NUTRITION MARKETS IMPACT

APPLICATIONS OF AI IN NUTRITION RESEARCH

White Paper

October 2025

DISCLAIMER

This white paper does not endorse, recommend, or promote any specific AI provider, platform, or technology. References to AI tools or processes are illustrative and should not be interpreted as commercial or institutional endorsement.

In preparing this publication, the author(s) used Microsoft Copilot (Microsoft Corporation, 2025) to assist with language editing and improving flow. The text of this publication has been thoroughly reviewed for veracity, authority, data protection and ethical considerations.

ACKNOWLEDGEMENTS

This white paper was written by Elwina Meylentia Halim, Aurélie Reynier, and Freddie von Kaufmann, and reviewed by Mark Wijne and Patrick de Regt from the ATNi AI Governance Committee.

ABBREVIATIONS

AI	Artificial Intelligence
ATNi	Access to Nutrition initiative
GDC	UN Global Digital Compact
IT	Information Technology
KPIs	Key Performance Indicators
LLMs	Large Language Models
OCR	Optical Character Recognition
RAG	Retrieval-Augmented Generation
SMEs	Small and medium-sized enterprises

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This white paper provides a thorough overview of ATNi's testing with artificial intelligence (AI) conducted over more than one year, potentially offering practical guidance and inspiration for small- and medium-sized research organizations considering AI adoption.[1] The document is structured to take readers through ATNi's journey, from the establishment of strategic priorities and frameworks, to real-world applications of AI in nutrition research. Two core case studies underpin this narrative: the use of AI to support with corporate benchmarking and data extraction, and analysis of promotional materials. By focusing on these real-world examples, the paper provides actionable insights for other researchers and development practitioners.

Key outcomes from this work include the establishment of robust governance structures for AI integration and a strong emphasis on accountability, transparency, and organizational alignment. We highlight the importance of having a structured and measured approach to innovation, ensuring that new technologies are implemented responsibly and with strong organizational buy-in.

Regarding specific use cases, this paper identifies potential supportive roles for AI in research, and possible benefits for efficiency of data collection and analysis. It also highlights the advantages of pilot projects which, while time-consuming, yield a wealth of data to inform the more targeted application of AI.

Looking forward, this white paper lays the groundwork for ATNi's continued AI journey, with particular attention to the Nutrition Transformation Hub initiative which will provide an AI-powered platform to easily compare the nutrition performance of companies and profile the healthiness of foods across markets.

Overall, this white paper serves two functions. First it shares as a roadmap for ATNi's use of AI for its work—transforming markets through data-driven accountability, strategic partnerships, and innovative tools that advance nutrition outcomes across food systems. Secondly, this paper is intended as a knowledge-sharing resource for similar research organizations. It aims to guide them in navigating the opportunities and challenges associated with integrating AI, while fostering responsible, evidence-based innovation.

[1] Small- sized organizations are defined by the European Commission as organizations with less than 50 staff members SME definition - Internal Market, Industry, Entrepreneurship and SMEs

1 INTRODUCTION

This white paper arrives at a pivotal moment for ATNi, marking the conclusion of our initial phase of artificial intelligence experimentation, which began in late 2023. With a renewed organizational focus on thought leadership and innovation, we intend to use this white paper to reflect on and share our learnings. We aim to provide practical insights for small- and medium-sized research organizations considering the integration of AI, highlighting ways to adopt these technologies responsibly and strategically in an era defined by rapid technological change.

For a global non-profit like ATNi, fostering innovation is a delicate balancing act. We seek to encourage fresh thinking while working within a modest budget. Typically, we outsource our IT services and operate with limited internal technological expertise. This also comes amidst turbulent changes for the development sector, requiring innovative approaches to ensure sustained and optimal impact.

In this context, our efforts to integrate AI have focused on enhancing the efficiency, comprehensiveness and complexity of ATNi's research. As we look to increase our impact in transforming food systems, it is worth considering how AI can support researchers in moving beyond routine tasks, expanding their capacity to benchmark more companies, or enabling more frequent and insightful reporting, among other potential advantages.

This white paper is designed not only to document our journey but also contribute to the growing body of evidence on the role of AI in nutrition research.[1] The structure of this whitepaper reflects our approach, from placing ATNi's AI readiness and strategic direction in context, to identifying real-world applications for AI in nutrition research. We present two case studies linked to our upcoming Retail assessment 2025: the use of AI for assessing the policies and practices of grocery retailers, and the content of retailers' promotional flyers. Through these examples, we seek to share lessons learned and inspiration for other organizations considering their own journeys with AI.

2 EFFECTIVELY INTEGRATING AI IN ORGANIZATIONS

2.1 ESTABLISHING A FIT-FOR-PURPOSE GOVERNANCE STRUCTURE FOR AI

From the initial piloting stage, ATNi has taken a strategic approach to integrating AI in its work. For example, rather than using existing AI tools, we co-designed a research-oriented platform with an AI start-up, tailored to ATNi's specific research requirements.[2]

Through this process, it became clear that systematically monitoring and evaluating the implementation of AI was crucial. Furthermore, as a research organization with firm commitments to accountability, transparency and rigour, it was essential to establish strong foundations for the responsible use of AI.

Our governance journey started with the establishment of a multidisciplinary AI Governance Committee in December 2024. At a very early stage, we decided to work exclusively from publicly available data and to always embed a layer of human verification. Our strategic approach to AI is guided by four core principles:

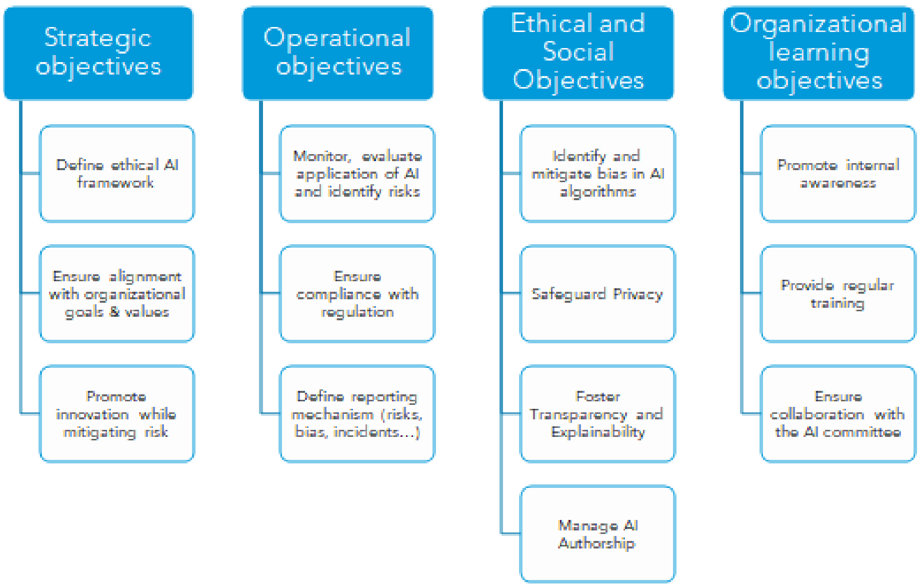

Figure 1 Objectives of the ATNi Governance Committee

- Keeping human expertise central by maintaining human oversight.
- Upholding research ethics and trust by protecting privacy and using AI consciously and responsibly.
- Fostering innovation through creative and forward-thinking solutions.
- Practicing caution by monitoring risks and complying with standards and regulations.

We defined the purpose of our committee as follows: "*to oversee and guide the ethical deployment of AI technologies within ATNi. This can include providing support and guidance to the team, reviewing AI projects for ethical concerns, advising on AI policy, and ensuring compliance with both internal standards and external regulations.*"

As we recognized that the success of the committee depended on assembling a diverse team with complementary skills, we established a group of individuals with expertise across research, operations, and data strategy. The specific objectives of the Committee are outlined in Figure 1.

Following its establishment, the Committee has taken a number of steps to build organizational alignment on the use of AI and ensure optimal adoption throughout the organization. First, the launch of the AI Governance Committee was communicated internally, accompanied by a rationale document outlining its purpose and objectives. A draft AI policy was then circulated, which staff were invited to share feedback on[3]. A staff questionnaire was conducted to capture perceptions of AI adoption in daily work, establish baseline knowledge, and determine training requirements. Training sessions were organised; a combination of in-house and external sessions, and with a focus on tools for productivity and operations, such as ChatGPT or Copilot. They also touched on more specialised applications such as Consensus, an AI-enabled academic search engine trained on millions of peer-reviewed papers; and Julius AI, a platform that enables users to upload datasets (Excel, CSV, etc.) and ask questions in natural language to produce visualizations, analyses, and insights.

2.2 ENSURING THE ETHICAL AND RESPONSIBLE USE OF AI

From an early stage, ATNi adopted a risk-based approach, following the four risk tiers set out in the *EU AI Act.*[2]

At present, ATNi's use of AI systems falls under the minimal risk tier. Our primary focus is on project-specific AI applications that enhance our ability to collect, process and analyse data (such as through information retrieval and optical character recognition (OCR). In this initial phase of AI adoption, we chose to avoid end-user interactions with outputs partially produced by Generative AI tools, such as chatbots. Any future integration of end-user AI interactions would involve careful consideration of data accuracy and reliability. This consideration is particularly pertinent in our context, as users may use ATNi's data to guide decisions regarding the performance and practices of private sector entities in the food system. More generally, transparent attribution of AI authorship is an important topic to consider, as it contributes to building accountability and trust in the information produced.

Moving forward, we are progressing towards using AI for generating and analysis of large data sets, particularly in the area of product healthiness measurement. In line with this, we are developing the concept of the Nutrition Transformation Hub, a digital platform designed to centralize data and tools to support comprehensive nutrition analysis, making it easier for organizations to access insights and compare nutrition performance across countries and companies.[5] To ensure responsible use of AI, we will use resources such as the AI Data Transparency Index[6], to inform our selection of AI models.
In addition to the sourcing of data, multiple parameters can be tested to assess the model's ethical implications. For in-depth reviews and evaluations, the *Holistic Evaluation of Language Models framework*[7] developed by the Stanford Center for Research on Foundation Models offers a comprehensive methodology. With regards to ethical considerations, the framework incorporates the dimensions of disinformation, bias, and toxicity.

[2] For definitions of specific risk levels and self-audit tools, see the EU AI Act Compliance Checker[3].

2.3 THE IMPORTANCE OF AI SOVEREIGNTY

Looking further ahead, as we aim to deepen our approach at the national level, the issue of sovereign AI will inevitably arise in the design and implementation of AI and data-driven solutions.

Sovereign AI is "*the ability for a nation to build AI with homegrown talent at different levels, based on its local policy or national AI strategy*".[4] This is an evolving area, with relevant literature and guidelines still emerging. Nonetheless we draw on two key sources for guidance. First, the *UN Global Digital Compact (GDC)*[5], adopted in September 2024, which sets out principles and commitments for the responsible governance of digital technologies, including AI. The second is the *2025 Technology and Innovation Report: Inclusive Artificial Intelligence for Development* [6] which emphasizes that strengthening AI sovereignty requires countries to invest in domestic digital infrastructure, skills, open data resources, and regulatory frameworks so they can design, deploy, and govern AI systems in line with national priorities. It also recommends enhancing regional and global cooperation, promoting open innovation, and ensuring inclusive participation in international AI governance to reduce dependency on external actors and safeguard national agency.

AI sovereignty[7] is particularly relevant when operating within the development sector and aiming to achieve meaningful, sustained change. ATNi for example is actively developing a variety of initiatives that build upon its initial food market assessments in East Africa, with a specific focus on Kenya. Should we decide to embed AI more systematically and at scale within our projects, it will be essential to take Kenya's national AI strategy into account at the project design stage.

Concretely, it is important for the project to gain a clear understanding of the local data and AI ecosystem. This will inform practical decisions regarding hosting and processing data locally, utilising available local infrastructure, and selecting locally relevant training data, including those in local languages.

Looking further ahead, considerations of AI sovereignty may also shape the types of partnerships we seek to establish, ensuring that our collaborations are aligned with local priorities and contribute to sustainable, locally driven progress.

2.4 RECOMMENDATIONS FOR NGOS LOOKING TO INTEGRATE AI IN THEIR WORK

Create a fit for purpose governance structure

From our perspective, the key is simply to take the first step. Even with a relatively light structure, an AI governance committee can play a crucial role by acting as a central point for all AI-related queries and initiatives, thereby reducing the risk of uncoordinated, ad hoc experimentation by staff.

Central to the success of the committee is establishing a clear mandate and purpose, defining a shared narrative to communicate both internally and externally, securing full endorsement at the management level, clarifying organizational values, identifying key deliverables for the committee, and monitoring staff perception and adoption.

Establishing such a committee signals that AI is a priority for the organization, not just a passing trend. We strongly recommend forming a multidisciplinary team, as this ensures that a range of viewpoints and expertise are brought to the table, which in turn fosters greater adoption, trust, and constructive debate. In practical terms, this approach also makes the oversight and management of AI issues more feasible and robust.

Ensure AI adoption aligns with organizational core values and those of the sector in which it operates

As it is often the case with technological shifts, the focus should not be solely on the technology itself, but rather on how it is integrated to drive change, transformation, and impact. To that end, it is essential to clearly reaffirm the organization's values.

For knowledge-based organizations similar to ours, credibility and reputation depend on trustworthy data. As a result, accuracy, transparency, explainability and continuous monitoring are essential when selecting and implementing AI tools.

In this context, to avoid inhibiting innovation, we recommend a 'cautious innovation' approach. While this may seem contradictory, it has in fact enabled us to embed innovation more effectively across the board while remaining in control.

Build the capacity of the team and monitor changes in the perception related to the integration of AI

The innovation lifecycle, often referred to as the Rogers' Curve, serves as a valuable framework for positioning team members at various stages of AI adoption. Identifying early adopters within the organization is essential, as these individuals can play a key role in driving internal change. By championing innovation, they have the capacity to influence colleagues who may be hesitant or resistant, and in the long run, can foster broader acceptance and engagement. In addition, developing AI capabilities within the workplace increases staff competence and enhances data literacy. These skills are directly applicable to work tasks and can also be valuable in other professional and personal contexts.

When selecting training providers, it is important to strike an appropriate balance between a strong theoretical foundation and practical, hands-on activities. In developing the training curriculum, ensure that all recommended tools and applications align with the organization's internal AI policies. This approach supports both effective learning and responsible adoption, in keeping with the broader principles outlined by the AI Governance Committee.

Monitor adoption and efficiency gain over time

To monitor progress in adoption and confidence in AI by the organization, we plan to conduct follow up questionnaires, for comparison with the initial audit. Moreover, while originally the committee only established output KPIs (for example, completion and publication of the organization's AI Policy; Delivery of planned training sessions), we recognize the need to also establish indicators to measure progress and performance. Regarding the measurement of efficiency gains, we are currently exploring options that are already integrated within the Microsoft environment, such as the Microsoft Adoption Score. This tool could, for example, help assess how effectively team members communicate and collaborate using applications such as Outlook and Co-pilot and provide analytics to track changes in productivity over time. As always, it is essential that any form of system monitoring is clearly communicated to staff prior to implementation.

Use existing tools such as the EU AI Act risk-based approach or other relevant regulations to adapt safeguards to risk level

Risk assessment and compliance management form one of the foundations of the governance Committee's work. It is important to regularly evaluate the internal risks associated with AI use, as well as those related to provider selection. Guardrails should evolve over time, recognising that increased risk levels may require changes in committee composition and adjustments to legal and technical requirements.

Risk can be assessed from multiple dimensions and from the viewpoints of various functions within the organization. Data sourcing, model selection, and legal and regulatory factors are mentioned in this white paper. Privacy and security, amongst others, should also be included when creating a risk matrix suitable for the respective organization.

Choose providers and partners with sustainability and responsibility in mind

If choosing to work with partners or providers for AI-related projects, it is important to consider which capabilities require outsourcing and which are best kept in-house. When evaluating the sustainability of a project, particular attention should be paid to the maturity of potential partners. Over the past two years, the rapid evolution and interest in AI have led to significant increase in the number of AI startups. However, this has also been met with significant turnover, with many no longer operating. This necessitates careful consideration of potential providers' longevity and resilience.

This is especially the case as large language models (LLMs) continue to advance, with any development work tied to a specific version eventually needing to be migrated to newer iterations. It is important to have early discussions with providers about how such transitions will be managed. Additionally, questions of data ownership should be prioritized, to ensure access to data and outputs is retained even after project conclusion.

Ensure strategic alignment and adequate resources to deliver on organization's ambitions

Transitioning from pilot and prototype stages to full AI integration requires strategic alignment across all areas of the organization[8]. The process begins by identifying the most suitable use-cases and piloting them, while ensuring that innovation is embedded within the broader strategy. What matters most is establishing a strong connection between programme priorities and a robust assessment of the potential changes AI systems can bring, recognising their complexity and subtle impacts. To support this, fostering a culture that encourages measured risk-taking is critical, and should become an integral part of the organization.

As illustrated by the two use cases in this paper, pilot outcomes are often nuanced and integrating AI into work processes requires careful consideration of its benefits and shortcomings. The challenge, however, is that this demands investment, time, and a certain degree of trust in the bigger picture. Having clear structures and processes in place is crucial to enabling this environment.

3 USING AI IN CORPORATE PROFILE ASSESSMENTS

A key output of ATNi's work is its Corporate Profile Assessments, which evaluate food and beverage companies on their nutrition-related policies and activities. These assessments cover a range of nutrition-related topics, including Governance, Portfolio Development, Nutrient Profiling, Affordability, Marketing, Workforce Nutrition, and Labelling.

The research approach for the Corporate Profile typically involves a combination of manual online searches for publicly available documents, and company-provided materials by those who choose to engage – often including confidential internal documents. This approach results in a high-quality evaluation, but is resource intensive, requiring significant personnel time for data collection and analysis.

ATNi's research is used by investors, governments and inter-governmental organizations, and companies themselves. While the number of companies assessed in ATNi's assessments has grown, with the Global Index increasing from 25 companies in the 2021 Global Index to 30 in the 2024 Global Index, now representing 23% of the global packaged F&B market, there is strong stakeholder interest in further expanding coverage, as well as conducting more frequent iterations of assessments.

3.1 METHODOLOGY

Motivation for AI pilot

Given the growing interest in the potential applications of AI, ATNi conducted a pilot to explore how AI could support and potentially enhance the Corporate Profile assessment process. Specifically, we sought to answer:

- Can AI identify and analyse relevant publicly available company documents?
- Does the use of AI in Corporate Profile assessments save time, and offer additional benefits to the current research approach?

Crucially, we aimed to evaluate whether AI-generated outputs could meet ATNi's research standards for specificity, accuracy, and comprehensiveness.

Approach

We partnered with an AI startup offering a research-focused retrieval-augmented generation (RAG) platform, built around Chat GPT-4. Features of the platform include:

- Web search option
- Document search function, with ability to upload and categorize files by folders and tags.
- Custom assessment templates that allow for packaging multiple pre-written prompts together, which can then be automatically generated for individual companies

For this pilot, we limited the scope to the web search function. This was for two reasons. Firstly, we wanted to see how well the AI locates relevant documents online, similar to work currently performed by analysts. Secondly, we only wanted to analyse publicly available documents, due to confidentiality concerns around how AI models store and use uploaded data.

We applied the AI pilot as an additional component to our 2025 Retail Assessment, which assesses 18 leading grocery retailers across six global markets. The assessment includes 48 indicators across various topic areas. These indicators were originally written just for use by analysts, so when compiled into a template on the AI platform, changes were made to the wording to ensure they could be understood as prompts by an AI model. Here we added additional information on the rationale for the assessment and guidance on information for the AI to locate.

Based on previous experience with using AI and wider literature on the topic, prompts often require fine-tuning to produce an optimal response. We therefore did an initial pre-pilot, testing the template for one market, evaluating its performance against the analyst's pre-assessment and iteratively adjusting the template until the responses stopped improving.

To compare AI performance with human analysts, we used the following process:

- AI assessments were generated for each retailer using the template, with analysts then comparing these against their previous assessment.
- Analysts scored the overall quality of each AI response in an Excel spreadsheet.
- Analysts also documented specific characteristics of AI responses.
- Analyst responses were compiled and analysed for overarching trends.

Evaluation Criteria

We used two evaluation frameworks:

- Scoring (0-3) based on overall quality of AI responses.
    - **Score 0**: Answer is completely incorrect or unreadable
    - **Score 1**: Answer provides some relevance to the question and answers part of the question correctly
    - **Score 2**: Answer mostly answers the question, but is missing or hallucinating a critical aspect
    - **Score 3**: Answer correctly and comprehensively answers the question, not missing any major aspect

- Multiple-choice reason codes to identify specific strengths or weaknesses.
    - **A:** Used correct sources and reached same conclusion
    - **B:** Identified additional sources (not found by analyst)
    - **C:** Used some correct sources but missed one or more key ones
    - **D:** Used correct sources, but missed/wrongly interpreted key info within them
    - **E:** Primarily used old sources
    - **F:** Missed key sources, using less relevant or third-party sources instead
    - **G:** Made up / hallucinated information
    - **H:** AI should have just said 'no information' but provided unnecessary additional information.

The 0-3 overall scoring framework was based on existing frameworks designed for the purpose of quantitively evaluating the performance of AI models. Meanwhile, the reason codes framework was an additional element we developed ourselves, to get a more detailed picture of AI's strengths and weaknesses. This was developed iteratively during the pre-pilot template testing, informed by observed issues. Regarding the reason codes framework, analysts were allowed to select multiple codes if relevant, for example, if both options B and C applied.

5 analysts participated in the assessment, each assessing 3 retailers, and one analyst assessing 6 retailers. For comparability, analysts assessed the AI responses for the same retailers that they had previously been responsible for assessing. As scoring decisions can be subjective, the participation of five analysts was useful for getting an overall picture of how useful the AI is for a research team, rather than solely from the perspective of one person.

3.2 RESULTS

Overall Score

In total, 766 AI-generated responses were evaluated across the 18 selected companies, with an average score of 2.38 out of 3. However, an important caveat to note is that for a significant proportion of indicators, analysts did not find any relevant publicly available information, and consequently in such cases a correct AI response of "No information" was sufficient to earn a top score. This therefore offered an easier route for the AI to attain a top score, than for questions where the analyst found more information, and therefore where the AI would also need to provide correct documentation and analysis.

When excluding all indicators where the analyst had originally selected 'no information,' the average score of AI responses drops significantly, to 1.6. This highlights that AI is generally finding some relevant information but struggles with providing a fully comprehensive assessment.

Additionally, 5.7% of responses scored 0, highlighting considerable variability in the quality of AI outputs, in some cases providing no information of relevance.

An important limitation to note is that analysts' assessment of AI was benchmarked against their preliminary assessments, conducted before the peer-review stage. This therefore allows for a scenario where an analyst may have misinterpreted or missed key documents, and therefore falsely evaluated the AI as having produced a 'correct' response based on issues with their initial analysis.

Key Features of AI Response

Although just over half of AI responses matched the analyst's own assessment, the high proportion of responses with missing or incomplete information means that AI cannot yet be relied upon as the primary or sole method for data collection.

Hallucinations were relatively low, at 1.3%. However, even this relatively low rate highlights the need for human oversight to carefully check AI generated outputs for accuracy and to prevent misleading content.

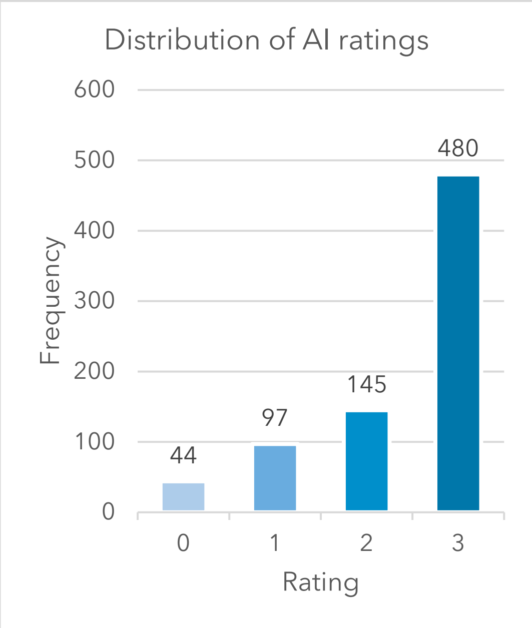

Figure 2. Distribution of AI ratings

In several cases, AI also provided extraneous or off-topic information, such as content unrelated to the company being assessed or outside the scope of the research question. While less critical than cases of incorrect or missing information, this tendency still uses up analyst time to review. We find this to be a common issue, whereby current AI models are too 'eager to please', tending toward over-inclusion, rather than limiting its answer only to information that is directly relevant.

Additionally, a common issue was that documents referenced by AI were outdated, or from less reputable sources. Even with specific prompt instructions to prioritize documents published in the last three years, and on the company's website the AI was often unable to follow these criteria. This may be due to a number of factors, including the AI not locating relevant metadata, not having access to recent documents outside of its training window, not correctly identifying the date on a report or webpage, or interpreting older documents as more relevant.

It was however interesting to note that AI often found additional relevant sources not identified previously by analysts. This was the case for approximately 7% of responses. In most cases these documents could not be found by conventional web searching, indicating that algorithms used by the LLM for determining webpage relevance differ from algorithms used by search engines to order search results. This indicates that the combination of AI and human searches could potentially result in a more comprehensive assessment of all publicly available information.

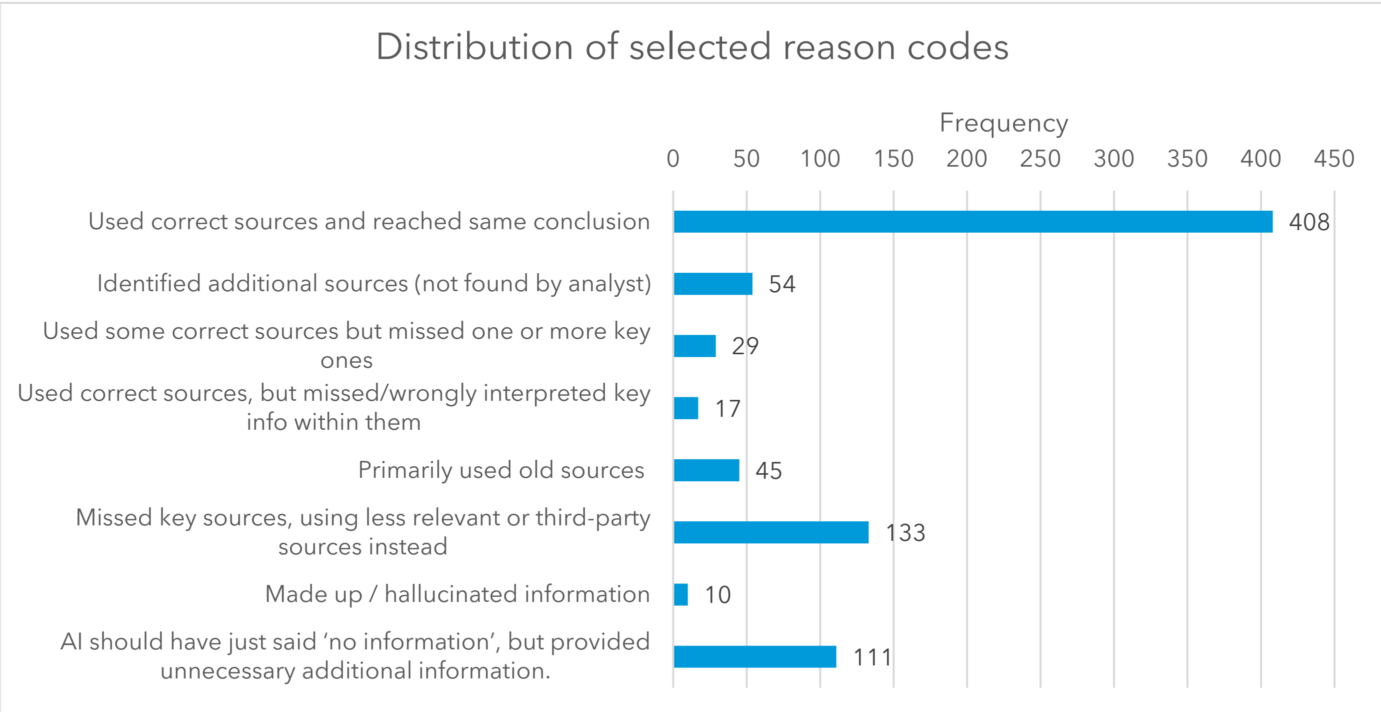

Figure 3. Distribution of selected reason codes

3.3 DISCUSSION AND FUTURE USE-CASES

For researchers wanting to conduct their own AI pilot

Frameworks are helpful for assessing AI's performance: Regarding the scoring frameworks we used for the pilot, analysts found the use of a scoring framework of 0-3 easy to use, providing more nuance than a binary 'Yes/No', while being more straightforward than larger scales such as 0-5 or 0-10. We also recommend researchers use a secondary framework such as our reason codes framework, which was effective in giving us a more nuanced understanding of specific strengths and weaknesses of the AI model used. Analysts felt that the reason codes framework was generally easy to use and covered all key areas.

It is important to test multiple models: For our pilot, the only AI model tested was Chat GPT-4. As there are a range of models currently available, it is possible that other models may perform better in some or all areas, and that their performance may improve with future iterations. We recommend that researchers test and compare different models' outputs, to identify which performs best for their needs.

Additionally, given the random nature of generative AI tools, whereby the same prompt may produce different outputs, researchers may wish to also monitor the reproducibility of outputs from AI models.

Conducting AI pilots is very time-consuming: Having analysts manually review LLM responses is inherently time consuming and difficult to scale effectively. For future AI pilots, researchers may want to explore the possibility of using 'autoraters', i.e. the use of trained AI models to assess the performance of LLM outputs.

Consider trade-offs between realism and control: As mentioned earlier, a limitation with the design of our pilot was that analysts benchmarked AI assessments against their own preliminary analysis, conducted prior to peer review. These initial assessments may have been incomplete or imperfect, introducing the risk of falsely validating AI outputs. This highlights a broader challenge in designing AI research pilots, between evaluating AI in actual research workflows, where human errors may be present, to more controlled studies where there is greater confidence in the accuracy and comprehensiveness of the baseline data.

For researchers wanting to use AI in their work

AI can't replace researchers, but it can assist them: Based on our pilot, AI shows a mixed picture with regards to performing web-based document retrieval and analysis tasks. In our experience it produces a mostly useful response, finding some relevant documents, but generally not providing sufficiently comprehensive information, or sufficient nuance in its interpretation. This is in line with existing guidelines for the responsible use of AI, such as by the European Commission, which emphasize that researchers must remain accountable for all research outputs, and that AI systems are neither authors nor co-authors.[9] We consider AI as showing promising potential as a complementary addition for analysts' work. Rather than replacing analysts, we see opportunities for AI to be embedded into the research workflow as a second-round of analysis, enabling analysts to identify any potential gaps, missing documents, or misinterpretations.

Furthermore, in addition to a more conventional format of a second-round assessment, which the analyst then compares to their own, we also see potential for other, more interactive forms of AI assistance in research. This includes:

- Research platforms with embedded chatbots that allow for more natural communication and clarification
- AI tools that are able to directly review and provide feedback on analyst's written assessments, giving suggestions for missing information.

This was a limitation of our pilot, that analysts only engaged with AI in a 'static' way, being given a set of generated AI responses, rather than being able to ask the AI follow-up questions. While this was intentional to contain the scope of the pilot, more interactive forms of AI-engagement may better suit researcher needs.

AI is not necessarily a time-saver, however it may offer other benefits: While AI has significant advantages, such as generating near-instant document-retrieval and analysis, this is significantly off-set by the need for analysts to manually check its work, and the fact that in our view it is more suited as a second pair-of-eyes, rather than for conducting first-round assessments which are more time intensive. Current models may not therefore be able to deliver on expectations of increasing the scalability of research (e.g. assessing more companies in the same timeframe), however it may offer other benefits, such as increasing the comprehensiveness of research.

AI may be less suited to research that includes confidential information: While the Retail project was well suited for our pilot as the assessment does not make use of company confidential information, AI may be less suited for other projects which do include confidential information (which is used to inform benchmarking and scoring, but not for published outputs), given data sensitivity concerns regarding current LLMs.

AI may support corporate benchmarking efforts through self-assessments: ATNi is exploring the potential to develop an AI-powered self-assessment tool for companies to evaluate their own policies and practices against an ATNi developed methodology. This could help organizations identify areas for improvement and prepare for formal ATNi assessments, while also promoting transparency and accountability. To enable this, we will continue to evaluate the performance of AI models, until we feel they meet suitable quality standards for use by third-parties. [3]

[3] Given the well-recognized issues regarding industry self-regulation, any such assessment platform would have guardrails to ensure it is used by companies for self-improvement rather than self-promotion. For example, the platform could provide feedback to companies on areas for improvement, without giving a score or ranking that companies could use for publicity.

4 AI-POWERED DATA EXTRACTION FROM SUPERMARKET FLYERS

Flyers, promotional catalogues, whether distributed in-store, online, or through letter boxes, remain a key marketing tool in the food retail sector despite the rise of digital channels. [10-12] Retailers use them to promote a selection of (discounted) products for a limited time, often employing strategies such as dollar-off, percentage-off, or bundling offers.[13] Research further shows that consumers are more likely to choose a brand simply because it is advertised in a flyer, even in the absence of discounts.[14,15] This underscores that flyers affect purchasing behaviour not only through economic incentives but also by shaping product perceptions.

Building on this understanding of flyers' influence on consumer behaviour, ATNi analysed the prevalence of healthy and unhealthy food promotions in weekly retailer flyers across countries with varying income levels. To date, existing studies have relied exclusively on manual data collection, with reviewers coding products in flyers by hand. Therefore, to increase the efficiency and scalability, we introduced a novel hybrid approach that incorporates Artificial Intelligence (AI) into the data collection process. This case study discusses the implementation of AI for data extraction from flyers, focusing on validity.

4.1 METHODOLOGY

Scope of flyers

Flyers were collected online from major retailers across six countries over a continuous three-week period between June and July 2025. To avoid promotional bias, flyers covering the week leading up to public holidays in each country were excluded from the analysis. The three-week minimum was selected based on findings from a previous study, which showed that after analysing three weeks of flyers, patterns in layout, design, and content reached saturation, with no notable changes observed beyond that point. [13]

All flyers were downloaded in PDF format from the retailers' official websites. When website access required a location, a major city where all selected retailers operate was selected. If a retailer did not publish PDF flyers, promotional content available on its website was used. However, since this webpage content was extracted using a non-AI method, it falls outside the scope of this paper. For this white paper case study, the analysis includes data from nine retailers, each providing a minimum of three weeks of observations (Table 1).

Table 1: Retailer and flyer information

Retailer	**Days collected**	**Format**	**Mean items per flyers**
Food Lion	21	PDF-Weekly	218
Kroger	21	PDF-Weekly	168
Indomaret	28	PDF-Biweekly	298
Super Indo	21	PDF-Weekly	201
Pick n Pay	28	PDF-Biweekly	214
Shoprite	28	PDF-Biweekly	228
Spar	29	PDF-Biweekly	199
Robinsons	28	PDF-Biweekly	30
Carrefour Kenya	28	PDF-Biweekly	34

Variables

To conduct the analysis, we extracted multiple data points, including product name (e.g. Corn Flakes), brand name (e.g. Kellogg's), unit (e.g. 1 kg), and page number (i.e. the page where the product promotion is located) (Figure 4). AI was then used to assign each product to one of four categories: fresh, packaged, alcohol, and non-food. Alcoholic and non-food products were outside the scope of ATNi's research and therefore excluded from analysis.

Figure 41. Example of retailer flyer and variables to extract

Choice of AI

There are numerous types of AI solutions available, each varying in features, limitations, and cost, and their suitability depends on the operational context. For our purposes, we required a large language model (LLM) capable of identifying and extracting text embedded within images, such as those contained in pictures or PDF documents. Several AI providers were explored and tested, after which ChatGPT was selected as the most appropriate for our needs. It should be noted, however, that the study did not involve a systematic benchmarking of AI tools. ChatGPT may therefore not be the most suitable choice in all cases.

Structure Instruction Prompt

Prompt engineering is a relatively new discipline concerned with developing and optimizing prompts to effectively leverage language models for a wide range of applications and research purposes. In practice, prompt engineering often involves interacting with LLMs through an API and adjusting model parameters. In our case, however, no parameter tuning was performed; instead, prompt engineering focused exclusively on modifying the elements and structure of the prompts. This approach was sufficient for the objectives of our study.

Instruction prompting builds on the capacity of generative AI models, particularly LLMs, to interpret and follow natural language instructions. Its purpose is to enable models to perform new, previously unseen tasks by understanding and executing instructions expressed in plain language, without the need for task-specific training data. For this study, a structured instruction prompt was developed following the best practices recommended for ChatGPT, to ensure consistency and comparability in the extraction of information from retailer flyers (Figure 5).

```
Act as an expert document interpreter specializing in extracting structured data from visual materials. I want you to transcribe information from a
supermarket flyer into a structured format. Your task is to identify and extract the product name, brand name, and unit of measurement (e.g., 500g, 1L,
3-pack) for each listed item. The flyer will be provided as [image/PDF], and the language is in [France/Bahasa Indonesia/English]. Keep the language when
transcribing.
Extract all product listings, ensuring that if multiple products appear on one line, each becomes its own row . For each product, include:
• product_name
• brand_name
• unit (e.g., "500g," "1L," "3-pack," "12-pack 11–12 fl oz")
• category
• ambiguous_category – a boolean value set to true for items where categorization might be unclear
• page_number – indicating the page the item was found on

Handle brand_name extraction as follows:
– If a line shows multiple brands separated by "/" or "or" (e.g., "Twisted Tea / Cayman Jack / Mike's"), and there are exactly as many brand segments as
product_name segments, assign each product its corresponding single brand by position (e.g., "Twisted Tea," "Cayman Jack," "Mike's").
– If there is one brand applied to all products on the same line, assign that single brand_name to each split row.
– If you cannot unambiguously match a brand to a given product, use "N/A."

Handle unit extraction as follows:
- If a line shows product_name with multiple unit (different volume or size), record them under a single entry using a unified format (e.g., product name
= "Milk", unit = "250 mL / 1 L"), rather than listing each variant separately.

Handle categorization as follows:
1. "Packaged" = packaged food and beverages - such as biscuits, milk, chips, juices, soda, canned food, ice cream, chocolate, drinks etc.
2. "Fresh" = fresh food and beverages - such as fruits, vegetables, unprocessed meat, eggs, etc. Some meat might be sold in packaged (for example: chicken
breast fillet, frozen seafood), but as long as they are not processed, they are considered fresh.
3. "Alcohol" - including alcoholic drinks like beer, wine, cider, etc.
4. "Non-food" - including toiletries, electronics, detergent, and other non-edible household goods.

Process the flyer in a consistent and orderly manner, following the logical visual sequence of the layout (e.g., top-to-bottom, left-to-right), rather
than jumping between unrelated sections.
Stay within the given flyer and don't add anything that's not there. If any field is missing for a product, use "N/A" as the value.
Display all transcribed data at once after processing the flyer—do not summarize or filter for brevity.
Return the result in markdown table
```

Figure 5 2. AI Prompt Template

Workflow

Data collection in this study followed a hybrid approach that combined automated extraction with manual review. For flyers available in PDF or picture format, product names were extracted using the optical character recognition (OCR) feature of a generative artificial intelligence tool, ChatGPT-o4-mini.

Flyers in .pdf, .jpg, or .jpeg format were first uploaded to the AI chat interface together with a designed prompt (Figure 6). The AI then generated an output file in markdown table format according to the extraction instructions. At this stage, a face validity review was conducted by comparing the AI output with the original flyers. The purpose of this review was to ensure that the AI has captured the expected variables at a basic level. Typical questions asked during this process include: "Are all pages represented?" and "Are all columns correctly extracted?" If the review identified gaps or errors, the prompt was revised, the AI was re-run, and the new output was reassessed. Once the AI output passed this review, it was saved as the 'AI Output' file. All data was compiled and organized in a Microsoft Excel spreadsheet.

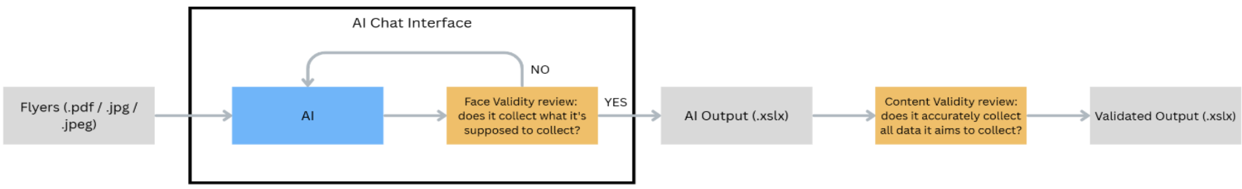

Figure 63. Data collection workflow

The next step involved a content validity review, which evaluated whether the AI had not only extracted the intended variables but also captured them with accuracy and completeness. To preserve the integrity of the AI output, a copy was created and renamed the 'Validated Output' file. A designated research team member then compared this copy against the raw flyers, manually correcting any inconsistencies or errors. This ensured that the validated dataset accurately reflected the source material. In this framework, the original 'AI Output' file was preserved as an unaltered record of the AI's performance, while the 'Validated Output' file became the gold standard dataset that reflects the flyer. This dual-file approach allowed the team both to measure the accuracy of the AI extraction and to maintain a rigorously validated dataset for downstream use.

Measure of Validity

To evaluate the validity of AI-based data extraction, two datasets were compared:

- AI Output dataset – containing product details as identified automatically by the AI
- Validated dataset – containing the same product details, but reviewed and corrected by human validators

Both datasets included the same type of variables: product name, brand name, unit size, category, and page number. Three error metrics were defined to measure validity:

- Proportion of omission – the proportion of products that were present in the flyers but not extracted by the AI.
- Proportion of hallucination – the proportion of products that were not present in the flyers but appeared in the AI extraction.
- Proportion of mis-extraction – the proportion of products that were present in the flyers and extracted by the AI but captured inaccurately or incompletely. Two levels of analysis were employed:
    - Row-Level – a product row was considered improperly extracted if any of its constituent fields (cells) contained an error. This provides a product-level perspective, reflecting the accuracy of the extraction process at the unit of the product entry.
    - Cell-Level – each individual field (cell) within a product row was evaluated separately. This offers a more granular view of error distribution within the dataset.

4.2 RESULTS

Flyers from nine retailers, each contributing at least three weeks of observations, resulted in 2,963 product rows extracted by the AI. Across the whole dataset, the proportion of omissions was 24.40%, while the proportion of hallucinations was 4.05%. Omission rates varied across retailers, ranging from a high of 36.8% at Pick n Pay to 0% at Robinsons and Carrefour Kenya (Table 2). Hallucination rates showed a similar spread, from 7.08% at Spar to 0% at Robinsons and Carrefour Kenya.

Table 2: Proportion by retailers (%)

	Omission	Hallucination	Mis-extraction on Row Level	Mis-extraction on Cell Level
Food Lion	28.9	3.7	54.1	16.1
Kroger	13.9	4.4	29.5	6.6
Indomaret	30.1	3.0	36.5	10.8
Super Indo	18.9	4.9	24.1	6.1
Pick n Pay	36.8	1.8	46.3	13.1
Shoprite	28.1	4.4	34.1	8.8
Spar	20.7	7.1	19.7	6.6
Robinsons	0	0	9.1	2.7
Carrefour Kenya	0	0	11.9	2.4

The reasons underlying these differences remain uncertain, though plausible explanations can be proposed. For example, the elevated omission rate at Pick n Pay may be attributed to the flyers' densely packed layout and small font size, which likely impeded accurate extraction. In contrast, the clean layout and shorter product lists of Robinsons and Carrefour Kenya flyers may have facilitated complete coverage with no omissions or hallucinations.

Regarding mis-extraction, 33.91% of rows contained at least one error, while the cell-level mis-extraction rate was 9.48%. This pattern indicates that, although the AI captures most information accurately at the attribute level, errors are dispersed across many rows, which inflates the row-level error rate. Among retailers, Food Lion exhibited the highest mis-extraction rates, both at the row level (54.1%) and the cell level (16.1%) (Table 2). A key factor contributing to this outcome appears to be the way product names were presented in the flyers. For example, product names were often written in continuous phrases such as "Breyers Ice Cream and Klondike Bars," which the AI recorded as a single entry rather than two distinct products. This not only increased mis-extraction but also contributed to a higher omission rate, as one product was effectively lost when merged with another.

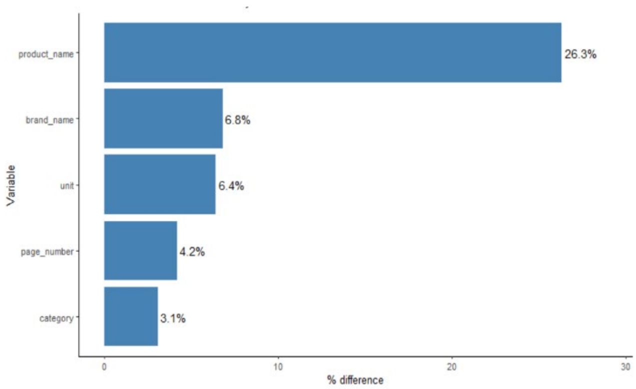

Figure 74. Percentage of cell-level mis-extraction by variable

Further investigation confirmed that the product name field was the variable most prone to mis-extraction (Figure 7). This may be explained by the relative length and complexity of product names compared to other fields, such as brand name. By contrast, the category variable exhibited the lowest mis-extraction rate (3.1%). This likely reflects the fact that it was assigned directly by the AI rather than extracted from text. These findings suggest that, while the AI struggles with parsing complex product names, it demonstrates stronger performance in categorizing products at a broader level.

4.3 DISCUSSION AND FUTURE USE CASES

The usefulness of AI will always depend on the context: This case study shows how AI can be applied to extract information from retail flyers, and to assess its validity using simple measures such as omissions, hallucinations, and mis-extractions that are easy to interpret without technical expertise. By comparing outcomes across different retailers and flyer designs, it highlights how AI performance varies by context, while also showing the limits of generalization. The analysis focused primarily on text fields, numeric values, and category flags, leaving out more complex content such as pricing or type of promotion, and should therefore be understood as an exploratory case study rather than a definitive evaluation of best practices.

Future research may benefit from further automation: At present, two components of the flyer extraction process still require manual effort: uploading materials and reviewing the outputs. Looking ahead, it may be worthwhile to explore options for automating the first component. One potential pathway is to batch upload via an API, where a script could loop through a folder of images or PDFs, submit them for extraction, and save the outputs automatically. While this approach requires some technical expertise, it would allow human effort to be concentrated on the most critical task: reviewing and validating the results.

AI can scale nutrition research with greater efficiency: These findings contribute to the growing body of literature on potential use-cases in nutrition-related research.[16] For example, AI could be used to extract nutrition facts panels from packaged food labels, automate data entry from food diaries or dietary questionnaires, or capture product and pricing information from supermarket shelves, menus, or restaurant boards to study food availability and affordability. Automating data extraction would enable researchers to gather larger and richer datasets, thereby increasing the validity and relevance of their findings. In some cases, such automation could even contribute to larger public datasets, where enrichment can be accelerated and expanded by the wider community of users. While specialized software could be developed for such tasks, it is often costly and requires substantial technical resources. AI-based solutions offer the potential to make automation more accessible and practical for researchers working in this space.

Human–AI combination ensures the most reliable results: Taken together, this case study shows that AI can be a powerful assistant in extracting and organizing data, but its effectiveness depends on both the type of material and the variables being extracted. Text, numbers, and categories present different levels of challenge, and layout strongly influences accuracy. Human review is still needed to catch errors and identify bias. The most reliable results come from human–AI combinations, where AI accelerates the work and humans safeguard accuracy. Rather than replacing researchers, combining human analysts and AI in research calls for new and adaptive skills that help draw the best from both human expertise and AI capabilities.

5 REFERENCE

1. Sosa-Holwerda A, Park OH, Albracht-Schulte K, Niraula S, Thompson L, Oldewage-Theron W. The Role of Artificial Intelligence in Nutrition Research: A Scoping Review. Nutrients. 2024 Jun 28;16(13):2066.

2. Reynier A. Exploring AI for F&B Benchmarking: What We Tried, What We Learned, and What's Next [Internet]. ATNi (Access to Nutrition initiative). 2024 [cited 2025 Sep 23]. Available from: https://accesstonutrition.org/exploring-ai-for-fb-benchmarking-what-we-tried-what-we-learned-and-whats-next/

3. EU AI Act Compliance Checker | EU Artificial Intelligence Act [Internet]. [cited 2025 Sep 22]. Available from: https://artificialintelligenceact.eu/assessment/eu-ai-act-compliance-checker/

4. World Economic Forum. How nations can build sovereign AI and homegrown talent for economic competitiveness [Internet]. Emerging Technologies. 2024. Available from: https://www.weforum.org/stories/2024/07/sovereign-ai-talent-improve-economic-competitiveness/

5. A Global Digital Compact — an Open, Free and Secure Digital Future for All: Our Common Agenda Policy Brief 5 [Internet]. 2023 May [cited 2025 Sep 23]. (UN Executive Office of the Secretary-General (EOSG) Policy Briefs and Papers). Available from: https://www.un-ilibrary.org/content/papers/10.18356/27082245-28

6. UNITED NATIONS CONFERENCE ON TRADE AND DEVELOPMENT. TECHNOLOGY AND INNOVATION REPORT 2025: inclusive artificial intelligence for development. S.l.: UNITED NATIONS; 2025.

7. How to build sovereign AI to boost economic competitiveness [Internet]. World Economic Forum. 2024 [cited 2025 Sep 23]. Available from: https://www.weforum.org/stories/2024/07/sovereign-ai-talent-improve-economic-competitiveness/

8. Hill A. Why 95% Of AI Pilots Fail, And What Business Leaders Should Do Instead [Internet]. Forbes. [cited 2025 Sep 22]. Available from: https://www.forbes.com/sites/andreahill/2025/08/21/why-95-of-ai-pilots-fail-and-what-business-leaders-should-do-instead/

9. Artificial Intelligence (AI) in Science - Research and innovation [Internet]. 2025 [cited 2025 Oct 6]. Available from: https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/artificial-intelligence-ai-science_en

10. Prediger M, Huertas-Garcia R, Gázquez-Abad JC. Store flyer design and the intentions to visit the store and buy: The moderating role of perceived variety and perceived store image. Journal of Retailing and Consumer Services. 2019 Nov 1;51:202–11.

11. Dillon M. The 2023 Global Shopper Survey: Understanding shopper habits, an international persepctive. 2023;

12. van Lin A, Keller K, Guyt J. Retiring the Store Flyer: Effects of Ceasing Print Store Flyers on Household Grocery Shopping Behavior [Internet]. Rochester, NY: Social Science Research Network; 2025 [cited 2025 Apr 11]. Available from: https://papers.ssrn.com/abstract=4966328

13. Tan PJ, Tanusondjaja A, Corsi A, Lockshin L, Villani C, Bogomolova S. Audit and benchmarking of supermarket catalog composition in five countries. International Journal of Advertising [Internet]. 2023 Apr 3 [cited 2025 Apr 14]; Available from: https://www.tandfonline.com/doi/abs/10.1080/02650487.2022.2079818

14. Gázquez-Abad JC, Sánchez-Pérez M. How Store Flyers Affect Consumer Choice Behaviour: National Brands vs. Store Brands. In: Morschett D, Rudolph T, Schnedlitz P, Schramm-Klein H, Swoboda B, editors. European Retail Research [Internet]. Wiesbaden: Gabler Verlag; 2009 [cited 2025 May 23]. p. 1–20. Available from: https://doi.org/10.1007/978-3-8349-8203-2_1

15. Boto-García D, Alvarez A. Modelling the effect of store flyers on supermarket sales: An application to olive oil demand. Journal of Retailing and Consumer Services. 2020 May 1;54:102057.

16. Sosa-Holwerda A, Park OH, Albracht-Schulte K, Niraula S, Thompson L, Oldewage-Theron W. The Role of Artificial Intelligence in Nutrition Research: A Scoping Review. Nutrients. 2024 Jun 28;16(13):2066.

ATNi (Access to Nutrition initiative)
Arthur van Schendelstraat 650
3511 MJ Utrecht
The Netherlands
+31 (0)6 429 51 655
info@atni.org
www.atni.org