

Documentação de Projeto

Otimizador de Prompts & Calculador de Tokens

1. Visão Geral do Projeto

O ecossistema de software focado em eficiência de inteligência artificial. O sistema atuará como um "Engenheiro de Prompt Automático" e "Calculador de Custos", reescrevendo textos para economizar tokens, melhorar a semântica e reduzir alucinações. A distribuição será feita em três plataformas: ferramenta Web, extensão de navegador e agente para VS Code.

2. O Motor Central (Inteligência e Negócios)

2.1. Contagem de Tokens (100% Offline / Client-Side)

A privacidade e a velocidade são garantidas calculando o peso do prompt localmente na máquina do usuário, sem requisições de rede.

- **OpenAI (GPT-4o, mini):** Utilização do pacote NPM `gpt-tokenizer` no front-end.
- **Anthropic (Claude 3.5):** Utilização do pacote oficial `@anthropic-ai/tokenizer`.
- **Google (Gemini):** Arquitetura Híbrida (Lazy Loading). A extensão carrega o motor `sentencepiece-js` e permite o download opcional do dicionário oficial (`.model`), armazenado no **IndexedDB**.

2.2. Motor de Otimização Semântica

Utilização do modelo **GPT-4o-mini** via API da OpenAI para reescrever prompts com foco em economia e clareza. Aplicação de frameworks como Ação/Contexto/Formato e remoção de redundâncias.

2.3. Dicionário de Personas e Especialidades

A ferramenta atuará também como consultor especializado, adaptando a engenharia de prompt à profissão do usuário. A funcionalidade oferece dois modos de operação:

- **Atuação Invisível (Padrão):** A persona (ex: Desenvolvedor Sênior, Analista de Web Analytics) atua exclusivamente no *System Prompt* do servidor. O modelo utiliza o jargão técnico e o

conhecimento do especialista para refinar o texto do usuário, sem declarar a persona no resultado final.

- **Injeção de Contexto:** Ativada por um *checkbox*, esta opção anexa a declaração da persona no topo do prompt otimizado gerado. É o modo ideal para estabelecer contexto ao iniciar uma nova conversa nas IAs de destino.

2.4. Modelo de Negócios e Monetização

O sistema suportará três níveis de acesso principais:

- **BYOK (Bring Your Own Key):** Gratuito para o projeto, o usuário usa sua própria chave.
- **Nuvem Free:** Cota limitada (ex: 15 otimizações/dia, acesso apenas à otimização Generalista) para atração de tráfego.
- **Nuvem Premium:** Venda de licenças para uso ilimitado da infraestrutura da aplicação e acesso exclusivo ao **Dicionário de Personas Especialistas**.

Projeção de Precificação Premium

Custo base estimado por power user extremo: R\$ 1,42 / mês.

Modelo	Valor	Taxas (Est.)	Lucro Líquido	Estratégia
Mensal Pão na Chapa	R\$ 4,90	R\$ 0,75	R\$ 2,73	Baixo atrito / Volume
Mensal Psicológico	R\$ 9,90	R\$ 1,00	R\$ 7,48	Margem de Segurança
Ticket Dev	R\$ 14,90	R\$ 1,20	R\$ 12,28	Padrão Corporativo
Passe Anual	R\$ 49,90	R\$ 3,00	R\$ 29,86	Caixa antecipado
Lifetime	R\$ 97,00	R\$ 5,00	R\$ 92,00	Capital imediato

3. Arquitetura das Frentes

3.1. Frente A: Aplicação Web (Astro + React)

Ferramenta de atrito zero no site principal. Foco no fluxo direto: Input → Otimização → Cópia. Implementação de Rate Limiting por IP para sessões gratuitas. **A interface inclui um dropdown para seleção do Perfil Especialista e um checkbox para Injeção de Contexto.**

3.2. Frente B: Extensão de Navegador

Injeção de interface (React Portals) diretamente no ChatGPT, Claude e Gemini. **Inclui seletores de Personas e checkbox de injeção.** Também apresenta o Estimador de Contexto que acumula tokens da sessão e avisa quando a janela de contexto está saturada (~80% do limite).

3.3. Frente C: Extensão VS Code

Integração nativa como *Chat Participant* do GitHub Copilot. **Suporta perfis especialistas através de subcomandos de menção (ex: @otimizador:senior).** Utiliza a SecretStorage API para guardar chaves de API com criptografia do sistema operacional.

4. Resumo da Stack Tecnológica

Camada	Tecnologia Selecionada
Front-end	React / TypeScript
Back-end (SSR)	Astro
Hospedagem	Node.js (Railway / Vercel)
Extensão Browser	Plasmo / Vite Web Ext
Tokenização	gpt-tokenizer / @anthropic-ai/tokenizer
LLM Otimizador	OpenAI API (GPT-4o-mini)