Gen Re

Harnessing the Transformative Power of Gen AI

Society for Insurance Financial Management (SIFM) Holiday Conference
December 5, 2024

Frank Schmid
Chief Technology Officer

Overview

- What can gen AI do for you?
- What specifically does a Large Language Model (LLM) do?
- Matters of note in the use of LLMs
- A brief history of gen AI
- From modeling causation to identifying connections
- Gen AI is general-purpose technology (GPT)
- Lessons for transformation from past arrivals of GPTs
- Three challenges of transformation at enterprise level
- Activities projected to be transformed by automation
- Change in workplace skills demand
- Potential implications for labor
- Implications for insurance professionals

Frank Schmid, "Generative Artificial Intelligence in Insurance - Three Lessons for Transformation from Past Arrivals of General-Purpose Technologies," Gen Re, March 18, 2024, https://www.genre.com/us/knowledge/publications/2024/march/generative-artificial-intelligence-in-insurance-2-en.

Frank Schmid, "Generative Artificial Intelligence in Insurance – Four Aspects of the Current Debate," Gen Re, February 1, 2024, https://www.genre.com/us/knowledge/publications/2024/february/generative-artificial-intelligence-in-insurance-en.

What Can Gen AI Do for You?

- Gen AI enables humans to have a *conversation with a body of knowledge*
  - Large Language Models (LLMs) enable the *conversation*
  - LLMs represent the *body of knowledge*
  - The body of knowledge of the LLM may be augmented by information from a retriever (*grounding*).
- Multimodal models enable human-like interaction with a machine
  - Multimodal models mimic the multidimensionality of human cognition (text, images, and audio)
  - Google's Gemini is natively multimodal—Google's project Astra develops multimodal AI assistants.[1]
- *Embodied* multimodal LLMs in robots
  - AI-enabled robots can handle the unexpected

    "…the ability to generalize to new tasks not seen during training…" [2]

1) See Google DeepMind, "Gemini," https://deepmind.google/technologies/gemini/, accessed May 23, 2024.
2) See Google Research, "PaLM-E: An embodied multimodal language model," March 10, 2023, https://blog.research.google/2023/03/palm-e-embodied-multimodal-language.html.

What Specifically Does an LLM Do?

- Text representation

  …is the process of converting textual data into machine-readable format (e.g., embeddings[1]) with the goal of capturing the meaning and context of the text.

  …is an important step in natural language processing and information retrieval tasks, such as text classification, sentiment analysis, and information extraction.

- Text generation

  …is a type of natural language processing task that involves generating text from a given input with the goal of resembling human-written context.

  …is used for summarizing and synthesizing content, creating draft responses, and translating languages.

- The difference between the tasks of *text representation* and *text generation* becomes apparent in the processing of legal documents—in text generation, LLMs are challenged in delivering sufficient precision.[2]

1) An embedding is a relatively low-dimensional space into which one can translate high-dimensional vectors. Embeddings make it easier to do machine learning on large inputs like sparse vectors representing words. Ideally, an embedding captures some of the semantics of the input by placing semantically similar inputs close together in the embedding space. An embedding can be learned and reused across models. See Google, "Embeddings," Machine Learning: Foundational Courses, https://developers.google.com/machine-learning/crash-course/embeddings/video-lecture.

2) Current OpenAI models are non-deterministic, even when the temperature parameter is set to zero. See "Non-determinism in GPT-4 is caused by Sparse MoE," https://152334h.github.io/blog/non-determinism-in-gpt-4/, published on August 5, 2023, last updated on June 23, 2024.

Matters of Note in the Use of LLMs

- LLMs are trained to represent and generate text—text is not synonymous with knowledge
  - The LLM training corpus has a cutoff date,[1] and the body of knowledge is static.[2]
  - Asking a gen AI chatbot questions beyond the training corpus poses the risk of hallucinations.
- Retrieval-augmented generation (RAG)
  - An information retrieval system provides *grounding data* to the LLM—there is no extra training[3]
  - The grounding data may be company-internal data or, in generative search, data from the public web.[4]
- Gen AI output is an approximation
  - Conceptually, the Xerox scanner-copier mystery, uncovered by David Kriesel in 2013, lives on.[5,6]

1) The data OpenAI's GPT-4 was trained on cuts off in September 2021. The latest version, GPT-4 Turbo, has an April 2023 cutoff date. "...GPT-4 does not learn from its experience." https://openai.com/research/gpt-4.

2) LLMs, such as GPT-4 or Google's Gemini, are subject to ongoing reinforcement learning by human feedback (RLHF). RLHF steers the behavior of the LLM rather than adding to the model's body of knowledge.

3) See Microsoft, "Retrieval Augmented Generation (RAG) in Azure AI Search," November 20, 2023, https://learn.microsoft.com/en-us/azure/search/retrieval-augmented-generation-overview.

4) Examples of generative search are Bing Chat in the Microsoft Edge browser and Google Gemini chat.

5) See Ted Chiang, "ChatGPT Is a Blurry JPEG of the Web," The New Yorker, February 9, 2023, https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web. For details on the Xerox copier mystery, which originates in data compression in the process of scanning and archiving, see https://dkriesel.com/blog/2013/0802_xerox-workcentres_are_switching_written_numbers_when_scanning.

6) Microsoft asserts that optical character recognition (OCR) integration allows GPT-4 Turbo with Vision to produce higher quality responses for dense text, transformed images, and number-heavy financial documents. See https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/gpt-with-vision?tabs=rest%2Csystem-assigned%2Cresource.

A Brief History of Gen AI

The deep learning revolution (2012)

- The deep learning revolution erupted when a neural network dubbed AlexNet won the 2012 ImageNet challenge by a wide margin
  - ImageNet was an annual contest where research teams around the world had their machine learning models compete in classifying a vast body of images
  - AlexNet was the only neural network among the seven contestants in the 2012 contest.[1]
- Traditionally, for an algorithm to perform a classification task, it had to be instructed what to pay attention to, a process known as *feature engineering*.
- Deep learning not only relates identified patterns in data to outcomes, but it also learns of the presence of patterns, a process known as *feature learning*
  - For instance, a deep learning algorithm learns on its own the defining features of a cat from labeled cat images.

1) See ImageNet, https://image-net.org/challenges/LSVRC/2012/results.html.

A Brief History of Gen AI

The role of NVIDIA Graphics Processing Units (GPUs)

- AlexNet was one of the first classification models to run on NVIDIA GPUs
  - Specifically, AlexNet was trained on two NVIDIA circuit boards[1]
  - In comparison, a neural network that Google had trained earlier that year to identify videos of cats required some 16,000 Central Processing Units (CPUs).[2]
- Developed for graphics processing, a GPU breaks complex mathematical tasks apart into small calculations, then processes them in parallel.[3]
- In 2006, NVIDIA launched CUDA, a proprietary parallel computing software layer that allows GPUs to be repurposed as *accelerators* to workloads beyond graphics.[4]
- Says NVIDIA's Ian Buck, in around 2012 "AI found us."[5]

1) See Stephen Witt, "How Jensen Huang's Nvidia Is Powering the A.I. Revolution," The New Yorker, November 27, 2023, https://www.newyorker.com/magazine/2023/12/04/how-jensen-huangs-nvidia-is-powering-the-ai-revolution.

2) Ibid.

3) See Tim Bradshaw and Richard Waters, "How Nvidia created the chip powering the generative AI boom," Financial Times, May 26, 2023, https://www.ft.com/content/315d804a-6ce1-4fb7-a86a-1fa222b77266.

4) Ibid.

5) Ibid.

A Brief History of Gen AI

The Transformer (2017)

- The Transformer was introduced by Google Brain in 2017.
- Now the dominant neural network architecture, the Transformer is based on a self-attention mechanism, which directly models relationships among all words in a sentence[1,2]
  - For example, deciding on the most likely meaning of the word "bank" in the sentence "I arrived at the bank after crossing the…" requires knowing if the sentence ends on "street" or "river"
  - To determine that the word "bank" refers to the shore of a river and not a financial institution, the Transformer can learn to attend immediately to the word "river" and make this decision in a single step.
- Unlike older neural networks, the Transformer is scalable, horizontally and vertically
  - The self-attention architecture of the Transformer lends itself to *parallelization*
  - The Transformer is a solution to the *vanishing gradient problem*, allowing the stacking of very high numbers of hidden layers.

1) For the original paper on the Transformer, see Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention is All You Need," *arXiv*, submitted June 12, 2017, last revised August 2, 2023, https://arxiv.org/pdf/1706.03762.pdf.

2) For a gentle introduction to the Transformer, see Visual Storytelling Team and Madhumita Murgia, "Generative AI exists because of the transformer," *Financial Times*, https://ig.ft.com/generative-ai/.

From Modeling Causation to Identifying Connections

Machine learning (and deep learning in particular) changes the way we make predictions

- Traditional predictive modeling is based on structural equations derived from causal relations, proven (physics) or hypothesized (human behavior).
- Advances in weather forecasting demonstrates the predictive power of deep learning
  - GraphCast is a weather forecasting model developed by Google's DeepMind[1,2,3,4]
    - GraphCast is a neutral network trained on 40 years of data of the European Centre for Medium-Range Weather Forecasts (ECMWF). The AI system produces a 10-day forecast based on the states of the atmosphere worldwide currently and six hours earlier.
    - The forecasts have proven more accurate than those of the world's most advanced physical model, which is operated by ECMWF. This physical model uses supercomputers to crunch equations based on scientific knowledge of atmospheric physics—this is an energy-intensive process that takes several hours.
    - GraphCast, once trained, is cheap to operate. GraphCast produces a 10-day forecast within a minute on a single Google TPU (Tensor Processing Unit) v4 cloud computer.

1) Andrew Blum, "The weather forecast may show AI storms ahead." *Financial Times*, November 18, 2023, https://www.ft.com/content/332d1c05-e3ef-4bb2-8e7b-53f0222239d5.
2) Clive Cookson, "AI outperforms conventional weather forecasting methods for first time." November 14, 2023, *Financial Times*, https://www.ft.com/content/ca5d655f-d684-4dec-8daa-1c58b0674be1.
3) For details on GraphCast see Remi Lam et al. "Learning skillful medium-range global weather forecasting." *Science* 382: 1416–1421, 2023, https://www.science.org/doi/epdf/10.1126/science.adi2336.
4) NeuralGCM by Google Research combines AI with physics-based models for long-range climate simulations. For an introduction to NeuralGCM, see Michael Peel, "AI helps to produce breakthrough in weather and climate forecasting." July 22, 2024, *Financial Times*, https://www.ft.com/content/78d1314b-2879-40cc-bb87-ffad72c8a0f4. GCM: General Circulation Model. For details on NeuralGCM, see Dmitrii Kochkov et al. "Neural general circulation models for weather and climate." *Nature* 632: 1060–1066, 2024, https://doi.org/10.1038/s41586-024-07744-y.

Gen AI is General-Purpose Technology (GPT)

- A GPT (1) is *widely used*, (2) is capable of *ongoing technical improvement*, and (3) enables *innovation in application sectors*.[1]
- The arrival of a GPT is a rare event, even in modern times
  - Examples of older GPTs are the steam engine, the electric motor, and the semiconductor.
- The adoption of a GPT is gradual, and its productivity benefits take time to materialize
  - The GPT complements innovation in production processes, organizational design, and products
  - The inventions of the electric motor (ca. 1890) and the personal computer (1981) have given rise to productivity booms in the United States with time lags of 25 and 15 years, respectively.[2,3]

1) See Timothy F. Bresnahan, "General Purpose Technologies," in: Bronwyn H. Hall and Nathan Rosenberg (eds.) *Handbook of the Economics of Innovation*, Vol. 2, 761-791, 2010, https://www.sciencedirect.com/science/article/pii/S0169721810020022.

2) See Martin Wolf, "The threat and promise of artificial intelligence," *Financial Times*, May 9, 2023, https://www.ft.com/content/41fd34b2-89ee-4b21-ac0a-9b15560ef37c. See also Dominic Wilson and Vickie Chang, "Markets around past productivity booms," *Top of Mind* 120: 18-19, July 5, 2023, https://www.goldmansachs.com/intelligence/pages/top-of-mind/generative-ai-hype-or-truly-transformative/report.pdf.

3) For two studies on the possible productivity-enhancing effect of generative AI see Joseph Briggs and Devesh Kodnani, "The Potentially Large Effect of Artificial Intelligence on Economic Growth," *Global Economics Analyst*, Goldman Sachs, March 26, 2023, https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html; and Michael Chui et al., "The economic potential of generative AI: The next productivity frontier," McKinsey & Company, June 2023, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier.

Lessons for Transformation from Past Arrivals of GPTs

The velocity of the feedback cycle of complementary innovation

- A GPT unleashes a feedback cycle of technical improvement and downstream innovation.[1]
- For generative AI, the velocity of the feedback cycle is expected to be higher than it was for the steam engine, electricity, or electronic computing.
- A seminal study on transition of U.S. corporations from mainframe computing to C/S (client/server) computing observed that the organizations slowest to transition were those with the *highest cost* of adoption rather than the *lowest benefit* of adoption.[2]
- Technology choices that allow the insurer to benefit from a high velocity of complementary innovation emphasize the importance of learning and reversibility.[3]

1) See Timothy F. Bresnahan and Manuel Trajtenberg, "General Purpose Technologies 'Engines of Growth'?" *Journal of Econometrics* 65 (1): 83-108, 1995, https://www.sciencedirect.com/science/article/pii/030440769401598T.

2) See Timothy Bresnahan and Shane Greenstein, "Technical Progress and Co-Invention in Computing and in the Uses of Computers" *Brookings Papers on Economic Activity. Microeconomics* 1996: 1-83, 1996, https://www.brookings.edu/articles/technical-progress-and-co-invention-in-computing-and-in-the-uses-of-computers/.

3) The concept of real options valuation delivers the theoretical foundation for principles of technology decision-making that account for learning and irreversibility.

Lessons for Transformation from Past Arrivals of GPTs

The potential of a J-curve effect in productivity

- The arrival of a GPT requires an organization to make complementary investments.
- At the early stage of a GPT adoption process, the organization builds intangible assets
  - These assets are output that the organization produces alongside its products and services
  - Once built, these intangible assets serve as input to the organization's production process.
- Intangible assets tend to escape traditional measures of productivity
  - Productivity may briefly be underestimated before being overestimated for some time.[1]
- A crude measure of the insurer's productivity is the expense ratio[2]
  - At first, intangible assets add to the numerator, then they contribute to the denominator.

1) See Erik Brynjolfsson, Daniel Rock, and Chad Syverson, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," *American Economic Journal: Macroeconomics* 13(1), 333-372, 2021, Working Paper (January 2020) at https://www.nber.org/papers/w25148.

2) Although the expense ratio can serve as a meaningful measure of productivity for any given insurer, differences in expense ratios across insurers are not necessarily indicative of differences in productivity. This is due to differences in business models.

Lessons for Transformation from Past Arrivals of GPTs

The benefits of augmentation of labor

- The arrival of generative AI offers opportunities for automation
  - Automation can *augment* labor by enhancing and complementing the skills of humans…
  - …and automation can *substitute* labor.[1]
- The history of GPT adoption processes shows that although substitution and augmentation both occur, the latter offers the greater economic benefit by far[2]
  - Substitution of labor is a *static* concept and delivers one-off and immediate gains, whereas augmentation of labor is a *dynamic* concept and keeps delivering productivity gains over time.
  - To illustrate the impact of technology in the augmentation of labor, the market value of an hour of human labor, as measured by median wages, has grown more than tenfold since 1820.[3]

1) See David H. Autor, "Why Are There Still So Many Jobs: The History and Future of Workplace Automation," *Journal of Economic Perspectives* 29(3): 3-30, 2015, https://www.aeaweb.org/articles?id=10.1257/jep.29.3.3.

2) See Erik Brynjolfsson, "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence," *Daedalus* 151(2), 272–287, 2022, https://direct.mit.edu/daed/article/151/2/272/110622/The-Turing-Trap-The-Promise-amp-Peril-of-Human.

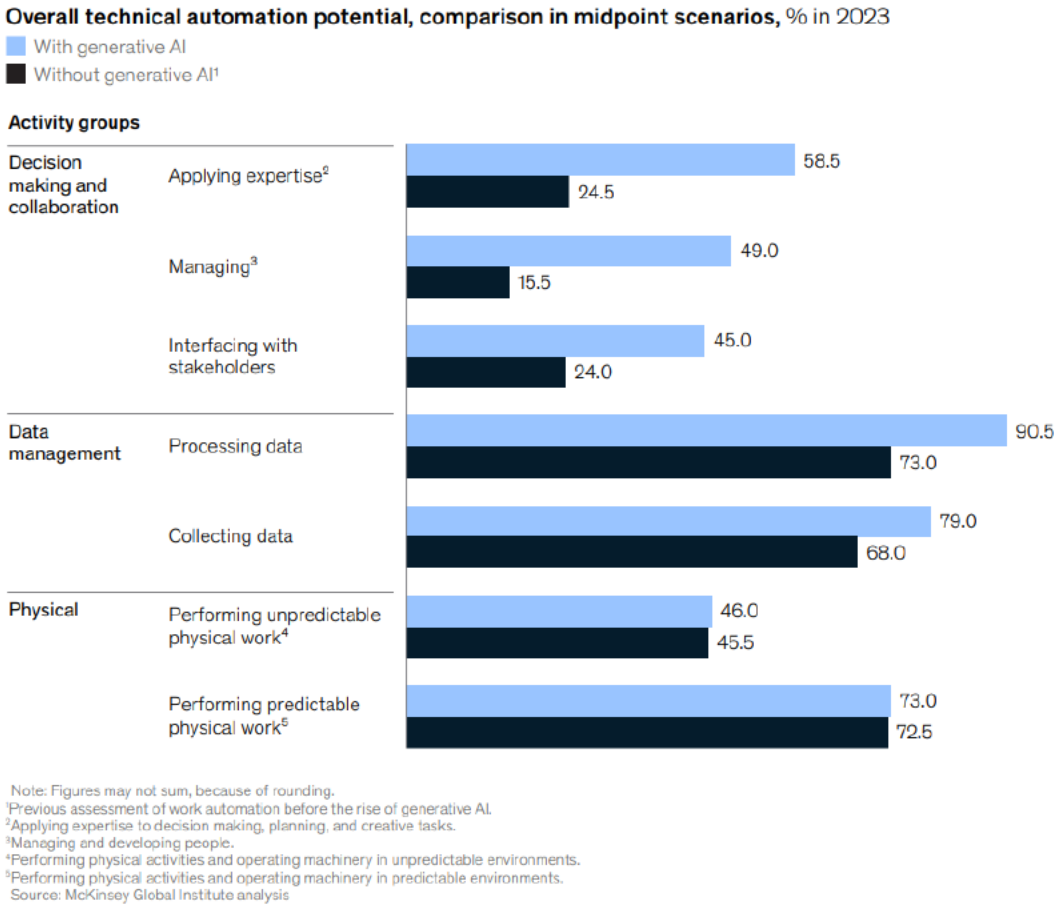
3) *Ibid.*

Three Steps of Transformation at Enterprise Level

1) A recent study estimates that task-level improvements (through automation, task complementarity, deepening of automation, and new tasks) will increase total factor productivity (TFP) in the U.S. economy over the next 10 years by a meagre cumulative 0.66 percent. See Daron Acemoglu, "The Simple Macroeconomics of AI," NBER Working Paper 32487, 2024, https://www.nber.org/system/files/working_papers/w32487/w32487.pdf.

Activities Projected to be Transformed by Automation

Gen AI is expected to make the greatest difference to decision-making and collaboration



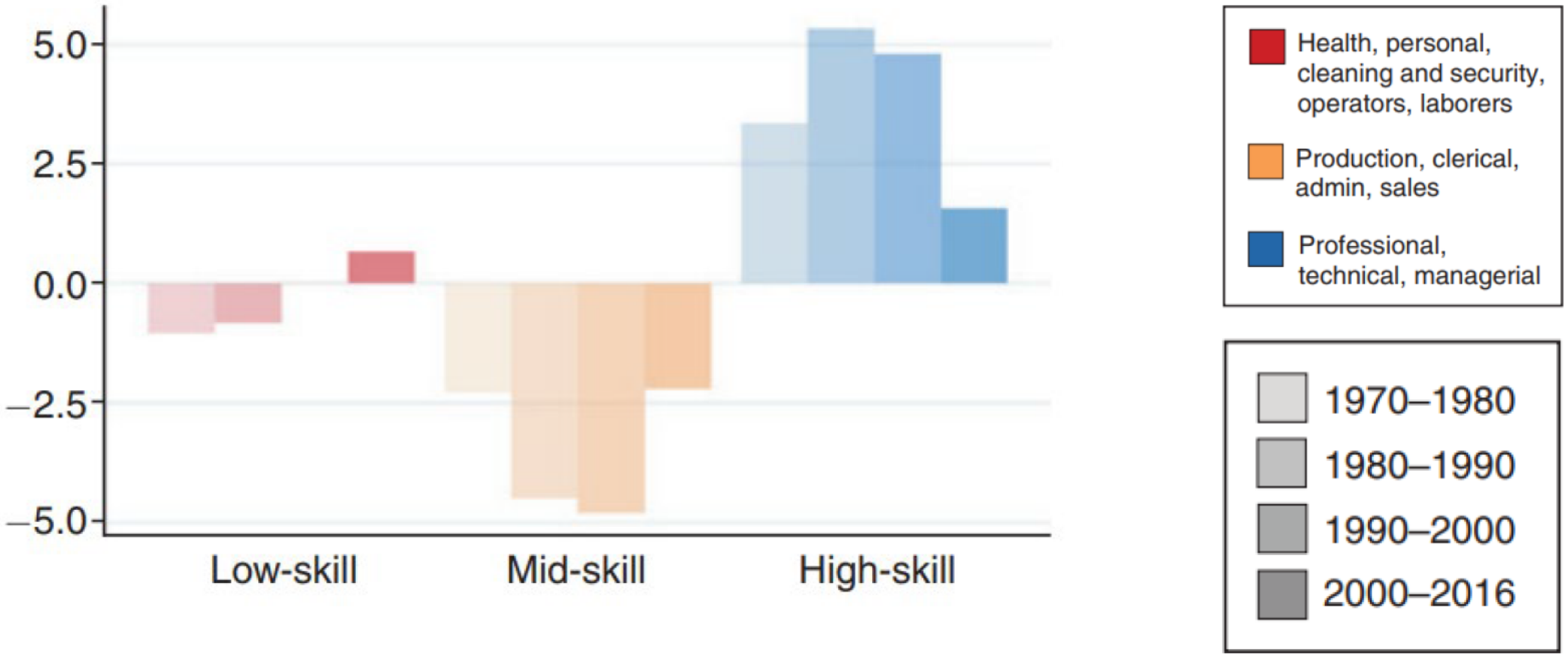
Note: Figures may not sum, because of rounding.
[1]Previous assessment of work automation before the rise of generative AI.
[2]Applying expertise to decision making, planning, and creative tasks.
[3]Managing and developing people.
[4]Performing physical activities and operating machinery in unpredictable environments.
[5]Performing physical activities and operating machinery in predictable environments.
Source: McKinsey Global Institute analysis

- "Without generative AI" projections (dark blue) refer to the 2017 McKinsey study "Jobs lost, jobs gained: Workforce Transitions in a time of Automation," available at mckinsey.com.
- Note that automation not only substitutes human activities in *decision-making and collaboration* but also enhances and complements these activities.
- Further, humanoid robots with embodied multimodal large language models may significantly automate physical tasks, even unpredictable ones.

Source: McKinsey & Company, "The Economic Potential of Generative AI: The Next Productivity Frontier," June 2023, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier.

Change in Workplace Skills Demand

Changes in occupational employment shares among work-age adults, 1980-2016



Source: David H. Autor, "Work of the Past, Work of the Future," *AEA Papers and Proceedings* 109: 1-32, 2019, DOI: 10.1257/pandp.20191110.

Potential Implications for Labor

- As automation has increasingly substituted labor in routine tasks, humans have allocated an increasing share of their working time to decision-making
  - The result is greater scarcity of labor at high skill levels, which are skills acquired through academic education, professional training, and industry experience
  - It has been posited that generative AI may reduce this scarcity by moving decision-making to less highly skilled labor, reallocating talent to decision-making from decision support.[1]
- About 60 percent of jobs in the United States represent new types of work that have been created since 1940[2]
  - Many new occupations have been created by new technologies—an example is the computer programmer
  - Some occupations (for instance in health care) emerge due to consumer demand, induced by rising incomes.

1) See Delphine Strauss, "David Autor: 'We have a real design choice about how we deploy AI'," *Financial Times*, August 10, 2023, https://www.ft.com/content/9c087da3-63d2-4d73-97dc-023025b529aa.

2) See David Autor, Caroline Chin, Anna Salomons, and Bryan Seegmiller (2024) "New Frontiers: The Origins and Content of New Work, 1940–2018," *Quarterly Journal of Economics*, 139(3): 1399–1465, 2024, https://doi.org/10.1093/qje/qjae008.

Implications for Insurance Professionals

- The insurance profession faces a change in skills demand

  "ordered a review of…coursework after…a venture capitalist…visited campus and told students that they wouldn't lose jobs to AI, but rather to professionals who are more skilled in deploying it."[1]

- The insurance professional has an entrepreneurial role in co-invention
  - The professional participates in the feedback cycle on the side of downstream applications
  - It is critical for the professional to understand the (changing) capabilities and limitations of gen AI
  - The benefits of a GPT originates primarily in system change rather than task-level improvement.
- As managers, insurance professionals have a role in change management
  - The arrival of a GPT tends to give rise to automation anxiety
  - There is augmentation of human skills, some of which involves task-level substitution.

1) See Lindsay Ellis, "Business Schools Are Going All In on AI'," *Wall Street Journal*, updated April 3, 2024, https://www.wsj.com/tech/ai/generative-ai-mba-business-school-13199631.

Proprietary Notice

The material contained in this presentation has been prepared solely for informational purposes by Gen Re. The material is based on sources believed to be reliable and/or from proprietary data developed by Gen Re. This information does not constitute legal advice and cannot serve as a substitute for such advice. The content of the presentation is copyrighted. Reproduction or transmission is only permitted with the prior consent of Gen Re.

Gen Re®

Thank you!

Frank Schmid
frank.schmid@genre.com
+1 203 461-1944

genre.com