Transforming OpenAI's ChatGPT into a Circular Business Model

Eirini Ornithopoulou – Design of Circular AI-Based Services (VT 2025)

Introduction

OpenAI is a leading organization in the artificial intelligence (AI) industry, known for developing advanced AI services and research breakthroughs. Its flagship product is ChatGPT, a state-of-the-art conversational AI launched in late 2022 that provides natural language dialogue to millions of users across various domains. ChatGPT is offered as an online service (accessible via web interface and APIs) and has experienced explosive and worldwide user growth – reaching an estimated 100 million users within two months of launch, making it one of the fastest-growing consumer applications in history (Reuters, 2023).

OpenAI's services, including ChatGPT, operate on a software-as-a-service model, where users access AI capabilities hosted on cloud infrastructure. The company's current business model combines free access (to attract a broad user base and gather data) with premium subscriptions (e.g. ChatGPT Plus) and API licensing for enterprise integration, generating revenue through monthly fees and usage-based charges. This model relies heavily on continuous software development and model training, large-scale hardware infrastructure (e.g. data centers), substantial energy consumption, and extensive data handling and processing to deliver AI responses. In essence, ChatGPT's present lifecycle is typical of a linear digital service: it "consumes" resources (computing power, electricity, hardware) to produce AI interactions, and as demand scales up, so do the inputs, with no inherent mechanism to recover or reuse spent resources. While OpenAI's innovations have delivered significant social value through AI capabilities, the current business model faces sustainability challenges. Training and running large AI models like ChatGPT incur significant environmental footprints. For example, the training of GPT-3 (a model underpinning ChatGPT) was estimated to emit around 502 metric

tons of CO_2 (Luccioni et al., 2022), partly due to the use of energy-intensive hardware; each query to ChatGPT also consumes notable electricity (roughly on the order of a few grams of CO_2 per query) (Smartly, 2024).

Moreover, maintaining ChatGPT's service involves frequent hardware upgrades to more powerful processors, contributing to electronic waste – servers and AI accelerators in data centers are often replaced every 2–5 years, and over 80% of this decommissioned equipment is currently discarded rather than recycled (Celion, C., 2024). This linear "take-make-dispose" trajectory is at odds with emerging global priorities for sustainability. As AI adoption grows, studies warn that generative AI could add 1.2–5 million tons of e-waste per year worldwide if no interventions are made (Celion, C., 2024).

In this context, reimagining ChatGPT's business model through circular economy (CE) principles becomes crucial. A circular business model aims to decouple growth from resource consumption and waste, designing operations to narrow resource use, slow the consumption loop, and close the loop by recovering value from by-products. This project develops an advanced plan to transform ChatGPT into a circular business model, integrating CE strategies with AI-driven innovations. It begins by identifying key aspects of ChatGPT's ecosystem that offer opportunities for circular transformation. Then, a detailed circular transformation plan is proposed, selecting appropriate resource strategies – namely narrowing, slowing, and closing resource loops (as defined by Bocken et al. 2022) – and an innovation strategy (open vs. closed innovation) suited to the AI service context. Finally, some anticipated challenges and barriers are discussed – including regulatory, technical, data privacy, and energy-related issues.

Key Aspects of the ChatGPT Ecosystem for Circular Transformation

The ChatGPT ecosystem extends beyond just the AI model – it encompasses how the software is developed and updated, how users engage with the service, the physical infrastructure powering the AI, the energy sources and consumption patterns, and how data is handled throughout.

Software Development & Model Lifecycle

The linear model of software/AI development often involves training increasingly larger models from scratch and discarding older versions. There is an opportunity to make this process more circular by reusing model components, sharing improvements, and extending the useful life of model versions. Currently, much of OpenAI's development is proprietary and internal (a relatively closed innovation approach), which can lead to duplicated efforts across the industry and less reuse of external innovations. Transforming this aspect might include adopting modular design for AI models (so parts can be upgraded without retraining the entire model) and embracing more open innovation or open-source elements to collaboratively improve efficiency. Another idea could be to allow for longer training times, which require less energy expenditure via the newest GPUs, or smaller parallel processing, etc.

User Engagement & Use Patterns

At present, users interact with ChatGPT on-demand, with no direct incentives to minimize resource use – e.g. complex queries are processed just as readily as simple ones, and users may not be aware of the computational cost. There is potential to engage users in ways that support

circularity. For instance, user feedback loops can be strengthened so that each interaction helps improve the system (retaining the knowledge value rather than “wasting” it).

Educating users about the environmental footprint of AI and encouraging efficient usage (for example, batching requests or avoiding redundant queries), or simply providing an easily accessible “green” user guide could narrow resource use. In a circular model, user engagement is not just about satisfaction, but also about creating a community that values and contributes to sustainability goals.

Hardware Infrastructure (Data Centers & Devices)

ChatGPT relies on large-scale cloud computing infrastructure – primarily data centers filled with servers and specialized AI hardware (GPUs/TPUs). The typical hardware lifecycle is linear: manufacture -> use -> disposal in a few years. This infrastructure aspect is perhaps the most tangible in terms of resource consumption (materials and minerals in hardware, water and energy in manufacturing, etc.) and waste generation (e-waste). Transforming this means focusing on longer hardware lifespans (through maintenance, upgrades, second-life uses) and materials recovery at end-of-life. It also means designing or selecting hardware with circular principles (e.g. modular components that can be replaced individually, or servers that are easier to repair).

There is also an ecosystem opportunity: collaborating with suppliers and recyclers to ensure components from retired AI hardware are recovered or reused.

Energy Consumption

Running AI models is energy-intensive – electricity powers the data center servers for both training (which can run for weeks on hundreds of GPUs) and inference (serving millions of queries daily). The current model draws power as needed from the grid or on-site generators, often without synchronization with renewable energy availability. A circular approach would aim to use 100% renewable energy sources and to optimize energy efficiency at every step. This includes both improving the energy efficiency of computation (doing the same AI work with less energy) and better aligning energy use with supply (for instance, scheduling non-urgent processing for times of renewable surplus). Managing energy in a circular way also involves recovering energy where possible – for example, using waste heat from servers to heat buildings. In essence, energy is treated as a looping resource: maximize use of sustainable inputs and recirculate by-products. Integration of AI for energy management (such as smart grids and adaptive cooling) can be a key enabler here.

Data Handling & Knowledge Management

ChatGPT is trained on massive datasets and continually processes user-provided data in conversations. In a linear fashion, data handling often involves collecting new datasets for each training cycle, storing vast amounts of information (some of which may never be used), and then data gets siloed or deleted after use due to privacy concerns. In a circular model, the aim is to “close the loop” on data as well: reuse and repurpose data wherever appropriate instead of constantly starting from scratch. This could include reusing training data and models for multiple purposes (avoiding duplicate efforts across projects), and employing privacy-preserving techniques so that even user interaction data can be fed back into improving the system without compromising confidentiality. Additionally, better data management (like

cleaning, deduplicating, and compressing data) can narrow the data storage and processing needs (an efficiency gain). Essentially, treating data as a resource means extracting maximum learning value from each bit of data collected, and sharing data insights (open data or model outputs) to benefit other systems, rather than discarding information.

Circular Economy Strategies for Transforming ChatGPT

The circular transformation plan for ChatGPT can be built on two strategic dimensions, as suggested in literature: resource strategy and innovation strategy. According to Bocken et al. (2022), a company pursuing a circular business model must decide how it will handle resources – by narrowing resource loops, slowing them, and/or closing them – and how it will approach innovation – through internal efforts or open collaboration (closed vs. open innovation). These choices can be combined to create a holistic circular strategy.

In the case of ChatGPT, an interesting approach is to incorporate all three resource-focused strategies in complementary ways. This combination is justified by the need to address the full lifecycle of the AI service: improving efficiency (narrowing) yields immediate reductions in resource use; extending lifespan (slowing) addresses the mid-term sustainability of infrastructure and models; and enabling reuse/recycling (closing) tackles the end-of-life and residual waste. Meanwhile, an open innovation mindset will bring in external expertise and broad engagement necessary for systemic changes (since AI services operate within a larger ecosystem of partners, users, and regulators). Bocken et al. (2022) define narrowing, slowing, and closing loops as three fundamental mechanisms to make resource use more circular.

Narrowing Resource Loops (Efficiency Improvements)

Narrowing the loop for ChatGPT involves improving the efficiency of resource use so that less energy, fewer materials, and less data are required to deliver the same (or improved) AI service to users. This strategy targets the input side of the system: by reducing consumption of resources per unit of service (per query, per model training, etc.), it directly lowers environmental impact and operating costs. According to Bocken et al. (2022), narrowing strategies focus on optimizations in design and production processes – in this case, design of AI models and the operation of data centers – such that waste and excess are trimmed out.

For ChatGPT, narrowing strategies could include:

- Optimize the ChatGPT model and code to perform computations with fewer operations and less overhead. This can involve techniques like model compression (distillation and quantization) – reducing the size of the neural network and using lower-precision calculations so that each inference uses less energy. OpenAI can leverage AI to achieve some of these gains: for example, using machine learning algorithms to search for more efficient model architectures or to automatically optimize code (an application of AI in software development).
- Apply AI-driven solutions to minimize energy usage in data center operations. A prime example is Google DeepMind's approach, where machine learning was used to optimize data center cooling systems, yielding up to a 40% reduction in cooling energy and 15% overall energy reduction (Vaughan, A., 2016). OpenAI (in partnership with cloud providers like Microsoft Azure) can deploy similar AI-based adaptive cooling and power distribution systems. These AI systems use predictive analytics to adjust

cooling parameters dynamically, anticipating server workload spikes and cooling needs with high precision.

- Reduce the amount of data that needs to be stored and processed by improving data management. This includes deduplicating training data (so we don't waste computation learning the same thing twice) and using synthetic data augmentation wisely to reduce the need for extensive real data collection. AI can assist in data cleaning – for example, intelligent systems can identify and remove low-quality or irrelevant data from the training corpus, which narrows the data needed without sacrificing model performance.
- Another aspect that is probably very effective in narrowing the loop is efficient inference: using AI to predict when a user's query can be answered by a smaller model or a cached result instead of the largest model. For instance, an AI-based dispatcher could route simpler queries to a lightweight model that consumes a fraction of the resources, and only use the big (resource-intensive) model for the complex questions. This tiered approach ensures high resource use (the big model) is only invoked when truly necessary, effectively reducing average resource consumption per query. The expected outcome is fewer servers and less processing time per user request, translating to energy savings and lower wear on hardware.

By implementing narrowing strategies, OpenAI can significantly cut down on the per-interaction resource footprint of ChatGPT. This not only reduces environmental impact (lower carbon emissions, lower water usage for cooling, etc.) but also has business benefits: decreased operating costs for cloud computing and an ability to serve more users or provide cheaper services with the same infrastructure. Efficiency gains also often have a multiplying effect – small percentage improvements scale up massively when you consider billions of queries (so even a 10% energy reduction per query can translate to enormous absolute savings).

Slowing Resource Loops (Extending Product and Asset Longevity)

Slowing the loop focuses on keeping products, components, and materials in use for a longer period, thus reducing the frequency of replacement and new production. Here are some ideas towards that goal:

- OpenAI can coordinate with its cloud providers to implement sustainable hardware lifecycle management. This involves choosing modular, upgradeable server designs where critical components (e.g. memory, drives, accelerator cards) can be replaced or upgraded without scrapping the entire server. Many modern servers support this, but often companies still do full refreshes for performance reasons. By slowing the refresh cycle, hardware is used to its full potential. One concrete practice is to perform predictive maintenance and refurbishment on servers: replacing worn parts (fans, power supplies) and cleaning/recalibrating systems to keep them running efficiently. AI can assist by monitoring signals from equipment (temperature, error rates, performance metrics) to predict failures or degradation (Soriano Sergi, 2025).
- By open-sourcing older model versions (once they are superseded by new ones), OpenAI effectively gives a second life to those models – they might be too outdated for premium service, but could be used by researchers, smaller companies, or in less critical applications (this idea touches on both slowing and closing loops, as it is reuse). The outcome is a longer productive lifespan for the R&D investment made in each model generation.
- From a business perspective, retaining users (and keeping them satisfied so they continue subscriptions rather than leaving) is another "longevity" aspect. While this is more of a marketing outcome, it connects to circularity in the sense that acquiring a

new customer typically has a higher resource and energy cost (advertising, onboarding, initial spike of usage) than maintaining an existing one. Thus, a circular ChatGPT model emphasizes high-quality service and trust to slow down the churn of users.

By slowing resource loops, OpenAI can dramatically reduce the material turnover associated with ChatGPT. If hardware refresh cycles extend from, say, 3 years to 5 or 6 years on average, this means fewer new servers need to be manufactured and installed (saving costs and reducing mining/manufacturing impacts), and less e-waste generated by disposing of old servers. Societally, extending hardware life and donating or selling refurbished equipment can make technology more accessible (old but functional hardware could be donated to educational or under-resourced institutions.

Closing Resource Loops (Recovery and Reuse)

Closing the loop means finding ways to recover value from materials or products after their primary use phase, thus reincorporating them into the production of new services or products (Bocken et al. 2022).

Suggested initiatives for ChatGPT:

- OpenAI should establish a robust program for IT asset disposition. Rather than treating used servers and electronics as waste to be shredded or landfilled (often done out of data security fear), the company can set targets that 100% of retired hardware is either reused or properly recycled. This involves processes like secure data sanitization of storage devices (using software/hardware routines to wipe data) so that servers and

drives can be safely passed on. OpenAI minimizes the waste footprint and can reduce costs by getting value back (selling used equipment or at least avoiding disposal fees). It also hedges against resource scarcity – recovering rare materials like cobalt, nickel, gold from old boards lessens dependency on mining.

- Data centers produce enormous amounts of heat as a waste by-product of running CPUs/GPUs. In a linear system, this heat is simply dissipated (often requiring more energy to cool it away). A circular approach is to treat waste heat as a resource that can be recovered for beneficial use. OpenAI could collaborate with data center operators to implement heat recovery systems – for instance, channeling hot air or water from the cooling system to nearby buildings (offices, residential heating, or industrial processes that need heat).
- Another non-material loop to close is the data and knowledge loop. Currently, user interactions with ChatGPT yield valuable feedback. In a circular model, this feedback shouldn't be wasted – it should be fed back into improving the AI (with user permission and privacy safeguards). OpenAI already does some of this through fine-tuning on user ratings (RLHF), but it can be taken further by creating a data donation program where users agree to let their anonymized chats be used to refine the model. I believe "new data" is generated in conversations with AI, faster than real world data. Therefore, it could be increasingly useful to retain that valuable information and feed it into production again.
- As a creative extension of closing loops, OpenAI could implement a take-back service model for AI models. Suppose in the future various organizations fine-tune their own versions of ChatGPT for specific uses; OpenAI could offer to host or integrate those fine-tunings back into its platform rather than each organization running separate instances. By doing so, the computational resources are pooled (closing the loop on

model usage – one model serving many purposes). This is analogous to product-service systems (PSS) in circular economy. In this case, ChatGPT as a service already means users don't each need to train their own AI (which is resource-saving), but expanding this concept means if a company trains a specialized model, OpenAI could "buy it back" or incorporate it instead of it sitting idle or being discarded when no longer needed by that company.

The closing loops strategy ensures that very little goes to waste at the end-of-life stage. If implemented, the company's contribution to e-waste pollution would be minimal despite operating cutting-edge IT equipment. The aforementioned closed-loop outcomes reinforce a positive image of OpenAI as a responsible innovator and may give it a head start on forthcoming e-waste. Business-wise, material recovery can lower net costs of hardware ownership (some money back at resale or savings from reusing parts). It also protects the company from resource supply risks, contributing to long-term viability. In a broad sense, closing loops contributes to reduced environmental toxicity (less heavy metals in landfills) and lower resource extraction burdens on communities where mining happens.

Challenges and Barriers, and Proposed Mitigations

Implementing a circular economy model for ChatGPT, especially with AI integration, is an ambitious undertaking that will encounter various challenges. It's important to acknowledge these potential barriers and outline strategies to mitigate them:

Regulatory and Compliance Barriers

As of now, there may be regulatory gaps when it comes to enforcing circular practices (e.g. lack of strong e-waste take-back laws in all jurisdictions), but conversely, data privacy regulations (like GDPR) and security standards can restrict some circular initiatives. For instance, reusing user data for improving the model might conflict with privacy laws if not handled properly, and reusing hardware drives could be impeded by data protection rules requiring destruction of data-bearing devices.

To navigate this, OpenAI should implement privacy-by-design and security-by-design in all circular processes. For data loops, techniques such as anonymization, aggregation, and differential privacy can allow learning from user data without exposing personal information. OpenAI can be transparent and obtain informed user consent for any data re-use in model training, aligning with regulations.

Technical and Operational Challenges

Achieving the level of efficiency and lifecycle extension proposed requires overcoming significant technical hurdles. For example, optimizing AI model efficiency without sacrificing performance is a challenge – extremely compressed models might perform worse, affecting user experience. Similarly, predictive maintenance AI must be very accurate to be effective; false predictions could lead to unnecessary downtime or missed failures. Implementing heat recovery or modular upgrades in data centers can be technically complex and costly initially.

OpenAI should treat this as an iterative R&D process. Invest in developing or adopting cutting-edge techniques for model efficiency (e.g. explore algorithmic improvements, neuromorphic hardware, etc., not just compress what exists). They can pilot these improvements on smaller

scale before full deployment (for instance, trial an efficient model for certain tasks and monitor quality). For predictive systems, continuously improve them by feeding them more data – as more maintenance events occur, the AI will get better. Partnering with hardware vendors to design AI-friendly modular systems can help (e.g., working with server manufacturers to ensure that upgrades don't drastically reduce performance).

Financial and Business Model Risks

Some circular investments have high upfront costs (like installing heat exchangers for heat reuse, or developing new AI optimization frameworks) and the returns might be long-term. There's a risk that if not managed, these could strain OpenAI's finances or distract from core business development.

OpenAI can pursue a phased approach and seek partnerships and funding for sustainability initiatives. For instance, there are increasing funds and grants for green technology – OpenAI could obtain sustainability-linked financing to fund data center retrofits or R&D on efficient AI. By tying certain investments to environmental outcomes, it could also benefit from incentives (like tax breaks for renewable energy use, where available). It's important to do a cost-benefit analysis for each major initiative.

Sources

Bocken, N., & Ritala, P. (2022). Six ways to build circular business models. *Journal of Business Strategy, 43*(3), 184–192. https://doi.org/10.1108/JBS-11-2020-0258

Celion, C. (2024, May 8). The hidden environmental cost of data-center growth: Millions of tons of e-waste. *Medium*. https://medium.com/@celions/the-hidden-environmental-cost-of-data-center-growth-millions-of-tons-of-e-waste-0bb4a18dbaa1

Luccioni, A., Viguier, M., Ligozat, A.-L., & Couronne, T. (2022). Estimating the carbon footprint of deep learning training: A case study of GPT-3. *Patterns, 3*(12), 100676. https://doi.org/10.1016/j.patter.2022.100676

Morseletto, P. (2020). Targets for a circular economy. *Resources, Conservation & Recycling, 153*, 104553. https://doi.org/10.1016/j.resconrec.2019.104553

Reuters. (2023, February 1). ChatGPT sets record for fastest-growing user base. *Reuters*. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Smartly.ai. (2024, June 7). What is the CO_2 emission per ChatGPT query? *Smartly.ai Blog*. https://smartly.ai/blog/the-carbon-footprint-of-chatgpt-how-much-co2-does-a-query-generate

Soriano Sergi, A. (2025, April 22). Tackling data center efficiency: Software in operations & maintenance. *Energize Capital*. https://www.energizecap.com/news-insights/tackling-data-center-efficiency-software-in-operations-maintenance

Vaughan, A. (2016, July 20). Google uses AI to cut data-centre energy use by 15%. *The Guardian*. https://www.theguardian.com/environment/2016/jul/20/google-ai-cut-data-centre-energy-use-15-per-cent