CORE BUILD DIRECTIVE

Mandatory: Route **100%** of all AI/LLM requests through **kAIxuGateway13**. Any direct provider call or SDK usage is an automatic build failure.

Standard ID: KXG13-CORE • Version 1.0 • Effective: February 19, 2026

Gateway Base	`https://kaixugateway13.netlify.app`
Non-stream	`POST /.netlify/functions/gateway-chat`
Streaming SSE	`POST /.netlify/functions/gateway-stream`
Auth	`Authorization: Bearer <KAIXU_VIRTUAL_KEY>`

1) Absolute Prohibitions (Zero Tolerance)

- Do **not** call any provider endpoints directly (OpenAI, Anthropic, Gemini).
- Do **not** use any provider SDKs (client or server).
- Remove/disable any direct HTTP calls to: `https://api.openai.com/...`, `https://api.anthropic.com/...`, `https://generativelanguage.googleapis.com/...`
- Ensure there are **no remaining call sites** (including legacy helpers, unused modules, dead code branches).

2) Required Streaming Method (POST + ReadableStream)

- Streaming must use `fetch` with POST and parse SSE frames from a `ReadableStream`.
- Do **not** use `EventSource` (it cannot POST).

3) Required Request Payload Shape (Must Match)

- Every gateway call must send the following JSON fields: `provider`, `model`, `messages`, `max_tokens`, `temperature`.

```
{
  "provider": "openai" | "anthropic" | "gemini",
  "model": "string",
  "messages": [{"role":"system"|"user"|"assistant","content":"text"}],
  "max_tokens": 123,
  "temperature": 0.7
}
```

4) Required Response Handling (Usage + Budget)

- Non-stream responses return `output_text` plus `usage` and `month` budget fields.

- Streaming responses emit SSE events: `meta`, `delta`, `done`, `error`.
- UI rule: show remaining budget `(cap_cents - spent_cents)` wherever usage is displayed.
- If HTTP **402** occurs, show “Monthly cap reached” and block further calls until upgrade/top-up.

5) Required Implementation Pattern (One Gateway Client)

- Implement a single shared gateway client module (example: `src/lib/kaixuGateway.js`).
- Expose `kaixuChat()` for non-stream and `kaixuStreamChat()` for streaming.
- Replace every AI call site with these gateway functions.
- Preserve current UX: chat history, system prompts, model selection, streaming behavior (if present).

6) Required Key Management (Kaixu Key Only)

- The **only** AI credential the app may accept/store is `KAIXU_VIRTUAL_KEY` (a “Kaixu Key”).
- Remove any UI/logic/storage for provider keys.
- Provider keys must not exist anywhere in the codebase.

7) Required Error Mapping (User-Visible)

- **401** → prompt for Kaixu Key / invalid key messaging
- **402** → “Monthly cap reached” + block calls
- **429** → rate limit message + retry guidance
- **500** → gateway/provider error message

8) Delivery / Acceptance Criteria (Build Fails if Any Misses)

- Zero direct provider endpoints/SDK usage remains anywhere in the repo.
- Gateway base + endpoints match exactly as specified in this directive.
- Payload shape matches exactly for all AI calls.
- Streaming uses POST fetch + ReadableStream SSE parsing (no EventSource).
- UI/UX is not broken; existing features remain intact.
- Deliver **full updated files** (no partial snippets, no TODOs).