Building a "girly assistant" agentic chatbot on Telegram

The most viable architecture combines Python (LangGraph + python-telegram-bot), GPT-4o-mini as the primary reasoning engine, and a curated stack of booking, affiliate, and discovery APIs — all deployable within a $100–500/month API budget. This plan covers every layer from Telegram bot framework through monetization, with specific API names, URLs, pricing, and sandbox availability. The critical insight: most lifestyle APIs (fitness booking, restaurant reservations, fashion products) are partner-gated or affiliate-only, making a hybrid approach of real transactions plus discovery-and-redirect not just strategic but necessary.

The Telegram bot layer: framework and capabilities

Python wins for this use case, and two excellent async frameworks exist. Medium `python-telegram-bot` Telegram (~28.6K GitHub stars, GitHub v22.6) offers the most mature conversation handling Rost Glukhov with built-in `ConversationHandler` for multi-step flows, while `aiogram` (~5K stars, v3.x) delivers better raw throughput with native async design Telegram Restack and a full middleware pipeline. Both support Telegram Bot API 9.3+. GitHub For TypeScript teams, `grammY` is the standout choice — TypeScript-first, serverless-friendly, Telegram Grammy with an official conversations plugin. Grammy

Telegram's platform provides three critical features for this bot. **Inline keyboards** with callback queries enable rich interactive menus (outfit selectors, booking confirmations, deal browsing). The **Payments API** supports both physical goods via Stripe and digital goods via **Telegram Stars** — the mandatory currency for in-app digital purchases. Telegram Stars work simply: send an invoice with `currency: "XTR"`, Telegram handles the payment flow, Telegram GitHub and developers receive proceeds minus a small administrative fee (~5%). telegram Recurring subscriptions are supported via `createInvoiceLink` with a `subscription_period` parameter. Telegram GitHub **Telegram Mini Apps** (Web Apps) open a WebView inside Telegram for rich UI experiences Telegram Creole Studios — useful for virtual try-on, outfit browsing, or trip itinerary visualization. Over **500 million users** interact with Mini Apps regularly. Magnetto

For conversation state management in production, **Redis is essential** — store per-user conversation state keyed by `telegram_id` for sub-50ms retrieval and horizontal scaling. Any bot instance can handle any user's request. Trapnest Pair with PostgreSQL for long-term persistence (preferences, booking history) and pgvector for semantic memory retrieval.

AI orchestration: the reasoning engine and agent framework

The agent layer sits between Telegram and external APIs, deciding what tools to invoke based on natural language input. Medium Two viable approaches exist within budget.

LangGraph (Python, ~12K GitHub stars) provides the most mature infrastructure for this use case. It models agents as graph-based state machines FASHN with nodes, edges, and conditional routing Turing — ideal for multi-step workflows [Langfuse] like "find me a yoga class near Union Square this Saturday, then book a restaurant for after." LangGraph's checkpointing enables durable execution, [FASHN] and LangSmith provides observability for debugging agent behavior. [LangChain] [Composio] The alternative for TypeScript teams is the **Vercel AI SDK** (v6, 20M+ monthly npm downloads) [Vercel] with built-in agent loops and streaming. [Strapi]

For simpler cases or teams wanting maximum control, a **custom agent loop** works: call the LLM with tool definitions, execute any requested tools, feed results back, repeat until the LLM produces a final response. [MachineLearningMastery] Both Claude and OpenAI APIs follow this pattern natively.

LLM pricing determines your budget ceiling. The critical comparison:

Model	Input / 1M tokens	Output / 1M tokens	Best for
GPT-4o-mini	**$0.15**	**$0.60**	95% of chatbot interactions — routing, simple recommendations, conversation
GPT-4o	$2.50	$10.00	Complex reasoning, multi-step planning, trip itineraries
Claude Sonnet 4.5	$3.00	$15.00	Nuanced styling advice, creative writing, detailed analysis
Claude Haiku 3.5	$0.80	$4.00	Budget alternative to GPT-4o-mini

GPT-4o-mini is the recommended default [Price Per Token] at roughly [Price Per Token] **$18/month for 1,000 daily active users** (assuming ~20 messages/user/day, ~2K input + 500 output tokens per call). Route only complex tasks to GPT-4o or Claude Sonnet. With prompt caching (available on both Anthropic and OpenAI), input costs drop by up to 90% for repeated system prompts and tool definitions. [Costgoat]

Context window management for Telegram bots should follow a **sliding window + summarization** pattern: keep the last 10–20 messages verbatim, summarize older context via LLM, and store extracted user preferences/facts in structured format [PyPI] (PostgreSQL JSONB). This achieves 80–90% token cost reduction [Mem0] compared to sending full history.

Yoga and fitness class booking: Mindbody is the only real option

The fitness booking landscape is surprisingly constrained. **Mindbody API** (https://developers.mindbodyonline.com) is the only viable option for programmatic class booking across studios, with access to **70,000+ businesses** worldwide.

Mindbody's v6 REST API provides `GetClasses`, `AddClientToClass`, and `CheckoutShoppingCart` endpoints that enable **real booking** — not just discovery. The free tier includes 5,000 API calls per billing cycle [Gymdesk] with a full sandbox environment (1,000 calls/day). Production pricing is **$0.002 per API call**, [Gymdesk] making moderate usage (~50,000 calls/month) roughly **$100/month**. Access requires creating a developer account, building in sandbox, then requesting go-live review from Mindbody. [Mindbodyonline]

ClassPass has no public consumer API — it was acquired by Mindbody and its Inventory API is exclusively for booking-system partners. Fitogram shut down December 31, 2024. [Fitogram] Momoyoga, Wellhub, [Gympass] and Vagaro [TEC] all lack consumer-facing booking APIs. No true cross-studio aggregator API exists.

The recommended architecture: use **Google Places API** (~$0–50/month, 10,000 free requests) [Nicola Lazzari] for studio discovery and search, then route booking through **Mindbody API** for Mindbody-powered studios, or generate redirect links for non-Mindbody studios.

Beauty and makeup: Perfect Corp leads, affiliates monetize

No major beauty retailer (Sephora, Ulta) offers a public developer API. The beauty feature stack must be assembled from specialized providers.

Skin analysis is best handled by **Perfect Corp's AI APIs** (https://perfectcorp.com/business/ai-apis) — they detect 15 skin conditions from photos, [Makeup AR] trained on 70,000+ medical-grade images, with a pay-as-you-go model starting at **1,000 free API units** for prototyping. [Business Wire] Their Essential plan runs ~$350/month. For a budget MVP, use **GPT-4 Vision** with structured prompts for skin type classification at ~$20–50/month — surprisingly effective for general guidance, though not medically certified. For clinical-grade results, **Skinive.Cloud** (€300/month) is CE-marked medical software. [GetApp] [Skinive]

Virtual try-on is Perfect Corp's strongest offering, with makeup try-on, hair color simulation, and foundation shade matching in a single API. [Business Wire] **Banuba** (https://banuba.com) offers a 14-day free trial with 1,000+ AR filters, supporting virtual makeup, hair recolor, and automated seasonal color analysis. [Banuba] ModiFace (owned by L'Oréal) is best-in-class quality but requires enterprise partnerships likely exceeding $1,000/month.

Product data comes from three sources. **Open Beauty Facts** (https://world.openbeautyfacts.org) provides a free, open database of 100,000+ products with ingredients and allergens [Open Beauty Facts] — quality varies but it's the only free option. **Amazon's Product Advertising API** (transitioning to Creators API by April 2026) gives access to the broadest beauty product catalog [Wikipedia] with **10% commission on Luxury Beauty**. [Shopify] **Sephora data** is accessible via RapidAPI scrapers [Apify] (~$0.01/request) or Apify actors [Apify] (free tier: 5 runs/month), though this sits in a legal gray area.

For beauty monetization, stack three affiliate programs:

- **Amazon Associates** — Luxury Beauty at **10% commission**, [HM Marketing] broadest catalog, migrating to Creators API [Amazon]

- **Sephora via Rakuten** — 5–10% commission, 13,000+ products, premium brand credibility
- **Ulta via Impact** — 2–10% commission, 25,000+ items, 30-day cookie window

---

Fashion recommendations and deals: affiliate networks are the backbone

Fashion APIs don't exist in the traditional sense. **No major fast-fashion retailer** (Fashion Nova, Shein, Zara, H&M, PrettyLittleThing) offers a public product API. All product data flows through affiliate network product feeds.

CJ Affiliate (https://www.cj.com) is the recommended primary network (Peak View Marketing) — it hosts Fashion Nova, Shein, Nike, Adidas, Nordstrom, Levi's, Macy's, and hundreds more. Its Publisher API provides product catalog search, deep link generation, and commission tracking. Free to join. **Awin** (which absorbed ShareASale) adds European brands including H&M, PrettyLittleThing, ASOS, and Etsy — also free for publishers with a $5 refundable deposit. (Join Brands)

ShopStyle API (via ShopStyle Collective) is the closest thing to a unified fashion product search API — a single endpoint to search products across many retailers with built-in affiliate links. (GitHub) Access requires joining the Collective as a publisher.

For price comparison, **SerpApi Google Shopping** (https://serpapi.com) returns structured JSON with product prices, merchants, and links. **250 free searches/month**, (AIMultiple) scaling to $50/month for 5,000 searches. This enables "find me the best price on these Nike Air Max" queries.

A purpose-built solution called **ChatAds** (https://www.getchatads.com) deserves special attention. It's specifically designed for AI chatbot monetization — it analyzes bot responses in real-time, detects product mentions, and inserts affiliate links in <500ms. Works with Amazon Associates, ShareASale, CJ accounts. SDKs available for Python and TypeScript. You keep 100% of affiliate commissions. (getchatads)

For outfit recommendations, **Marqo-FashionCLIP** and **Marqo-FashionSigLIP** (Hugging Face) (free, open-source on Hugging Face) provide state-of-the-art fashion product embeddings for visual similarity search. Combined with a vector database and the **neuralwork/fashion-style-instruct** dataset (Hugging Face) for training, this enables "find outfits similar to this" or body-type-aware styling recommendations.

Web scraping for sale alerts is technically feasible using Playwright (handles JavaScript-heavy sites) with residential proxy rotation, but legally risky and operationally expensive. The preferred approach: use affiliate product feeds (updated daily) with price tracking, supplemented by **Keepa API** (€50/month) for Amazon price history. (TraceFuse)

Trip planning and flight booking: Amadeus plus Duffel

Three APIs support **actual flight booking** (not just search):

API	Booking type	Cost model	Free tier	Airlines
Amadeus Self-Service	Real PNR creation	€0.01/search call	200–10K free/month	400+ airlines
Duffel	Real PNR creation	$3/order + 1% managed	Free sandbox	300+ airlines
Kiwi.com Tequila	Kiwi as merchant of record	Commission-based	Search is free	Virtual interlining

Amadeus (https://developers.amadeus.com) is the best value starting point. Self-Service APIs require no IATA certification, Zoftify Traveltekpro provide Python and Node.js SDKs, Traveltekpro and cover the full Search → Price → Book flow. The free tier includes 200–10,000 calls/month depending on endpoint; production overage is **€0.001–0.025 per call**. GP Solutions Traveltekpro At ~€0.01/call for flight search, $100/month buys ~10,000 searches. Amadeus also covers hotels (booking endpoint is free), destination experiences, and airport data. AltexSoft

Duffel (https://duffel.com) offers the most modern developer experience with seat maps, baggage selection, and managed IATA accreditation. RapidAPI At **$3 per flight order plus 1% of order value**, it's cost-effective for moderate booking volumes. duffel Their sandbox includes "Duffel Airways" — a fake airline for reliable test bookings. GitHub Official SDKs in Python, JavaScript, Ruby, C#, and Java. RapidAPI

Kiwi.com Tequila (https://tequila.kiwi.com) uniquely offers "virtual interlining" — connecting flights from different airlines that don't normally codeshare. Search API is free; booking earns commission with Kiwi handling ticketing and customer service.

For trip planning, there's no dedicated itinerary API. The recommended approach: use **Google Places API** for attraction discovery, feed structured data into the **LLM agent** (GPT-4o or Claude Sonnet), and have it generate personalized itineraries. Add **Viator Partner API** (300K+ tours, commission-based) Phptravels and **GetYourGuide** (8%+ commission) Zoftify for bookable activities.

Google Flights has no API Apify — QPX Express was deprecated in 2018. SerpApi's Google Flights scraper SerpApi ($75/month for 5,000 searches) is the best proxy for price comparison data.

Restaurant reservations: a partner-gated landscape

Restaurant booking APIs are almost universally **partner/contract-based** with no self-service access:

- **OpenTable** (https://docs.opentable.com) offers real reservation booking via API but requires partner approval. A newly launched API Sandbox enables testing. The affiliate program provides reservation links with tracking — the most accessible entry point.

- **Resy** (owned by American Express) has **no public developer API** whatsoever. Internal APIs exist but are completely closed.
- **TheFork** (https://docs.thefork.io) covers thefork 80,000+ European restaurants with full booking capability (create reservation with party size, date, customer data) — requires partnership. TheFork Manager API Tracker

For discovery, combine **Google Places API** (~$0–50/month, 10,000 free Essentials requests) Nicola Lazzari with **Yelp Fusion API** (https://docs.developer.yelp.com, 30-day trial with 5,000 free calls, then ~$240/month). Neither enables booking — they provide restaurant search, ratings, reviews, and photos.

The realistic architecture for restaurants: **Google Places for discovery → inline keyboard with restaurant options → deep link to OpenTable/Resy/Google Maps for actual reservation**. This redirect-based approach avoids the partner approval barrier while still delivering value.

Recommended architecture and tech stack

The complete stack, optimized for a solo developer or small team at $100–500/month in API costs:

Layer	Technology	Why
Language	**Python 3.12+**	Best AI/ML ecosystem, LangGraph maturity, larger hiring pool
Telegram SDK	**python-telegram-bot v22+**	Most mature conversation handling, async, 28K+ stars
Agent framework	**LangGraph**	Stateful graph orchestration, tool-based agents, checkpointing
Primary LLM	**GPT-4o-mini** ($0.15/$0.60 per 1M tokens)	Best cost/performance for 95% of interactions
Complex LLM	**GPT-4o** or **Claude Sonnet 4.5**	Multi-step planning, creative styling advice
Database	**Supabase** (PostgreSQL + pgvector + Auth)	$25/month Pro tier, all-in-one, vector search built-in
Cache	**Redis** (Railway or Upstash)	Sessions, rate limiting, API response caching, ~$5–10/month
Deployment	**Railway.app**	Best DX, one-click databases, $5–20/month base
Monitoring	**LangSmith**	Agent tracing, debugging, evaluations
Payments	**Telegram Stars** (digital) + **Stripe** (physical)	Native Telegram integration

Agent architecture pattern

Structure the bot as a **single agent with a tool registry** — not multiple specialized agents: Speakeasy

```
Telegram Update → Rate Limiter (Redis) → LangGraph Agent Controller
  → LLM classifies intent + selects tool(s)
  → Tool execution (with circuit breaker + cache check)
  → LLM synthesizes result into conversational response
  → Telegram sends formatted message with inline keyboard
```

Each API integration registers as a tool with name, description, JSON schema, rate limit config, and fallback behavior. Medium LangGraph's graph structure handles multi-step workflows FASHN (e.g., search flights → confirm selection → book → send confirmation). This is far more maintainable than state machines for 10+ skills — adding a new tool requires registering one function, not rewiring transition logic.

Error handling follows a fallback chain: Primary API → Cached result → Alternative API → Graceful degradation message. Use pybreaker for circuit breakers (open after 5 failures, 30-second reset) Box Piper and aiolimiter for client-side rate limiting per API.

Budget breakdown and monetization math

Monthly API costs at moderate usage (1,000–5,000 daily active users)

Category	Service	Est. monthly cost
LLM (GPT-4o-mini primary)	OpenAI API	$18–150 (scales with users)
Fitness booking	Mindbody API	$0–100
Discovery	Google Places API	$0–50
Restaurant reviews	Yelp Fusion API	$0–240
Flight search	Amadeus Self-Service	$0–100
Price comparison	SerpApi Google Shopping	$0–50
Beauty AI	Perfect Corp or GPT-4V	$0–100
Infrastructure	Railway + Supabase + Redis	$35–60
Total		**$53–850**

The budget is feasible at 1,000 users and requires careful API selection at 5,000+ users. **LLM costs are the dominant expense** — prompt caching, model routing (GPT-4o-mini for simple tasks), and response caching are critical optimizations. Costgoat Finout

Revenue projections

Stream	1K users	10K users
Fashion/beauty affiliate (CJ + Amazon + Sephora)	$200–400	$2,000–4,000
Telegram Stars (tips + premium unlocks)	$100–200	$1,000–2,000
Premium subscription (3–5% conversion at $4.99/mo)	$150–250	$1,500–2,500
Booking referral commissions	$50–100	$500–1,000

Stream	1K users	10K users
Total revenue	**$500–950**	**$5,000–9,500**

Break-even is achievable at approximately 2,000–3,000 active users, assuming moderate affiliate conversion rates and 3–5% premium subscription uptake. Fashion affiliate revenue alone (via ChatAds or Skimlinks auto-linking) can cover API costs at scale.

Conclusion

This architecture works because it embraces the reality of the API landscape rather than fighting it. **Real transactions** are possible for fitness classes (Mindbody), flights (Amadeus/Duffel), and activities (Viator/GetYourGuide). **Discovery plus redirect** is the pragmatic path for restaurants (OpenTable/Resy links), fashion purchases (affiliate deep links via CJ/Awin), and beauty products (Amazon/Sephora affiliate). The LLM agent layer — powered by GPT-4o-mini with LangGraph orchestration — makes this hybrid approach feel seamless to users by handling intent classification, API routing, and conversational synthesis in natural language. [Medium]

Three technical decisions matter most: choosing GPT-4o-mini as the default model (20x cheaper than GPT-4o with adequate quality for most interactions), [Costgoat +2] using Redis for conversation state (enabling horizontal scaling from day one), [Trapnest] and implementing ChatAds or Skimlinks for automatic affiliate monetization (turning every product recommendation into potential revenue without manual link management). [FasterCapital] Start with the Mindbody + Amadeus + CJ Affiliate core, ship to users, and expand tool integrations based on actual usage patterns.