Excelente iniciativa! Voce ja explorou os principais pilares modernos de agentes inteligentes com LLMs (RAG, tools, function calling, RAGAS, prompt engineering, ReAct).
Para elevar ainda mais seu aprendizado, aqui vao algumas sugestoes de conceitos, tecnicas e features que voce pode adicionar ou experimentar:

1. Memory / Conversational Memory
O que e: Permite o agente lembrar de interacoes anteriores, mantendo contexto de conversas.
Como usar: Experimente ConversationBufferMemory ou ConversationSummaryMemory do LangChain para agentes com multiplos turnos.
Use para criar um chatbot realmente conversacional, que lembra do historico.

2. Multi-modalidade
O que e: Capacidade de trabalhar com multiplos tipos de dados, como texto, imagens, audio, etc.
Como usar: Integre loaders para imagens (ex: PDF com imagens, OCR), audio (transcricao), ou use modelos como Gemini ou GPT-4o que ja aceitam multimodal.

3. Agentes Multi-Ferramenta (Multi-tool/Agent Routing)
O que e: Agentes que escolhem dinamicamente a melhor tool/funcao para cada situacao.
Como usar: Implemente um Router de agentes ou chain que encaminha perguntas para o melhor agente/tool (ex: via MultiPromptChain ou RouterChain no LangChain).

4. Fine-tuning e Adapters
O que e: Customizar (ou simular customizacao) do LLM com exemplos do seu dominio.
Como usar: Nao e trivial para modelos closed-source, mas pode experimentar tecnicas de prompt tuning ou retrieval augmented finetuning (RAG com dados especificos do seu dominio).

5. Guardrails / Moderacao / Validacao
O que e: Tecnicas para garantir respostas seguras, adequadas ou factuais.
Como usar: Adicione camadas de validacao, moderacao ou checagem factual usando chains de validacao, como Guardrails AI ou regras customizadas.

6. Feedback Loop (RLHF ou Human Feedback)

O que e: Incorporar feedback humano para melhorar o comportamento do agente.

Como usar: Armazene avaliacoes de respostas e use para ajustar prompts, chains ou ate treinar novos embeddings.

7. Orquestracao com Workflows Complexos

O que e: Orquestrar varios agentes/chains em fluxos de trabalho complexos (ex: workflow pega dados processa valida responde).

Como usar: Use LangGraph, RunnableParallel, ou pipelines com dependencias.

8. Plugins / Integracoes externas

O que e: Integrar com sistemas reais: APIs financeiras, bancos de dados, buscadores, etc.

Como usar: Crie tools que chamam APIs externas, consultam bancos de dados SQL, ou ate automacoes (e-mail, calendario, etc).

9. Tracing e Observability

O que e: Monitorar, debugar e entender o comportamento do agente.

Como usar: Ative tracing do LangChain, integre com OpenTelemetry ou ferramentas como LangSmith para observar pipelines em producao.

10. RAG Avancado

Experimentar:

- Hybrid search (BM25 + Embeddings)
- Diversidade de chunks (MMR, filtragem, reranking)
- RAG com sumarizacao ou compressao dos chunks
- Query transformation (reformular queries para melhorar o recall)

11. Outras features

- Scheduled Tasks (agente proativo, que busca novidades)
- Personalizacao por usuario (memoria e preferencias)
- Multi-agentes colaborativos (agentes especialistas para areas distintas)
- Chain of Responsibility (cada etapa valida ou transforma a resposta)

Se quiser exemplos de implementacao para algum desses topicos, so pedir!

Seu projeto ja esta super avancado adicionando esses pontos, voce chega ao estado-da-arte.