GOOGLE AI IN 2026: THE SEAT SIDE, THE CONSUMPTION SIDE, AND THE COMMIT THAT JOINS THEM

Two things every Google customer needs to understand this year: how Gemini is sold per user through Workspace and Gemini Enterprise, and how agents, avatars and media are sold per unit through Vertex AI and the Google Cloud commitment

> *A research note for CIOs, CISOs, IT asset managers, digital workplace owners and procurement leaders with a Workspace renewal, a Gemini Enterprise proposal, a Vertex AI agent program or a Google Cloud commitment on the table in the next twelve months. It is not tied to any single client situation. Written from the buyer's side of the table. Redress Compliance, 2026.*

Key takeaways

- Google sells AI on two sides of the same house. On the seat side, Gemini is folded into Workspace editions and sold again as Gemini Enterprise, per user, whole population, with consumption overage on top. On the consumption side, Vertex AI, agents, Veo, Imagen and the avatar and video presenter features bill per token, per second and per generation, inside a Google Cloud commitment.
- The two sides are negotiated by different teams on the customer side and by one account team on Google's side. Google sizes the commit from both. The buyer that negotiates Workspace in one room and Cloud in another funds the same floor twice.
- Seat side: the standalone Gemini add ons are gone, the base price rose to absorb them, and Gemini Enterprise adds a second seat with an included quota that heavy users exhaust in days. The trap is paying for every user and leaving the overage open.
- Consumption side: introductory token rates on the newest models reset upward on 1 January 2027, media and avatar generation has no forecast, and Gemini Enterprise commit plans carry a symmetric downside. The trap is committing to Google's forecast on a rate that is about to change.
- The response is one exercise across both sides: a bottom up forecast in Google's own meters, a pilot that decides who gets a seat, ceilings on every line that cannot be forecast, unit rates fixed in the order form past the January reset, and a single ramped commit sized to your number rather than Google's.

PART ONE: THE SEAT SIDE, WORKSPACE AND GEMINI ENTERPRISE

1. HOW GEMINI IS SOLD PER USER

- **Workspace with Gemini included.** The Gemini Business and Enterprise add ons were retired in early 2025 and the AI was folded into the Workspace editions, with Business Standard and Plus and Enterprise Standard and Plus repriced upward to carry it. Every seat on the plan pays for Gemini whether the user opens it or not. The uplift arrived as a price change on the base edition, not as a line the buyer could decline.
- **AI Expanded Access.** A per user add on that raises usage limits and opens video and image generation, deep research and the higher models to users who exhaust the included quota. It is the first place the seat side starts to behave like consumption.
- **Gemini Enterprise.** The agent platform, sold as a second seat in Business, Standard, Plus and Frontline editions with annual commitment pricing below month to month. The seat includes a quota of agent actions; beyond it, consumption

is billed against a credit pool. Google positions this as the enterprise front door for search across company data, prebuilt agents and custom agents built by business users.

- **Gemini Code Assist.** Per developer per month, bundled with a Google Cloud subscription, with API calls above the included allowance billed through Vertex.
- **Plan mechanics.** Workspace runs on Annual or Flexible plans. Annual plans commit the seat count for the year with additions allowed and reductions blocked until renewal. Flexible plans allow month to month changes at a higher unit price. Google proposes Annual for the whole population and prices the flexibility out of reach.
- **The compliance layer.** Workspace licenses are enforced technically, so there is no audit in the Oracle sense. The compliance motion is the true up: users provisioned above the committed count are billed at the then current rate, and unused seats are not credited. Directory sprawl, shared mailboxes and service accounts on full editions are the recurring findings.

2. WHERE THE SEAT SIDE MONEY SITS

Line	Google's opening structure	The trap	Buyer side structure
Workspace edition	Enterprise Plus for all, Annual plan, list uplift accepted	AI uplift paid on every seat including light users and shared accounts	Edition mix by role; light users on Business Standard or Frontline; base rate protected for term
AI Expanded Access	Proposed for the whole population	Add on for users who never hit the included limit	Only for users with measured usage above the quota; right to remove at anniversary
Gemini Enterprise seats	All knowledge workers, Standard or Plus, year one	Sixty percent adoption pays for one hundred percent of seats	Pilot population first; seats staged to adoption thresholds; Frontline for light use; additions at the same net rate
Seat consumption quota	Included quota per seat, overage at card rate	Heavy users exhaust the quota; overage uncapped	Quota pooled across the tenant; overage rate fixed; monthly ceiling
True up	Provisioned seats billed at then current rate	Directory sprawl billed as users	Quarterly reconciliation of provisioned to active; true up at the committed rate

3. HOW TO RESPOND ON THE SEAT SIDE

- **1. Reconcile provisioned to active.** Pull the admin console reports and separate real users from shared, service and dormant accounts before any count is committed.
- **2. Map editions to roles.** Decide who needs Enterprise Plus, who needs Business Standard and who is a Frontline worker. The edition mix is the largest seat side lever and Google will not propose it.
- **3. Run a Gemini Enterprise pilot with a threshold.** Ninety days, a defined population, an adoption measure that decides who gets a seat. Stage the rollout to that measure and get the staging schedule and the net rate for additions into the order form.
- **4. Pool and cap the quota.** Consumption above the included quota is pooled across the tenant, the overage rate is fixed, and a monthly ceiling with an approval gate applies.
- **5. Protect the base rate.** The Workspace edition price is held for the term, with any future AI uplift excluded for seats already committed.
- **6. Keep true down.** A right to reduce seats by a stated percentage at each anniversary, on Annual plans, without losing the discount tier.

PART TWO: THE CONSUMPTION SIDE, VERTEX AI, AGENTS, AVATARS AND THE COMMIT

4. HOW AGENTS, AVATARS AND MEDIA ARE SOLD PER UNIT

- **Vertex AI and the Gemini API.** Per million tokens, input and output priced separately, reasoning tokens billed as output. The newest Flash generation carries introductory rates until 31 December 2026 and roughly doubles from 1 January 2027. Context caching, batch processing and provisioned throughput each carry their own discount or premium.
- **Agents.** Prebuilt agents inside Gemini Enterprise draw on the seat quota and then on credits. Custom agents built with the agent development kit or Agent Builder bill as tokens, tool calls, grounding requests and, where they run continuously, compute. A production agent that runs several reasoning passes per task bills the output rate on every pass.
- **Avatars and video presenters.** AI presenter and avatar features in Vids and the Gemini app, and video generation through Veo on Vertex, are metered per second of output. A minute of generated video costs more than a day of text for a typical user. Marketing, learning and communications teams adopt these fastest and forecast them least.
- **Image, music and speech.** Imagen, Lyria and text to speech bill per image, per generation or per character. Small lines individually, unforecastable collectively.
- **Gemini Enterprise commit plans.** Since August 2026 Gemini Enterprise consumption can be bought pay as you go, on a Flexible Savings Plan, or on a committed plan with deferred execution. The plans discount the rate in exchange for a floor. The downside is symmetric: a plan sized to Google's forecast that covers less usage than expected costs more than paying as you go.
- **The Google Cloud commitment.** All of the above sits inside a multi year committed spend agreement with tiered discounting, resource based and flexible committed use discounts on compute, GPUs and TPUs, and Marketplace drawdown subject to caps. Since June 2026 the default CUD scope is the billing account rather than the project, which changes how utilization is measured and how shortfall is calculated.

5. WHERE THE CONSUMPTION SIDE MONEY SITS

The illustration below is a mid sized enterprise with roughly 5,000 knowledge workers, a Gemini Enterprise rollout, two production agents and a media program, over three years at approximate list. Figures are for structure, not quotation.

Line	Google's opening structure	Approx. three year value	Buyer side structure	Approx. three year value
Vertex AI tokens	Forecast on introductory rates; standard rates from January 2027	USD 3.6M	Rates held for term in the order form; volume tier from month one; batch and cached rates specified	USD 2.1M
Production agents	Estimated at demo loop; tool calls and grounding omitted	USD 1.4M or open	Forecast on the production loop; tool calls and grounding priced; monthly ceiling	USD 0.9M
Avatars, video, image	Open consumption	Unforecast	One media ceiling with alerts and approval gate; unit rates in the order form	USD 0.4M
Gemini Enterprise commit plan	Three year committed plan sized to Google forecast	USD 9.0M against 6.7M of usage	Flexible Savings Plan at 80 percent of buyer forecast; remainder pay as you go	USD 6.9M
Google Cloud commit	USD 36M flat annual floor	USD 36M	USD 27M ramped, shortfall rollover, CUD scope confirmed, Marketplace drawdown to cap	USD 27M

Line	Google's opening structure	Approx. three year value	Buyer side structure	Approx. three year value
Total		**USD 50.0M plus open lines**		**USD 37.3M, all lines capped**

- **The commit is the money.** Every token and seat discount is small next to the difference between a flat floor sized to Google's model and a ramped floor sized to yours.
- **The January reset is the largest silent change.** Usage that does not move at all costs twice as much in year two unless the rate is in the paper.
- **The plan can cost more than paying as you go.** A committed plan that covers less usage than forecast is the one line in the proposal where the discount is negative.
- **Open lines are the surprise.** Media, avatars and agent tool calls are individually modest and collectively the difference between a budget and an invoice.

6. HOW TO RESPOND ON THE CONSUMPTION SIDE

- **1. Forecast in Google's meters.** Tokens by model, actions, seconds of video, images, tool calls and grounding requests, from your own billing exports and pilot telemetry, for three years. Compare with the account team's forecast and expect a large gap.
- **2. Establish the alternative.** A second model provider running a real workload through your other hyperscaler or directly. It is the only leverage Google respects on tokens and the only thing that makes a rate hold negotiable.
- **3. Fix the unit rates.** Token rates held past 31 December 2026, successor models at equivalent tier priced no higher, ninety days notice on deprecation with a migration credit, media and agent unit rates in the order form.
- **4. Cap what you cannot forecast.** Monthly ceilings on media, avatars and agent consumption, with alert thresholds and an approval gate. A ceiling you never reach costs nothing.
- **5. Size any plan below the forecast.** Flexible Savings Plan or committed plan at no more than 80 percent of your own number, with the balance pay as you go, and the symmetric downside modeled before signing.
- **6. Size the commit from both sides together.** One ramped floor covering Workspace, Gemini Enterprise, Vertex and infrastructure, from your combined forecast, with shortfall rollover, coterminous end dates, CUD scope confirmed in writing and Marketplace drawdown to the permitted cap.
- **7. Secure the terms that decide year two.** No training on customer data, zero data retention for prompts and outputs where available, the generative AI indemnity confirmed for the products in scope, data residency by region, support as a fixed amount rather than a percentage of the commit.
- **8. Sign on your calendar.** Google's quarters close in March, June, September and December. The proposal, the credits and the executive escalation cluster there. Set the timeline from your renewal date and hold it.

7. MISTAKES THAT COST THE MOST

- Accepting the Workspace AI uplift on every seat, including shared, service and dormant accounts, because it arrived as a base price change.
- Signing Gemini Enterprise for the whole workforce on a demo, before a pilot has decided who uses it.
- Building the three year token budget on introductory rates and discovering the standard rates on the first January invoice.
- Leaving avatar, video and agent consumption uncapped because nobody could forecast them.
- Committing to a Gemini Enterprise plan sized to Google's forecast and paying more than pay as you go would have cost.

- Negotiating Workspace and Google Cloud in separate rooms with separate teams while one Google account team sizes a single commit from both.
- Treating the Google proposal as the only proposal. Without a second provider on a real workload, every discount is a favor.

8. HOW REDRESS HELPS

Redress Compliance runs the exercise from the buyer's side across both sides of the Google relationship: the Workspace edition and seat reconciliation, the Gemini Enterprise pilot design and staging, the consumption forecast in Google's meters, the plan and commit model, the benchmark on seats, tokens and media rates, the contract terms that decide year two, and the negotiation itself on the client's calendar. Our fees are fixed, and every dollar removed from Google's number belongs to the client.

REDRESS COMPLIANCE LLC

1314 E Las Olas Blvd, Fort Lauderdale, FL 33301 | redresscompliance.com | +1 (239) 402-7397
100 percent buyer side. Zero vendor affiliations. No reseller agreements. No referral fees.