CAS Essays on the Intersection of Artificial Intelligence and Actuarial Science

The New Insurance Toolkit: Human-AI Partnerships

By Sergey Filimonov

In the span of just a decade, computers have gone from failing to differentiate cats from dogs, to outperforming 92 percent of humans on the SAT. It's no longer difficult to imagine a world where AI models, trained on vast datasets, can outperform humans in most tasks we consider quintessentially human.

This technological leap has profound implications for the insurance industry, which has historically relied on a combination of heuristics, machine learning models, and human judgment to assess and price risk. The advent of generative AI equips carriers with new tools to reimagine countless aspects of their business. This will result in more personalized pricing, enhanced customer service, new strategies to combat fraud, and countless further innovations yet to be imagined.

Building upon this, our startup has spent the last year developing AI tools to analyze insurance filings. Working closely with carriers, we've gained insights into the practical challenges of deploying AI within the industry. This essay aims to share those findings, along with strategies for successful human-AI collaboration, and explore their broader implications for the insurance sector.

To fully appreciate the implications of this shift, it's helpful to explore the progression of AI technologies.

A Brief History of Language Models

The evolution of artificial intelligence has progressed through distinct phases. Early symbolic AI, emerging in the 1950s, excelled in domains where all necessary knowledge could be explicitly encoded and problems were well-defined, paving the way for achievements in expert systems and chess-playing programs like IBM's Deep Blue.

However, these models struggled with tasks requiring nuanced understanding or handling ambiguous or incomplete information. The fundamental issue was that computers were far too slow to do anything interesting. As a result, this led to an overarching negative sentiment in the field that computers were fundamentally constrained in the types of problems they could solve.

This sentiment dominated until the mid-2010s. However, an unlikely catalyst was propelling a revolution — the gaming industry's insatiable demand for powerful graphics processing units (GPUs). These GPUs proved unexpectedly well-suited for the parallel computations required by neural networks. This technological shift, coupled with Moore's

Law's relentless march, set the stage for early breakthroughs like AlexNet's groundbreaking performance in image recognition and AlphaGo's mastery of the complex game of Go.

These advancements inspired a new wave of ambitious AI research efforts, including the founding of OpenAI in 2015. Attracting top AI talent, they quickly established themselves at the forefront of natural language processing (NLP) research.

A critical breakthrough came in 2018 with Google's release of the seminal "Attention Is All You Need" paper, which introduced the transformer architecture. The transformer's unique design significantly reduced the computational requirements for training large language models (LLM) compared to previous approaches. This would set the stage for OpenAI and others to push the boundaries of what was possible.

OpenAI's initial models were small and could only learn basic patterns, such as the presence of characters or punctuation rules. But early experiments revealed a surprisingly straightforward formula for achieving new capabilities - simply take a transformer-based language model, progressively scale up its parameters, and train it on ever-growing datasets. Each iteration produced powerful new capabilities as the models grew more eloquent, contextual, and insightful.

The release of GPT-3 in 2020 marked a watershed moment demonstrating that a language model could generate near-human quality content. In blind tests, humans could barely distinguish articles written by the model from those written by other humans. This milestone galvanized OpenAI's efforts towards the process of bringing GPT-3 to the commercial market.

By late 2022, the company would embark on training its most ambitious model yet, spending an estimated 100 million dollars. GPT-4's performance was unprecedented, reaching human-level scores on rigorous exams like the BAR, MCAT, and passing nearly all AP exams. This advancement propelled LLM's from niche academic pursuits and limited practical applications into a powerful tool capable of tackling complex cognitive challenges.

Coincidentally, they also launched an experimental chat interface, ChatGPT, for interacting with their models. To their complete surprise, this quickly went viral. By January 2023, it had become the fastest growing consumer application of all time, gaining over 100 million users. The launch of ChatGPT brought the profound capabilities of GPT-4 directly into the hands of millions, effectively democratizing access to advanced AI and illustrating its broad applicability.

Figure 1. Amortized hardware and energy cost to train frontier AI models over time (Source: Cottier et al., 2024).

The implications of these powerful models are vast. However, for organizations whose core expertise lies outside AI, it can be difficult to assess potential applications amidst the rapid pace of innovation.

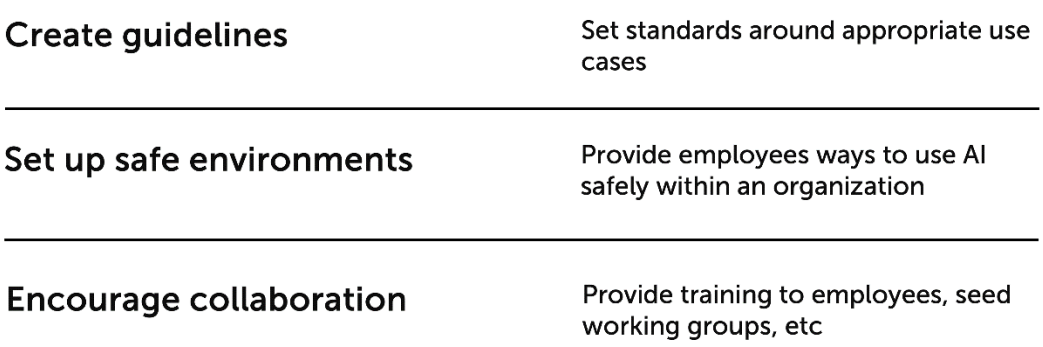
Seeding AI Bets

Fundamentally AI applications within an enterprise fall into two main areas: internal process improvements and external, customer-facing initiatives.

Internal improvements are an ideal starting point for AI experimentation due to the lower stakes, yet formulating a comprehensive AI strategy remains challenging given the vast space of possible solutions.

Determining the optimal applications of AI requires a collaborative approach - no single individual can definitively answer the question of what to build. Successful strategies leverage diverse perspectives and expertise.

Figure 2. AI within an enterprise



Companies often begin with internal experimentation, fostering partnerships between technical resources and domain experts while providing dedicated budget and time for exploration. This allows teams with deep domain knowledge to apply AI directly to the problems they know best, while also gaining familiarity with AI. Hackathons, working groups, newsletters, etc., can be the initial catalyst needed to get ideas flowing from within different parts of the organization.

Stripe's AI Evolution

Stripe, a leading online payment processing platform, has been an early OpenAI partner and has seen much success with the following approach.

Stripe's strategy began with deploying a generative AI tool — specifically, a ChatGPT-like interface — across the company, creating an internal prompt library that employees could contribute to and upvote for usefulness. This initiative quickly saw widespread adoption, with one-third of Stripe's workforce utilizing the tool in a matter of days. By analyzing usage logs and identifying common use cases, Stripe was able to pinpoint effective applications of AI across different parts of their business.

Recognizing the challenges of diverting engineering resources from existing projects, Stripe established small teams equipped with a dedicated budget, and a six-month timeline to explore targeted AI initiatives. Importantly, these accelerators offered growth and development opportunities for their internal talent.

This grassroots approach directly contributed to multiple products that are now live. An example is "Radar Assistant" — a tool that empowers their customers to create custom fraud rules using natural language. For example, it can allow a business owner to block transactions from countries where they have no customers using plain English. This enables fraud analysts and less-technical users to directly implement rules without extensive developer support.

Stripe's experience highlights a replicable pattern for other domains like underwriting: Combining natural language policy descriptions with AI offers significant potential to streamline decision engines and reduce reliance on engineering resources.

Navigating the AI Landscape

As an organization shifts from various prototypes to production-ready solutions, it's critical to figure out what's unique to them and what's a commodity that should be purchased externally. This often comes down to a tradeoff between leveraging unique data and expertise or finding solutions with broad applicability.

Careful selection of foundational models, which are large, pre-trained language models that serve as the base for more specialized applications, is also a key part of a long-term AI strategy. This can be challenging, given the ever-increasing landscape of options.

Offerings from major cloud providers like OpenAI or Google are the most powerful options and can provide SOC2 and HIPAA compliance out of the box. However, these solutions can also lead to higher costs and often require extensive legal and procurement approvals.

Conversely, self-hosted open-source solutions can offer greater flexibility, but often demand powerful hardware and currently underperform commercial offerings.

Choosing the right foundational model involves balancing your organization's specific needs, budgetary limits, and compliance requirements. It's worth noting that experimenting with different models isn't overly burdensome, allowing organizations to pivot and adapt as they refine their AI strategy.

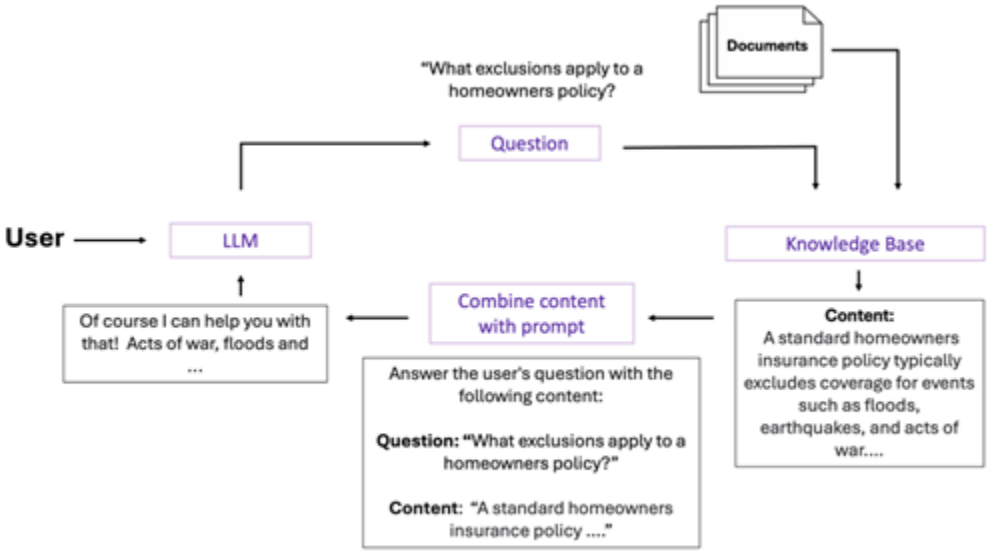
Mitigating Risk in AI Deployments

Beyond the technical considerations, designing effective AI solutions demands a strong focus on user experience (UX). This involves recognizing the iterative nature of interaction with generative AI. Allowing users to refine or retry requests is crucial, as the ideal output may not always be achieved on the first attempt.

Clearly communicating model limitations, coupled with suggestive prompts, also proactively guides the user towards better results and sets realistic expectations.

To mitigate the risk of generating inaccurate information or "hallucinations," it's essential to ground the model in the specific input context by providing it with explicit, factual information. This process involves a strategic approach where the user's prompt is not fed directly into the model. Instead, it undergoes a preliminary step known as AI grounding, where the prompt is carefully analyzed and matched against a search index containing company-specific information. This enriched context is then presented to the model, ensuring that its responses are anchored in verifiable data.

Figure 3. An overview of retrieval augmented generation



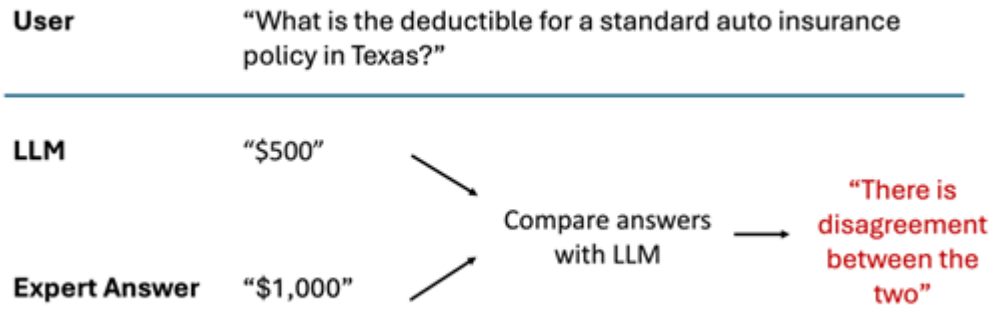
At this point, it's also crucial to establish guardrails: constraints, or preventative controls that sit between the user interface and the generative model. These safeguards aim to prevent harmful content from reaching users and add an essential layer of control when deploying AI solutions. This is especially critical in regulated industries like insurance. Rather than simply blocking problematic prompts, smart guardrails can suggest alternative prompts that guide users towards safer and more appropriate interactions.

However, even with powerful tools in place, operationalizing them presents a distinct challenge within this new paradigm. The non-deterministic nature of GenAI outputs makes traditional quality assurance measures largely inapplicable. This often represents a key challenge companies face in deploying to production.

As a result, a robust evaluation process is essential. This can take the form of human experts rating model responses against established criteria. However, when human feedback becomes impractical due to cost or scale, automated evaluations offer an alternative — generative AI itself can be used to monitor progress and detect regressions. By automating portions of the feedback process, human experts can focus their attention on the most complex edge cases.

For example, to evaluate the factual accuracy of AI-generated answers, you could submit both the AI's response and an expert-provided answer to a powerful model like GPT-4. The model could then compare the content and identify any discrepancies. This systematic approach becomes critical as your team experiments with different prompts and retrieval techniques.

Figure 4. A sample evaluation prompt



Powerful, large models like GPT-4 can be costly for frequent evaluations, and as a result, it's important to note that smaller models can be orders of magnitude cheaper. By fine-tuning smaller models on high-quality data curated by GPT-4, you can create specialized "judge" models. These models become highly adept at evaluating your specific use cases, offering significant cost savings compared to using larger models.

Human-AI Collaboration: The Key to Success

The core principle we've found of successful AI deployment lies in augmentation, not replacement. The AI assistant should augment a user's capabilities rather than replacing human judgement.

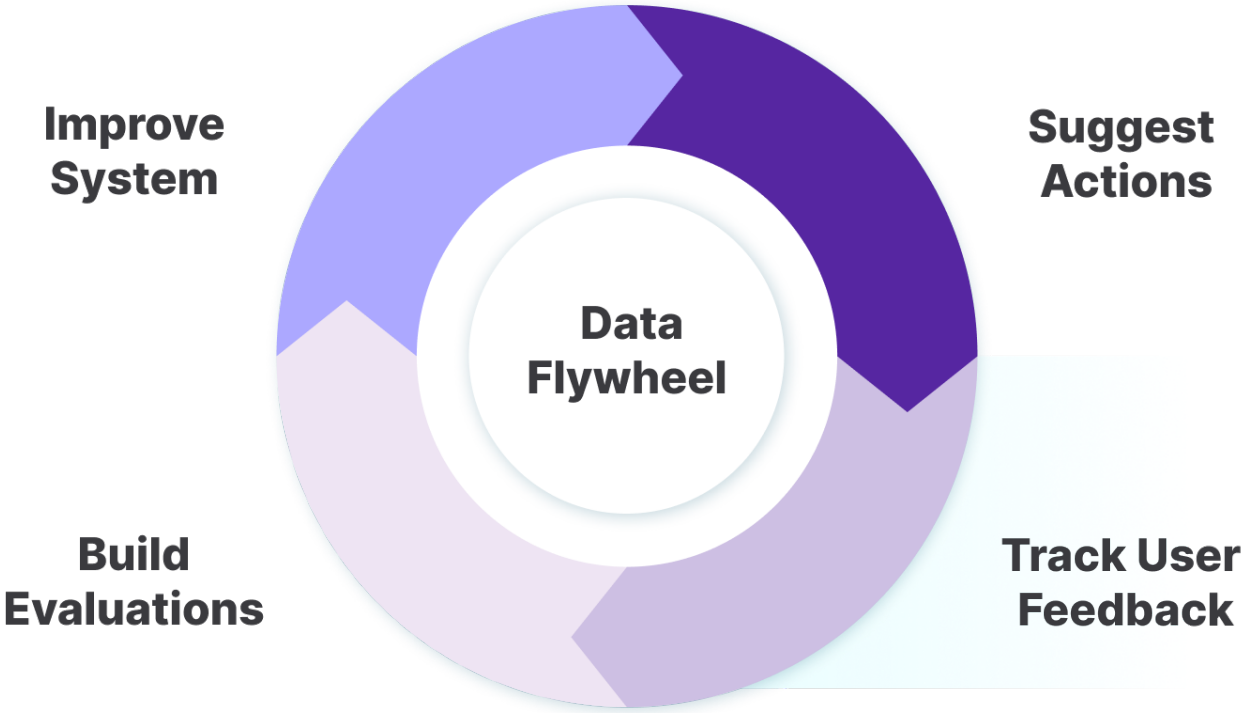
A tool should propose a change, but it requires human approval to accept, reject, or modify the suggestion. This human-in-the-loop approach provides a critical layer of control for addressing issues with model reliability. Microsoft describes this as having a co-pilot, a white-collar worker available to help you with various tasks.

Copilots are exceedingly powerful. They can equip non-technical users with the ability to generate solutions that traditionally demanded engineering resources. They also help technical users move significantly faster.

Most importantly, by collecting data on accepted, modified, and rejected suggestions, the system creates a feedback loop that can be used to continuously improve the product.

This data-driven approach even opens the possibility that if completions are consistently accepted above a certain threshold, full automation could be considered. As AI technology advances, the potential for agents to take on increasingly complex roles grows. These agents could proactively perform tasks, monitor processes, and even interact with customers, always learning and refining their actions based on the established feedback loop.

Figure 5. Feedback loop for continuous product improvement



The rapid evolution of AI presents both immense opportunities and challenges for the insurance industry. As AI capabilities continue to expand, this data-driven, collaborative approach offers the most promising path forward. The insurance industry stands at a pivotal moment; embracing human-AI partnerships isn't merely a competitive advantage, it's the key to redefining the future of insurance.

Reference

[1] Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, and David Owen. 'The rising costs of training frontier AI models'. ArXiv [cs.CY], 2024. arXiv. https://arxiv.org/abs/2405.21015.