifts in the underlying populations.

Second, the study relies on a single foundation model, and the extent to which results are robust across model families is unknown. Different foundation models, trained on different corpora with different fine-tuning procedures, may exhibit different baseline biases and respond differently to theory-grounded prompts. Replication with alternative models would strengthen confidence in the generality of the findings.

Third, the simulation design treats each firm as an independent agent, ignoring network effects, competitive interactions, and supply chain dependencies that may influence innovation decisions. Real innovation processes are often embedded in relational structures – firms are network structures where connections are limited and complex, and income–expenditure flows between systems and their components present both opportunities and limitations on what is possible (Granovetter, 1985; Powell et al., 1996). The current framework does not capture these inter-firm dynamics. Incorporating network structure, potentially through multi-agent architectures or by enriching firm contexts with relational data, represents a natural extension of this work – one that could draw on the hyperlink networks and co-mention patterns that WebAI methods are beginning to make available (Dahlke, Schmidt, et al., 2025).

Finally, the present study does not examine the sensitivity of results to prompt formulation within a given theory. The theory pool was encoded as single prompt variants per theory, and alternative formulations of the same theoretical mechanism might produce different prediction distributions. Systematic exploration of prompt sensitivity, building on emerging prompt robustness methods, would be valuable for establishing the stability of theory-grounded simulation

approaches.

5.6 Implications

The findings of this study carry implications for the measurement of firm behaviour, notably in relation to firms' innovative activities, as well as for innovation policy.

Implications for innovation research methodology. The approach developed here can offer a new instrument for studying innovation at scale that is complementary to existing methods. Questionnaire-based innovation surveys such as the CIS provide high-quality, nationally representative data but are costly, information exhibit a relevant time lag, and are subject to strict confidentiality constraints. The theory-grounded LLM simulation approach occupies a distinct methodological niche as it requires no firm-level innovation labels from the target period, drawing instead on firm characteristics observable from survey and web-based sources combined with established theoretical knowledge encoded as natural language prompts. This makes it potentially applicable in settings where survey infrastructure is absent or where rapid preliminary assessments are needed – for instance, in monitoring innovation dynamics in real time or in countries that lack established innovation survey programmes. At the same time, it should not be understood as a replacement for direct measurement. The approach may be best conceived as a "virtual survey infrastructure" for exploratory and preliminary analysis, capable of generating testable hypotheses about firm behaviour distributions that can subsequently be verified against actual survey data where available.

The the extension from individual agents (Manning and Horton, 2025) to organisational agents demonstrates that LLM-based simulation is not confined to modelling individual human behaviour. Organisations – with their more complex decision structures, multiple stakeholders, and institutional constraints – can also serve as the unit of simulation, provided the information envelope and theoretical framing are appropriately adapted. In particular, our analysis shows that LLMs, when appropriately constrained, can serve as a vehicle for augmenting meso-level analyses by identifying the dominant behavioural logics that shape innovation outcomes at the sectoral level. The regime-specific weight patterns constitute empirical evidence, albeit provisional, for the existence of distinct innovation rule regimes across the technological taxonomy.

Policy implications. With appropriate further validation, theory-grounded firm agents could enable counterfactual analysis of relevant policy scenarios, e.g. in relation to innovation policy – including in relation to for instance R&D subsidies, regulatory changes, skills policies – that are currently challenging to study experimentally. The regime-specific weight structure appears particularly promising for policy applications, as it allows the framework to capture the possibility that the same policy intervention may operate through different theoretical channels in different sectoral contexts. However, two caveats apply. First, the current validation is limited to aggregate distributional accuracy; firm-level predictive validity – essential for targeting policies at specific firm populations – has not been established. Second, counterfactual simulation requires not merely that the model predicts current innovation rates accurately, but that it responds correctly to hypothetical changes in the conditions that theory identifies as causally relevant.

Establishing this stronger form of validity is a task for future research, and would require either experimental policy variation or carefully designed quasi-experimental comparisons.

6 Conclusion

This study set out to determine whether large language models, grounded in established innovation theories and calibrated against empirical data, can approximate population-level measures of firm behaviour, notably focusing on innovation distributions as measured by the Community Innovation Survey. Uncalibrated LLM agents predict near-universal innovation among firms – a 63 percentage point overestimation that may reflect systematic biases in training corpora and instruction-tuning. Six innovation theories, encoded as natural language mechanism descriptions toghether with further contextual information, help substantially reduce this bias by enabling the model to reason toward non-innovation. The calibrated regime-specific model reduces mean absolute prediction error to 10.5 percentage points and achieves cross-industry correlations of $r = 0.58$ with CIS data, generalising across both time (2017–2019 to 2022–2024) and data source (WebAI metrics to survey self-reports). The empirical findings provide a new understanding of how established theories of innovation and firm behaviour can function within LLM-based simulation. Toghether with relevant empirical context, they appear to play a role of mechanistic scaffolds that debias the model's prior, structure its heterogeneity, and anchor its predictions along dimensions that innovation economics has identified as stable. The next steps of research will aim at testing alternative operationalisations of firms as *generative socio-economic agents*, e.g., leveraging different information sets, carrying out alternative validation exercises, while exploring the potential of further empirical applications, such as the analysis of firm-level knowledge production or competition among firms.

Use of AI assistance. We employed large language models in two distinct capacities during this research. First, OpenAI's GPT models served as the object of study, generating simulated firm-level innovation predictions via the OpenAI Batch API as described in the methodology. Second, Anthropic's Claude was used as a writing assistant during the drafting and revision of this manuscript. In the latter capacity, the AI was used to support iterative text drafting, editing, and formatting under continuous author direction. All substantive intellectual contributions – including research design, theoretical framing, data analysis, interpretation of results, and final editorial decisions – remain solely the responsibility of the authors, who reviewed and revised all AI-assisted output.

Acknowledgement of competing interest. The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Johannes Dahlke serves as Head of Science at the company ISTARI.AI in an unpaid, advisory capacity. ISTARI.AI provided data resources used in this study but no financial support. Computing costs were funded by the Strategic Research Fund of the University of Twente [2026.16]. The remaining authors declare no competing interests.

References

Acerbi, Alberto and Joseph M Stubbersfield (2023). "Large language models show human-like content biases in transmission chain experiments". In: *Proceedings of the National Academy of Sciences* 120.44, e2313790120.

Ameisen, Emmanuel et al. (2025). "Circuit tracing: Revealing computational graphs in language models". In: *Transformer Circuits Thread* 6, pp. 16318–16352.

An, Li et al. (2021). "Challenges, tasks, and opportunities in modeling agent-based complex systems". In: *Ecological Modelling* 457.C. DOI: 10.1016/j.ecolmodel.2021.109685.

Anthis, Jacy R. et al. (2025). "Position: LLM Social Simulations Are a Promising Research Method". In: *Proceedings of the 42nd International Conference on Machine Learning*. Vol. 267. Proceedings of Machine Learning Research. Vancouver, Canada.

Arthur, W Brian (2021). "Foundations of complexity economics". In: *Nature Reviews Physics* 3.2, pp. 136–145.

Bergeaud, Antonin et al. (2025). *Natural language processing and innovation research*. Tech. rep. National Bureau of Economic Research.

Binz, Marcel et al. (2024). "How should the advent of large language models affect the practice of science?" In: *Proceedings of the National Academy of Sciences* 122.5, e2401227121. DOI: 10.1073/pnas.2401227121.

Bommasani, Rishi et al. (2021). "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258*.

Castellacci, Fulvio (2008). "Technological Paradigms, Regimes and Trajectories: Manufacturing and Service Industries in a New Taxonomy of Sectoral Patterns of Innovation". In: *Research Policy* 37.6–7, pp. 978–994. DOI: 10.1016/j.respol.2008.03.011.

Cerina, Roberto and Raymond Duch (2025). "The 2024 US presidential election PoSSUM poll". In: *PS: Political Science & Politics* 58.2, pp. 286–297.

Chang, Serina et al. (2025). "LLMs generate structurally realistic social networks but overestimate political homophily". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 19, pp. 341–371.

Charness, Gary et al. (2025). "The next generation of experimental research with LLMs". In: *Nature Human Behaviour* 9.5, pp. 833–835.

Cohen, Wesley M. and Daniel A. Levinthal (1990). "Absorptive Capacity: A New Perspective on Learning and Innovation". In: *Administrative Science Quarterly* 35.1, pp. 128–152. DOI: 10.2307/2393553.

Cyert, Richard M. and James G. March (1963). *A Behavioral Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall.

Dahlke, Johannes, Mathias Beck, et al. (2024). "Epidemic effects in the diffusion of emerging digital technologies: evidence from artificial intelligence adoption". In: *Research Policy* 53.2, p. 104917.

Dahlke, Johannes, Sebastian Schmidt, et al. (2025). "The WebAI Paradigm of Innovation Research: Extracting Insight from Organizational Web Data Through AI". Working paper.

Dopfer, Kurt et al. (2004). "Micro-meso-macro". In: *Journal of evolutionary economics* 14.3, pp. 263–279.

Dosi, Giovanni (1988). *Sources, Procedures, and Microeconomic Effects of Innovation.*

Foster, John (2006). "Why is economics not a complex systems science?" In: *Journal of Economic Issues* 40.4, pp. 1069–1091. ISSN: 00213624. DOI: 10.1080/00213624.2006.11506975.

Friedman, Milton et al. (1953). "The methodology of positive economics". In: *Essays in positive economics* 3.3, pp. 145–178.

Gao, Chen et al. (2023). "Large language models empowered agent-based modeling and simulation: a survey and perspectives". In: *Humanities and Social Sciences Communications* 11. DOI: 10.1057/s41599-024-03611-3.

Granovetter, Mark (1985). "Economic action and social structure: The problem of embeddedness". In: *American journal of sociology* 91.3, pp. 481–510.

Grant, Robert M. (1996). "Toward a Knowledge-Based Theory of the Firm". In: *Strategic Management Journal* 17.S2, pp. 109–122. DOI: 10.1002/smj.4250171110.

Griliches, Zvi (1979). "Issues in Assessing the Contribution of Research and Development to Productivity Growth". In: *The Bell Journal of Economics* 10.1, pp. 92–116. DOI: 10.2307/3003321.

Gupta, Himanshu et al. (2023). "TarGEN: Targeted Data Generation with Large Language Models". In: *ArXiv* abs/2310.17876. DOI: 10.48550/arXiv.2310.17876.

Hajikhani, Arash and Arho Suominen (2022). "Mapping the sustainable development goals (SDGs) in science, technology and innovation: application of machine learning in SDG-oriented artefact detection". In: *Scientometrics* 127.11, pp. 6661–6693.

Hansen, Anne Lundgaard et al. (2024). "Simulating the survey of professional forecasters". In: *Available at SSRN 5066286*.

Horton, John J. et al. (Feb. 2026). *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* Working Paper 31122. Revised February 2026. National Bureau of Economic Research. URL: https://www.nber.org/papers/w31122.

Ji, Ziwei et al. (2023). "Towards mitigating LLM hallucination via self reflection". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843.

Kinne, Jan and Janna Axenbeck (2020). "Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study". In: *Scientometrics* 125.3, pp. 2011–2041.

Kinne, Jan and David Lenz (2021). "Predicting Innovative Firms Using Web Mining and Deep Learning". In: *PLoS ONE* 16.4, e0249071. DOI: 10.1371/journal.pone.0249071.

Korinek, Anton (2025). "AI Agents for Economic Research". In: *Journal of Economic Literature* 34202, pp. 1–56. DOI: 10.3386/w34202.

Kozlowski, Austin C and James Evans (2025). "Simulating subjects: The promise and peril of artificial intelligence stand-ins for social agents and interactions". In: *Sociological Methods & Research* 54.3, pp. 1017–1073.

Li, Nian et al. (2023). "Large language model-empowered agents for simulating macroeconomic activities". In: *Available at SSRN 4606937*.

Lindsey, Jack (2026). "Emergent introspective awareness in large language models". In: *arXiv preprint arXiv:2601.01828*.

Long, Lin et al. (2024). "On LLMs-driven synthetic data generation, curation, and evaluation: A survey". In: *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11065–11082.

Ludwig, Jens and Sendhil Mullainathan (2024). "Machine learning as a tool for hypothesis generation". In: *The Quarterly Journal of Economics* 139.2, pp. 751–827.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan (2025). *Large language models: An applied econometric framework*. Tech. rep. National Bureau of Economic Research.

Macal, Charles M. and Michael J. North (2009). "Agent-Based Modeling and Simulation". In: *Proceedings of the 2009 Winter Simulation Conference* 74.?, pp. 86–98. DOI: 10.1109/WSC.2009.5429318.

Macy, M. and R. Willer (2001). "From Factors to Actors: Computational Sociology and Agent-Based Modeling." In: *Annual Review of Sociology* 27.1, pp. 143–166. DOI: 10.1146/annurev.soc.28.110601.141117.

Manning, Benjamin S. and John J. Horton (2025). "General Social Agents". In: *arXiv preprint arXiv:2508.17407*. URL: https://arxiv.org/abs/2508.17407.

Manning, Benjamin S., Kehang Zhu, et al. (2024). "Automated Social Science: Language Models as Scientist and Subjects". In: *arXiv preprint arXiv:2404.11794*. URL: https://arxiv.org/abs/2404.11794.

Nadăș, Mihai et al. (2025). "Synthetic data generation using large language models: Advances in text and code". In: *IEEE Access*.

Naveed, Humza et al. (2025). "A comprehensive overview of large language models". In: *ACM Transactions on Intelligent Systems and Technology* 16.5, pp. 1–72.

Nelson, Richard R. and Sidney G. Winter (1982). *An Evolutionary Theory of Economic Change*. Harvard University Press.

OECD and Eurostat (2018). *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation*. 4th. Paris / Luxembourg: OECD Publishing / Eurostat. DOI: 10.1787/9789264304604-en.

Pakes, Ariel and Zvi Griliches (1984). "Patents and R&D at the Firm Level: A First Look". In: *R&D, Patents, and Productivity*. Ed. by Zvi Griliches. Chicago: University of Chicago Press, pp. 55–72.

Park, Joon Sung et al. (2023). "Generative Agents: Interactive Simulacra of Human Behavior". In: *arXiv preprint*. arXiv: 2304.03442 [cs.HC]. URL: https://arxiv.org/abs/2304.03442.

Pavitt, Keith (1984). "Sectoral Patterns of Technical Change: Towards a Taxonomy and a Theory". In: *Research Policy* 13.6, pp. 343–373. DOI: 10.1016/0048-7333(84)90018-0.

Powell, Walter W et al. (1996). "Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology". In: *Administrative science quarterly*, pp. 116–145.

Rae, Jack W et al. (2021). "Scaling language models: Methods, analysis & insights from training gopher". In: *arXiv preprint arXiv:2112.11446*.

Rogers, Everett M. (2003). *Diffusion of Innovations*. 5th. New York: Free Press.

Schmidt, Sebastian et al. (2022). "Greenwashing in the US metal industry? A novel approach combining SO2 concentrations from satellite data, a plant-level firm database and web text mining". In: *Science of the Total Environment* 835, p. 155512.

Schwierzy, Julian et al. (2026). "Mapping technology diffusion with AI: A web-based approach for tracking additive manufacturing adoption". In: *International Journal of Information Management Data Insights* 6.1, p. 100387.

Teece, David J. et al. (1997). "Dynamic Capabilities and Strategic Management". In: *Strategic Management Journal* 18.7, pp. 509–533. DOI: 10.1002/(SICI)1097-0266(199708)18:7<509::AID-SMJ882>3.0.CO;2-Z.

Tranchero, Matteo et al. (2024). *Theorizing with Large Language Models*. DOI: 10.2139/ssrn.4978831.

Wei, Jason et al. (2022). "Emergent Abilities of Large Language Models". In: *ArXiv* abs/2206.07682. DOI: 10.48550/arXiv.2206.07682.

Williamson, Oliver E. (1985). *The Economic Institutions of Capitalism*. Free Press.

Winter, Sidney G. (1987). "Knowledge and Competence as Strategic Assets". In: *The Competitive Challenge*, pp. 159–184.

A Illustrative prompt variants

The following example illustrates the two prompt variants used in the simulation study for a single firm (`weyermann.de`). Panel A shows the baseline prompt (V1) without theoretical grounding; Panel B shows the theory-grounded prompt (V2) under the Absorptive Capacity framework with an optimised weight of $w^* = 0.60$.

Panel A: V1 Baseline Prompt

System instructions:

> You are simulating the behavior of a representative of a real German firm. You will be given a brief description of the firm. Based on this description and your general knowledge of firms in the German economy, you must answer a survey question exactly as a representative of this type of firm would.
>
> Important rules:
>
> - Answer the survey question based **only** on the firm description provided and your general knowledge of similar firms in the German economy.
> - Do **not** search the web or use any external information.
> - Provide your answer in the exact JSON format specified.
> - Think carefully about what is typical for firms with these characteristics.

User input (firm context):

> **Firm Description**
> This firm operates in the Beverages industry (NACE code 1106). It is classified as a large enterprise. The firm is 141 years old and is located in the German region DE241 – Bamberg.
>
> **Survey Question**
> It is the year 2024. During the years 2022 to 2024, did your enterprise introduce any new or significantly improved good or service onto the market? (This is a question from the Community Innovation Survey.)
>
> **Response Format**
> Respond with **only** a JSON object in this exact format, nothing else:

```
{
  "answer": "yes" or "no",
  "confidence": <0-100>,
  "rationale": "<one sentence explaining your reasoning>"
}
```

Panel B: V2 Theory-Grounded Prompt
(Absorptive Capacity, $w^* = 0.60$)

System instructions:

You are simulating the behavior of a representative of a real German firm. You will be given a brief description of the firm. Based on this description, the theoretical perspective below, and your general knowledge of firms in the German economy, you must answer a survey question exactly as a representative of this type of firm would.

Theoretical Perspective
When assessing this firm's innovation behavior, reason from the perspective of the *Absorptive Capacity* framework. A firm's ability to innovate depends critically on its prior related knowledge, which determines its capacity to recognize the value of new external information, assimilate it, and apply it commercially. Firms with deeper existing knowledge bases – typically reflected in prior R&D, technical staff, and accumulated experience – are better positioned to absorb and exploit new knowledge from their environment. This creates path dependence: firms that have innovated before are more likely to innovate again, while firms without a knowledge base struggle to benefit from even abundant external opportunities. Industry context shapes the availability of external knowledge, while firm size and age proxy for accumulated internal knowledge. Consider how this firm's likely knowledge base – given its characteristics – shapes its ability to absorb external knowledge and generate innovations.

Important rules:

- Use the theoretical perspective above to guide your reasoning about this firm.
- Answer the survey question based on the firm description, the theory, and your general knowledge of similar firms in the German economy.
- Do **not** search the web or use any external information.
- Provide your answer in the exact JSON format specified.

User input (enriched firm context):

Firm Description
This firm operates in the Beverages industry (NACE code 1106). It is classified as a large enterprise. The firm is 141 years old and is located in the German region DE241 – Bamberg. The firm holds no patents. The firm shows no significant use of AI technologies. It shows no significant engagement with sustainability practices. It does not use additive manufacturing.

Survey Question
It is the year 2024. During the years 2022 to 2024, did your enterprise introduce any

new or significantly improved good or service onto the market? (This is a question from the Community Innovation Survey.)

Response Format

Respond with **only** a JSON object in this exact format, nothing else:

```
{
  "answer": "yes" or "no",
  "confidence": <0-100>
}
```

B Model validation

Table 7: Predicted vs. actual product innovation rates by CIS industry group (%) and model variant

Industry	**CIS**	**V1**	**V2**	**V2-R**	n
Pharmaceuticals	62.0	100.0	30.0	30.0	20
Software/IT/Telecom	53.0	100.0	73.4	71.3	188
Electronics/Instruments	51.0	100.0	63.2	47.4	95
Chemicals	50.0	100.0	68.6	68.6	51
Rail/Ship/Aircraft	46.0	95.0	25.0	25.0	20
Machinery	44.0	100.0	73.3	50.0	150
Electrical equipment	44.0	100.0	45.6	45.6	68
Motor vehicles	39.0	100.0	48.3	48.3	29
Textiles/Apparel/Leather	39.0	95.4	29.2	21.5	65
Engineering/R&D services	34.0	100.0	40.5	32.9	173
Rubber/Plastics	33.0	98.7	36.4	36.4	77
Advertising/Creative serv.	33.0	97.2	26.8	29.6	71
Publishing/Film/Broadcasting	32.0	100.0	50.0	24.3	74
Legal/Mgmt. consulting	28.0	99.1	37.1	36.2	116
Repair/Installation	24.0	81.5	11.1	11.1	27
Energy supply	23.0	98.0	27.5	31.4	51
Wood/Paper/Printing	21.0	97.4	22.8	22.8	114
Furniture/Sports/MedTech	21.0	100.0	43.8	26.0	96
Glass/Ceramics/Stone	21.0	91.5	23.4	23.4	47
Water/Wastewater	20.0	41.7	4.2	0.0	24
Metal production	18.0	81.8	18.2	18.2	33
Wholesale trade	14.0	92.9	18.7	18.7	182
Metal products	14.0	95.6	25.0	25.0	160
Food/Beverages/Tobacco	12.0	100.0	68.7	49.5	99
Waste mgmt./Recycling	11.0	87.9	20.7	19.0	58

Note. Industries sorted by CIS 2024 actual rate (descending). n = number of firms in the simulation sample.

Table 8: Predicted vs. actual product innovation rates by technological regime (%) and model variants

Regime	**CIS**	**V1**	**V2**	**V2-R**	n
MPG – Science-Based	48.8	100.0	51.8	51.8	139
AKP – Specialised Suppliers	46.7	100.0	69.4	49.0	245
SIS – Network Infrastr.	41.5	87.9	31.5	20.3	232
AKP – KIBS	39.1	99.5	47.4	44.5	553
PGS – Supplier-Dom. Serv.	24.0	92.2	11.8	11.8	346
MPG – Scale-Intensive	23.0	94.9	28.3	28.3	371
PGS – Supplier-Dom. Goods	20.3	97.0	38.7	28.9	432
SIS – Physical Infrastr.	14.0	91.3	17.2	17.2	378

Note. CIS actual is computed as the weighted average of constituent industry groups' CIS rates. Regimes sorted by CIS 2024 actual rate (descending). AKP = Advanced Knowledge Providers; MPG = Mass Production Goods; SIS = Supporting Infrastructure Services; PGS = Personal Goods and Services.