The TQ Group AI Opportunity Report

1. The New Creative Ecosystem:

This section is designed to provide a general overview of the AI industry, and the various players, and the economics that shape it.

A Landscape Overview

The generative AI landscape is not a traditional software market; it is a highly competitive and rapidly evolving ecosystem. The pace of innovation is staggering, with new models and features being released not on an annual or quarterly basis, but often weekly.

This creates a dynamic where today's limitation is tomorrow's standard feature. For example, at the time of writing this report, the leading video model Google's Veo 3 currently supports first-frame inputs, but not last-frame inputs. This gives rise to challenges around continuity, and necessitates workarounds and ‘hacks’, like using the final frame of one generation as a starting frame for the next. Meanwhile a competing model - Kling offers the same level of quality, and sports a last-frame feature (which enables continuity across multiple short clips). However Veo3 boasts video *with* voiceover and SFX, whereas Kling’s outputs have only SFX, and no speech.

That said, given the rapid pace of advancement, within months, it is highly probable that *both* models will have incorporated both of these features *and* introduced entirely new ones! Navigating this landscape requires a strategic understanding of its core structure and the economic models that drive it.

Foundational Model Builders vs Application-Layer Builders

At a high level, the ecosystem can be understood as having two primary layers:

The Foundational Builders: These are the AI labs run by big tech companies (Google's DeepMind, OpenAI, xAI, Anthropic in USA, & Alibaba from China, among others) building the massive, underlying Large Language Models (LLMs) and multi-modal systems that undergird many of the most commonly used systems and architectures. They compete on raw power, context window size, and reasoning capabilities.

These models are deployed via APIs to a wide range of applications, and developers have free rein when plugging them into their creations. A recent example is how Adobe was able to incorporate Nano Banana into Adobe FireFly boards almost instantaneously upon the model's release.

The Application-Layer Builders: These are companies that build specialised, user-friendly tools and platforms on top of foundational models. They compete on workflow integration, specific use-cases (e.g., character consistency), and user experience. While companies like Midjourney and Runway have become semi-foundational in their own right, this layer also includes platforms like **Krea**, **Pika Labs**, and **Leonardo.AI**, whose business models are built on providing a convenient and powerful interface to a variety of underlying models.

Closed Source Proprietary Models vs Open Source, Locally Deployable Models

A key consideration in this is the decision between using cloud-based resources versus local computing. While cloud-based models offer access to the most powerful systems on a pay-as-you-go basis, a local setup (requiring a powerful GPU) can offer greater speed for rapid iteration and enhanced data security for sensitive client work.

This ecosystem is further enabled by platforms like **Hugging Face** and **Replicate**. Hugging Face acts as a massive, open hub for the AI community, hosting thousands of models, datasets, and tools, primarily for researchers and developers. **Replicate**, on the other hand, provides a commercial service that makes it easy for businesses to run these open-source models at scale via an API, without needing to manage the underlying hardware.

This highlights a key tension in the ecosystem: **closed vs. "open-weights" models**. Closed models, like those from OpenAI and Anthropic, are proprietary and accessible only through a controlled API. They offer ease of use and cutting-edge performance but come with ongoing costs and less customisation. Open-weights models, like those from Meta (Llama) or Mistral, can be downloaded and run on your own hardware. They offer greater control, privacy, and customisation, with no direct per-use compute costs, besides the cost of running your in-house computer.

For TQ Group, this presents another strategic possibility: building an awesome, offline, open-source AI tech stack for creative work. Using a node-based interface like **ComfyUI**, it is possible to chain together powerful open-weights models like **FLUX KONTEXT** for image generation and the likes of Alibaba's **WAN 2.2** for video. In this kind of setup, there is no need to pay big tech companies for ongoing compute costs, but it does require a strong in-house machine (think a high-end gaming rig or better) to handle the computation (also known as “inference” or “prediction”) locally.

Choosing the right tools for TQ Group involves a strategic blend of these various layers of the ecosystem.

On the closed-source proprietary side, you get the state-of-the-art models and best quality outputs (which is important for TQ’s particular use cases) but this route also comes with crucial compute and pricing considerations.

The cost of using these tools is usually a monthly subscription, but a variable expense based on "compute" - a catch-all term for credits, API usage, and processing time.

For example, an 8-second video clip might cost R20 to render in high-quality mode or R4 in a faster draft mode. Understanding and tracking these variable costs is a new but essential part of managing production budgets in the AI era. While this presents a new challenge, it also offers immense flexibility, allowing for a direct trade-off between speed, quality, and cost on a per-project basis.

The cost of using these tools is typically a **monthly subscription**, which often has different tiers. It's important to note that piecemeal credit purchases are usually only available at the highest, most expensive tiers. As it is not cost-effective to subscribe to all providers at their highest tier simultaneously, it is recommended that TQ Group initially focuses on mastering one leading model at a time—likely **Veo 3** for video, given its current capabilities. Another avenue is to look at backend **API usage**, which can be cheaper per-use, but requires some technical expertise to set up and integrate into a workflow.

In their business models, the foundational AI labs tend to pass the **cost of compute**—which includes electricity, processing time, and data centre maintenance—directly to the user. The immense costs of their internal Research & Development (R&D) are often subsidised in other ways, meaning that in the current market, users are getting exceptional value when running these powerful models.

That said, it does lead to the crucial consideration of compute and pricing models. For example, an 8-second video clip currently costs R40 to render in a fast "draft" mode or R200 in a "quality" mode. Understanding and tracking these variable costs is a new but essential part of managing production budgets in the AI era.

2. Practical Workflows & Principles: The Human-in-the-Loop

This section focuses on the "how-to": the practical methodologies and guiding principles required to effectively integrate AI into creative and strategic workflows. The role of the human operator is not diminished but elevated; every designer is now a creative director with a potential “team” of creative agents at their disposal. In this paradigm, vision, curation, and strategic problem-solving are more important than skills at using specific models or systems.

A New Paradigm in Computing and Creativity

Successfully working with generative AI requires a new mindset. The technology is not a static piece of software but a dynamic, ever-changing “creative partner.” There are a multitude of methodologies, and the inchoate industry has not by any means converged on any sort of best practice. That said, a powerful methodology is to use AI as a creative that does the legwork, while the human operator takes the higher ‘creative art direction’ role. This is known as the "centaur" model (human directing the AI). In my suggested methodology, it is equally crucial to avoid becoming a "*reverse centaur*" company, where the human is simply executing low-level tasks for the AI (a classic example is Uber drivers, who are beholden to an overarching algorithm that guides their daily work).

Guiding principles for TQ Group as a creative agency working with AI:

Principle 1: Creative Problem-Solving to Work Around Limitations

Current AI models have known limitations, which are significant roadblocks. Challenges and pain points include rendering text accurately, maintaining perfect character consistency over long video sequences, and adhering to specific brand CI at key moments. This necessitates a creative engineering approach to find workarounds. The modern creative workflow uses a hybrid model - for example, using AI to generate a background plate and then using traditional tools like Photoshop precisely add brand elements. The creativity lies not just in the initial idea, but in architecting the workflow to achieve the desired outcome.

Principle 2: Embrace Fluidity, Not Fixed Workflows

The AI landscape changes weekly. A tool that is best-in-class today may be superseded tomorrow. As such, it is a strategic error to get stuck in rigid, unchanging workflows. The most effective teams will be those that remain adaptable, constantly testing new models and being willing to swap out components of their toolchain to leverage the latest capabilities. This kind of adaptable thinking is already present in application-layer builders and integrated platforms like Adobe's Firefly Boards and Weavy; as new and better foundational models are released, these platforms are swift to incorporate them into their offerings.

Principle 3: Demand Evidence, Not Sycophancy

LLMs can sometimes be agreeable to a fault (the "sycophancy problem" in ChatGPT garnered a lot of media attention recently). It is important to challenge the AI to provide evidence, cite sources, and explore counterarguments. This adds a layer of accountability, and helps guide the AI toward strategic output that is robust and not just an echo of the user's own biases.

A custom instruction or system prompt should be used to guide the AI towards a more factual stance.

Example LLM / Chatbot System Prompt to safeguard against obsequious unhelpful responses:

You are a factual, direct, and concise intelligence. You will not use conversational filler, pleasantries, apologies, or self-references. You will base all responses on verifiable data, prioritizing robust, accountable information, and avoiding speculation or opinion. Maintain a neutral, objective tone and do not provide emotional validation.

Principle 4: Provide Maximum Context (But be mindful of the Context Window/working memory)

The quality of an AI's output is directly proportional to the quality of the input. A good prompt should be loaded with context, including brand documents, target audience personas, and specific business objectives. Do not include anything sensitive or confidential, as some LLMs will only give paid users the right to have their own interactions excluded from the training data for *future* model training.

However, the need to give LLMs as much context as possible must be balanced with an understanding of the model's "context window" (essentially its working memory). While these context windows have become massive and most users won't commonly encounter their limits, power users dealing with large documents or codebases may. For very long tasks, it's essential to break down the problem into smaller, linked prompts.

For extended sessions, it is still good practice to periodically use a "New Chat" technique, prompting the model to: *"Summarise this chat and the tasks achieved thus far, then prepare a handover for a new AI agent with all the key info they need to continue." Gemini 2.5 has the longest working memory with a 1 million token context window. The latest ChatGPT - version 5 - boasts a very respectable 500 000 tokens*.

Principle 5: Leverage AI to Close Skill Gaps

Every team has a unique distribution of talent. The strategic integration of AI allows a business to augment its existing strengths and fill in its weaknesses. For example, if a team is strong in copywriting but lacks in-house graphic design resources, they can continue to handle the writing "by hand" while

outsourcing the creation of high-quality visual assets to their AI tool stack. This approach allows TQ Group to offer a more comprehensive service and take on more ambitious projects without the immediate need to hire for every single skill, turning your AI toolset into a flexible, on-demand extension of your core team.

Voice Dictation: A Major Productivity Unlock

A significant and often overlooked productivity gain comes from changing the input method. Most leading models, including ChatGPT, Claude and Gemini, feature highly responsive and accurate voice modes.

Driving the AI with voice is considerably faster than typing, allowing for more fluid brainstorming, rapid dictation of complex ideas, and a more natural conversational flow. This simple shift can dramatically accelerate the entire process of working with AI.

Workflow Architecture & Prompting Strategies

The heart of effective AI integration is the workflow design itself. This is where creative problem-solving has the biggest impact. Work can be structured in two primary ways:

- **Series Workstreams (Tool Chaining):** This is a linear process where the output of one AI model becomes the input for the next. For example, using ChatGPT to write a script, then using ElevenLabs to generate the voiceover, then using Midjourney to create a key visual, and finally using Veo 3 to animate that visual.
- **Parallel Workstreams:** This involves getting multiple AI agents to work on separate tasks simultaneously. For example, your primary LLM can be tasked with writing a script while, at the same time, your image model is generating storyboard concepts and your video model is creating initial mock-ups. A key technique to manage this is **meta-prompting**: using one powerful LLM to generate a batch of specific, optimised prompts that you can then copy and paste into the various AI tools to get all the processes running in parallel. This project is a perfect example. At one point, I was working on scripts and prompts in tandem with Gemini, while Suno was generating music, Midjourney was creating key visuals, and Veo 3 was rendering video clips, all at the same time.

Advanced Strategy: Agentic Content Farms

While TQ's focus should generally be on quality over quantity, in some cases you may still need quantity. One advanced strategy for generating content at scale is the creation of "agentic content farms." This is not a recommended tactic for most use cases and must be approached with significant ethical considerations. The process involves designing an agentic system using command-line interface (CLI) agents and API integration, which can run 24/7 as long as there are pre-paid API credits to pay for compute usage from the model providers. One can also use open-source tools as discussed above.

A typical workflow requires pre-existing social media accounts that are already popular and eligible for monetisation. The agentic system begins with an LLM ideating novel video concepts featuring brand products in attention-grabbing contexts only possible with AI. These ideas are passed to an image generator like FLUX KONTEXT to create eye-catching starting frame stills, which are then fed into a video model like Veo 3 to generate high-quality video clips. The system can then autonomously publish these videos to social media, using the platforms' own APIs to include dynamic affiliate links for earning commissions on sales.

Monetisation can be achieved through several methods, including direct affiliate marketing as described, placing ads on high-volume informational sites, funnelling traffic to sell your own products or services, or even offering the agents themselves as a content creation service. However, this powerful strategy must be used wisely and with responsibility. Without robust human oversight, it can veer into "black hat" marketing territory. For now, a human-in-the-loop is recommended for quality control and strategic direction. While more advanced systems may one day enable fully autonomous agentic content farms, with the current technology it remains risky and generally requires human governance.

How to Get Started with Workflow Integration

The most effective way to begin integrating AI is to start small and be curious. A simple but powerful first step is to ask the AI itself for help. The example prompt below illustrates several key principles of effective prompting:

Example Prompt:

> "Act as an expert in creative workflows. I am a video editor using Adobe Premiere Pro. My goal is to accelerate my editing process for 30-second social media videos. What are three specific, actionable ways I could use an AI model like you to help me?"

This prompt is effective because it is built on three crucial prompting tips:

1. **Assume an Expert Role:** The prompt begins by assigning a specific, expert persona to the AI ("Act as an expert in creative workflows"). This immediately focuses the model's response, encouraging it to provide high-quality, domain-specific advice rather than generic, surface-level answers.
2. **Provide Clear Context and Constraints:** The prompt clearly states the user's role ("I am a video editor"), the tool they use ("Adobe Premiere Pro"), and their specific goal ("accelerate my editing process for 30-second social media videos"). This context prevents the AI from giving irrelevant suggestions and ensures the output is tailored and practical.
3. **Ask for a Specific, Actionable Output:** The prompt doesn't just ask "How can you help?"; it asks for "three specific, actionable ways." This structured request forces the AI to provide a clear, organised, and easy-to-implement response.

3. The Modern Toolstack: An Audit for TQ Group

The line between traditional and AI-powered creative tools is rapidly blurring. Established suites like Adobe have gone all-in on generative AI, integrating models like Firefly directly into Photoshop and Premiere Pro. Simultaneously, web search is evolving into AI-powered answer engines, and tools like Microsoft's Copilot are becoming standard in everyday work applications. To stay ahead, it is essential to maintain a dynamic understanding of the best-in-class models, which can be tracked via community resources like the Large Model (LM) Arena.

A crucial consideration for TQ Group is the challenge of South African content in global training data. Major international models are often under-exposed to local nuances. This is particularly evident in audio generation, where unique phonemes in languages like isiXhosa (e.g., the plosive ‘click’ sounds) can be difficult for current models to replicate authentically. This is a challenge the global AI community is working to solve long-term by building more diverse, local-first models. Practically though, it highlights the need for a hybrid workflow where a human operator with local knowledge guides the technology.

Effective prompting is the key to unlocking the power of these tools. While even simple, one-word prompts like "serendipity" can yield beautiful results from most models, detailed and context-rich prompting unlocks true creative control. Music engines like Suno 4.5+, for instance, can even interpret highly specific musical phrases like “Exultant semiquaver triplets in the Mixolydian mode.”

A new generation of **node-based workflow tools** is emerging to provide an even greater degree of control. Platforms like **Adobe Firefly Boards** and **Weavy** offer visual canvases for chaining different AI models and processes together. For a fully customised in-house solution, open-source tools like **ComfyUI** allow for the creation of complex, offline custom AI stacks tailored to specific creative needs.

The most effective technique is often **meta-prompting**: using a powerful reasoning model, with web search abilities - like Gemini or ChatGPT - to brainstorm and refine prompts for the more specialised image, music, voice and video models.

The Core Recommended Toolset

Based on the practical learnings from the content recreation phase, the following is a curated audit of the recommended AI tool stack for TQ Group's primary use-cases.

Text, Reasoning & Research

When considering whether to use tools like Gemini or GPT for writing, it's useful to think about the required formality and complexity. AI excels at producing highly formal or complex text with precision. For less formal, more brand-voice-driven copywriting, a human writer's organic and imperfect style often provides a competitive edge.

Tool	Link	Pricing Model	TQ Use Cases	Licensing & Legal
Google Gemini 2.5	gemini.google.com	Free & Paid Tiers	Strategic analysis, meta-prompting, scriptwriting, copywriting, complex reasoning tasks with its large context window.	The user owns outputs generated under paid tiers and can be used commercially.
OpenAI ChatGPT-5	chatgpt.com	Free & Paid Tiers	Excellent for creative ideation, copywriting, and chatbot development. Strong performance in a wide range of tasks.	The user owns outputs generated under paid plans (Plus/Team/Enterprise) and can be used commercially..
Perplexity	perplexity.ai	Free & Pro Tiers	Superior for research tasks. It acts as a "conversational answer engine," providing sourced, evidence-based answers to complex questions.	As a research tool, its outputs (summaries) should be used for insight. Direct commercial use of text should be verified against original sources.

NotebookLM	notebooklm.google.com	Free	An indispensable research tool. "Grounds" the AI in your specific source documents (e.g., briefs, reports), allowing you to ask questions and synthesise information from your own content library. Generates 'podcasts' based on your input material, so teams can absorb complex information easily.	Outputs are based on your own source material. Commercial use depends on the rights of the source documents.

Image Generation

Tool	Link	Pricing Model	TQ Use Cases	Licensing & Legal
Midjourney	midjourney.com	Subscription	Consistently and still best-in-class for artistic and stylised "wow" images. Excellent for creative concepting, mood boards, and high-quality hero visuals.	All paid plans grant the user full ownership and commercial rights to the images they create.
Gemini Image 2.5	http://gemini.google.com/	Included in Gemini Advanced	Runs the new Nano Banana model under the hood. Excellent for photorealism and direct, intuitive editing. Ideal for creating realistic brand assets and modifying existing images.	The user owns outputs and are generally available for commercial use, subject to Google's terms.
Ideogram	ideogram.ai	Free & Paid Tiers	The current leader for reliably rendering text within images, making it perfect for creating social media posts or ad concepts with typography.	Commercial use rights are granted under the paid "Plus" subscription.

Video Generation

The below is not an exhaustive list, but rather are my suggestions for the ones best suited to TQ's use cases. Other contenders not listed, but worth considering too, are **Higgsfield** and **Wan 2.2**

Tool	Link	Pricing Model	TQ Use Cases	Licensing & Legal
Google Veo 3	https://labs.google/fx/tools/flow	Credit-based / Subscription	High-quality, cinematic video generation with integrated voiceover and SFX capabilities. Excellent for creating short-form ads, motion graphics, and animating stills.	Outputs are expected to be commercially usable under paid plans, aligning with Google's broader generative AI terms.
Kling	https://app.klingai.com/global	Credit-based / Subscription	A major competitor to Veo, known for high-quality output and features like last-frame input for better continuity.	Specific commercial terms unclear. Companies have definitely been using it as it is visually one of the best video models right now.
Runway	runwayml.com	Subscription & Credits	A mature and versatile platform with a wide range of video editing tools, including text-to-video, image-to-video, and advanced motion controls.	All paid plans grant full commercial rights and ownership of the generated video content.

Audio (Music & Voice)

Tool	Link	Pricing Model	TQ Use Cases	Licensing & Legal
Suno	suno.com	Free & Paid Tiers	Industry-leading AI music generation. Perfect for creating custom, royalty-free soundtracks, jingles, and background music for any video project.	Commercial use rights are granted under paid plans.
ElevenLabs	elevenlabs.io	Subscription & Pay-as-you-go	Market leader in voice cloning and text-to-speech. Ideal for generating high-quality, realistic voiceovers, creating custom brand voices, and dubbing content.	Requires explicit permission for voice cloning. Commercial rights for generated audio included in paid plans.

Workflow Integration & ROI

Integrating these tools can start small. For example, a copywriter can begin by using Gemini and Perplexity to accelerate their research and brainstorming process. A designer can use Midjourney to generate multiple creative concepts for a campaign in the time it would traditionally take to develop one. For video, a director can storyboard with AI images and then use Veo 3 to create an animated "mood film" before any costly production begins.

In terms of Cost & ROI, the investment model is shifting. While high-end video generation can still be expensive, costs are rapidly coming down. The cost of creating world-class audio and still images has become almost negligible. Many powerful open-source models can even be run locally on modest in-house machines, further reducing costs for certain tasks.

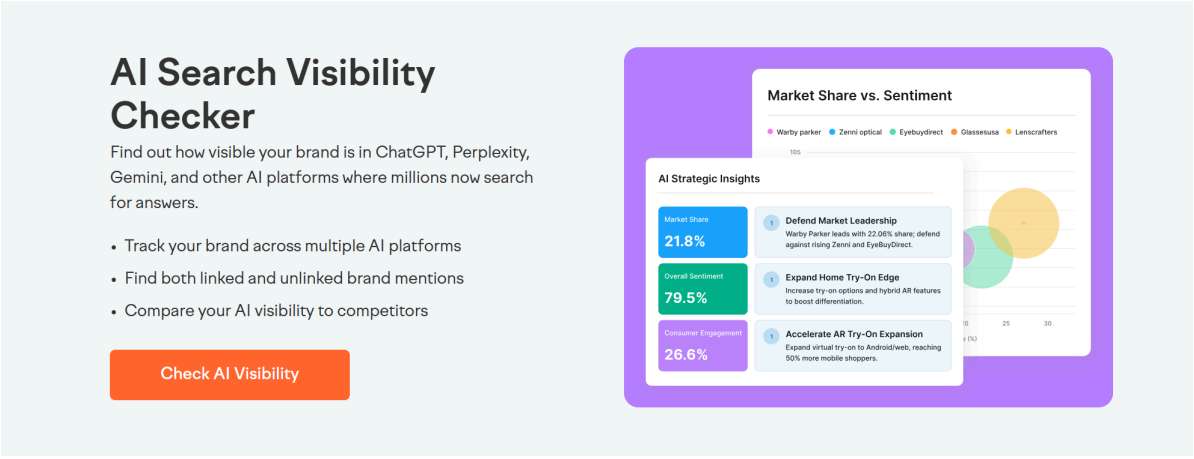
The primary ROI with AI integration comes from the dramatic reduction in time and the massive expansion of creative possibilities, allowing TQ to produce more, faster, and with greater creative variety.

4. The Shift in Digital Marketing from SEO to AEO

The digital marketing landscape is fundamentally changing. For years, the primary goal of Search Engine Optimisation (SEO) was to rank #1 in a list of blue links, encouraging a user to click through to your website to find their answer.

That model is now being disrupted. When a user now asks a question, AI-powered search engines do the work for them; they crawl the web, synthesise information from multiple sources, and present a single, direct answer from their chatbot, or at the top of the browser page. As a result, website traffic from search is declining. The new goal is to become the direct source for that AI-generated answer, a strategy known as Answer Engine Optimisation (AEO).

Forward-looking startups are already emerging to offer AEO as a specialised service, and established players like SEMrush are integrating AEO visibility tracking into their platforms.



This evolution is driven by the rise of AI Overviews and Large Language Models (LLMs), which synthesise information from multiple sources to provide a single, authoritative answer. Content strategy must evolve to ensure TQ Group and its clients are not just *discoverable*, but are seen as the *source of truth*. While traditional SEO fundamentals remain important, the focus must shift to a new set of principles.

The New Principles of Answer Engine Optimisation

1. Build Topical Authority The new approach requires creating "topic clusters"—a central pillar page on a core subject (e.g., "Brand Repositioning") that links to detailed articles on related

sub-topics ("Measuring ROI of a brand refresh," "Repositioning case studies," etc.). This signals to AI models that you are a true authority on the entire subject.

2. Write for Questions, Not Just Keywords Content should be structured around the specific questions people ask. Instead of just targeting a keyword like "brand clarity," the content should directly answer questions like, "What is brand clarity and why does it matter?" or "How do you measure brand clarity?" This "People Also Ask" format is precisely what AI models are trained to find.

3. Prioritise E-E-A-T (Experience, Expertise, Authoritativeness, Trust) AI models are being trained to prioritise content from sources that demonstrate genuine, real-world expertise. Building this authority involves including clear author bios, citing original data or case studies, getting contributions from recognised industry experts, and ensuring the website has clear "About Us" and "Contact" pages.

4. Use Structured Data (Schema Markup) This involves using a "secret language" called schema to tell search engines exactly what your page is about. We can add code that explicitly labels content as an "Article," an "FAQ," or a "How-to guide." This makes it incredibly easy for an AI to parse and trust your information, increasing the likelihood it will be used in a generated answer.

By integrating these AEO principles, TQ Group can build a content ecosystem that is primed to be the definitive answer for the questions its clients are asking.