Graduation Project Proposal

# AI Company Knowledge Hub

## مساعد معرفة مؤسسي ذكي للشركات الصغيرة والمتوسطة باستخدام Hybrid RAG وVerified Citations

منصة بسيطة تتيح للشركة رفع مستنداتها الداخلية، ثم يستطيع الموظفون سؤال النظام والحصول على إجابات موثقة بالمصدر بدل البحث اليدوي داخل ملفات كثيرة.

# 1. الملخص التنفيذي

فكرة المشروع هي بناء منصة ذكاء اصطناعي داخلية تساعد الشركات على تحويل ملفاتها وسياساتها وإجراءاتها التشغيلية إلى مساعد معرفة يمكن للموظف سؤاله مباشرة. بدل أن يبحث الموظف في ملفات PDF وWord وSOPs وسياسات HR لفترات طويلة، يقوم النظام باسترجاع الجزء المناسب من المستندات ثم يولد إجابة واضحة مع ذكر المصدر.

**الفكرة في جملة واحدة:**
AI Knowledge Assistant للشركات الصغيرة والمتوسطة، يدعم العربي والإنجليزي، ويجيب من مستندات الشركة فقط مع مصادر موثقة.

## لماذا الفكرة مناسبة كمشروع تخرج؟

### حل قابل للعرض

يمكن عمل Demo واضح: رفع ملفات، سؤال النظام، ثم ظهور إجابة بمصدر.

### مشكلة حقيقية

الشركات لديها معرفة داخلية متفرقة بين ملفات وسياسات ومجلدات ومصادر مختلفة.

### قابل للتطوير

يمكن تحويله بعد التخرج إلى SaaS للشركات الصغيرة والمتوسطة.

### تقنيات حديثة

المشروع يجمع بين RAG، Vector Database، FastAPI، Document Processing، وDashboard.

## النتيجة المتوقعة من الـ MVP

نسخة أولى تسمح للأدمن برفع ملفات الشركة، وتحويلها إلى قاعدة معرفة قابلة للبحث، ثم تسمح للموظف بسؤال النظام والحصول على إجابات مدعومة بمصادر واضحة، مع Dashboard بسيط يوضح استخدام المعرفة والأسئلة غير المجابة.

# 2. المشكلة التي يحلها المشروع

المشكلة ليست في نقص المعلومات داخل الشركات، بل في صعوبة الوصول إليها بسرعة وبشكل موثوق. كثير من الشركات تمتلك سياسات وإجراءات وملفات تدريبية، لكن الموظفين لا يعرفون أين توجد المعلومة الصحيحة أو هل النسخة التي لديهم محدثة أم لا.

### إهدار وقت الموظفين

الموظف يقضي وقتًا طويلًا في البحث بدل تنفيذ المهام الأساسية.

### تشتت المعرفة

ملفات HR وSOPs والتدريب موزعة بين PDF وWord ومجلدات وWiki داخلي.

### فقدان المعرفة

عند مغادرة موظف خبير، تفقد الشركة جزءًا من المعرفة غير الموثقة.

### إجابات غير موحدة

كل موظف قد يحصل على إجابة مختلفة حسب الشخص الذي سأله أو الملف الذي وجده.

### ضعف الدعم العربي

الشركات العربية تحتاج تجربة عربية/إنجليزية مهيأة للمصطلحات المحلية وواجهة RTL.

### مستندات غير محدثة

الإدارة لا تعرف دائمًا أي ملفات أصبحت قديمة أو ناقصة.

**مثال عملي:** موظف جديد يسأل: "إزاي أقدم طلب إجازة؟". بدل البحث في ملف HR Policy، يسأل المساعد، فيرجع النظام الإجابة من قسم الإجازات في الملف مع اسم الملف ورقم الصفحة أو رقم الجزء المسترجع.

# 3. الحل المقترح

الحل هو منصة مركزية ترفع عليها الشركة مستنداتها الداخلية. يقوم النظام بتحليل المستندات، تقسيمها إلى أجزاء صغيرة، إنشاء Embeddings، وتخزينها في Vector Database. عندما يسأل الموظف سؤالًا، يستخدم النظام Hybrid Retrieval لاسترجاع أفضل المقاطع، ثم يولد إجابة موثقة بالمصدر.

## رحلة الاستخدام الأساسية

1) **Admin Upload:** الأدمن يرفع ملفات الشركة مثل PDF وDOCX.

2) **Document Processing:** النظام يستخرج النص، يقسمه إلى chunks، ويضيف metadata مثل اسم الملف والقسم والصفحة.

3) **Indexing:** يتم إنشاء embeddings وتخزينها في Qdrant مع البيانات الوصفية.

4) **Employee Chat:** الموظف يسأل سؤالًا بالعربي أو الإنجليزي.

5) **Hybrid Retrieval:** النظام يبحث دلاليًا وبالكلمات المفتاحية للحصول على أدلة من المستندات.

6) **Answer + Citations:** النظام يولد إجابة ويربطها بالمصدر: اسم الملف، الصفحة، أو chunk reference.

7) **Analytics:** الأدمن يرى أكثر الأسئلة، أكثر الملفات استخدامًا، والأسئلة التي لم يجد لها النظام دليلًا كافيًا.

## القيمة المقدمة للشركة

- تقليل وقت البحث داخل المستندات.
- توحيد الإجابات بناءً على مصادر موثقة.
- تحسين تجربة الموظفين الجدد في onboarding.
- كشف فجوات المعرفة والملفات الناقصة.
- تحويل المعرفة الداخلية إلى أصل قابل للبحث والتحليل.

# 4. المستخدمون المستهدفون وحالات الاستخدام

### Operations Manager

يراقب SOPs والأسئلة المتكررة ويعرف أين توجد فجوات في المعرفة.

### Employee

يسأل عن السياسات والإجراءات ويحصل على إجابة مباشرة مع مصدر.

### Admin / HR Manager

يرفع المستندات، يدير الملفات، ويتابع الأسئلة والـ analytics.

## أمثلة على الأسئلة داخل النظام

| القسم | سؤال الموظف | شكل الإجابة المتوقع |
|---|---|---|
| HR | إزاي أقدم طلب إجازة سنوية؟ | خطوات التقديم + الشروط + المصدر من HR Policy. |
| Operations | What is the client onboarding process? | Workflow من 5 خطوات + مصدر من SOP. |
| IT | مين يوافق على طلب لابتوب جديد؟ | اسم الدور المسؤول + شروط الموافقة + مصدر. |
| Finance | إيه سياسة السلف أو المصروفات؟ | ملخص policy + الاستثناءات + مصدر. |
| Training | فين خطوات تدريب الموظف الجديد؟ | Checklist مستخرجة من training manual. |

**ميزة مهمة:** في حالة عدم وجود دليل كافٍ في المستندات، لا يقوم النظام بتأليف إجابة، بل يرد برسالة واضحة: "لم أجد معلومات كافية داخل مستندات الشركة المرفوعة".

# 5. تحليل المنافسين

توجد حلول عالمية قريبة من فكرة المشروع. هذا لا يقلل من قيمة الفكرة، بل يثبت أن المشكلة حقيقية. التمييز المقترح هو التركيز على الشركات الصغيرة والمتوسطة، التجربة العربية/الإنجليزية، البساطة، والإجابات الموثقة من الملفات المرفوعة فقط.

| الشركة | ماذا تقدم؟ | ما الذي يميزها؟ | الفجوة التي نستهدفها |
|---|---|---|---|
| **Glean** | Work AI platform متصلة ببيانات المؤسسة للبحث والإجابات والأتمتة. | قوية في enterprise search وربط مصادر كثيرة داخل الشركة. | موجهة أكثر للمؤسسات الكبيرة، بينما مشروعنا يبدأ بتجربة upload-and-chat بسيطة للشركات الصغيرة. |
| **Microsoft 365 Copilot** | مساعد AI داخل Microsoft 365 ويبحث في بيئة العمل ومصادر Microsoft. | تكامل قوي مع Teams وSharePoint وOutlook وOneDrive وصلاحيات Microsoft. | مرتبط بقوة ببيئة Microsoft؛ مشروعنا مستقل ويعمل حتى لو الشركة لديها ملفات فقط. |
| **ChatGPT Company Knowledge** | يستخدم سياق المؤسسة من apps متصلة داخل ChatGPT مع citations. | تجربة ChatGPT قوية ومألوفة، ومناسبة لمشتركي Business/Enterprise/Edu. | مشروعنا يقدم dashboard مستقل لإدارة الملفات، gaps، وجودة المستندات داخل منتج مخصص. |
| **Guru** | Knowledge management platform يركز على governance ومعرفة موثوقة. | قوي في تنظيم المعرفة، التحقق منها، وتحسينها باستمرار. | قد يكون أكبر من احتياج SME بسيط؛ مشروعنا يركز على رفع الملفات والسؤال مباشرة. |
| **Sana AI** | AI agents/search للتعامل مع معرفة ومحتوى المؤسسة. | قوي في بيئات learning/workplace knowledge. | مشروعنا يركز على local Arabic-English SMEs وتجربة أبسط. |
| **Coveo** | Generative AI search وAnswering لتجارب enterprise وcustomer service وweb/ecommerce. | منصة واسعة للبحث والتخصيص والتوصيات. | واسعة ومعقدة؛ مشروعنا narrower وأكثر وضوحًا كـ Knowledge Hub داخلي. |
| **Hebbia** | منصة لتحليل مستندات ضخمة وworkflows خاصة بالقطاعات المالية والقانونية. | قوية في document-heavy analysis وfinancial workflows. | موجهة لقطاعات عالية التخصص؛ مشروعنا عام للشركات الصغيرة والمتوسطة. |
| **Notion AI Search** | يبحث داخل Notion workspace وتطبيقات متصلة مثل Slack وGoogle Drive وJira. | ممتاز للشركات المستخدمة Notion بالفعل. | مرتبط ببيئة Notion؛ مشروعنا مستقل ومبني حول الملفات الداخلية. |

ملاحظة: التفاصيل مبنية على صفحات المنتجات الرسمية والمراجع المذكورة في نهاية الملف.

# 6. نقطة تميز مشروعنا

**Positioning المقترح:**
نحن لا نحاول منافسة Microsoft أو Glean مباشرة. نحن نبني مساعد معرفة خفيف ومناسب للشركات الصغيرة والمتوسطة، يدعم العربي والإنجليزي من البداية، ويعتمد على إجابات موثقة من مستندات الشركة.

## عناصر التميز الرئيسية

### SME-focused

بدون تكاملات Enterprise معقدة. الشركة ترفع مستنداتها وتبدأ الاستخدام.

### Arabic-first

تصميم RTL، أسئلة عربية، مستندات عربية/إنجليزية، ومصطلحات أعمال محلية.

### Citation Verification

طبقة تحقق تتأكد أن الإجابة فعلاً مبنية على المقاطع المسترجعة.

### Evidence-first

كل إجابة يجب أن تكون مدعومة بمصدر من ملفات الشركة، وإلا يرفض النظام الإجابة.

### Document Quality Score

تقييم مبدئي لجودة المستند: قابلية القراءة، وجود أقسام واضحة، واحتمال أنه قديم أو ناقص.

### Knowledge Gap Detection

تسجيل الأسئلة التي لم يجد لها النظام إجابة ليوضح للإدارة ما ينقصها من مستندات.

## الفرق بين دعم العربي Arabic-firstg

| Arabic Support | Arabic-first في مشروعنا |
|---|---|
| العربي يكون لغة ضمن لغات كثيرة. | العربي جزء أساسي من تصميم المنتج والواجهة والاختبار. |
| قد يجيب بالعربي فقط. | يفهم مستندات عربية/إنجليزية ويجيب بلغة المستخدم مع مصادر عربية واضحة. |
| لا يركز على مصطلحات محلية. | يدعم قاموس مصطلحات للشركة مثل إخلاء طرف، عهدة، اعتماد، إجازة سنوية. |

# 7. نطاق الـ MVP المقترح

حتى يكون المشروع قابلًا للتنفيذ كمشروع تخرج، يجب عدم محاولة بناء منصة Enterprise كاملة من أول نسخة. النسخة الأولى ستركز على الوظائف الأساسية التي تثبت قيمة الفكرة.

## داخل نطاق الـ MVP

| | |
|---|---|
| ✓ | Authentication بسيط للأدمن والموظفين. |
| ✓ | رفع ملفات PDF وDOCX. |
| ✓ | استخراج النص وتقسيم المستند إلى chunks. |
| ✓ | إنشاء embeddings وتخزينها في Qdrant. |
| ✓ | Chat interface للموظفين. |
| ✓ | إجابات مع source document وpage/chunk reference. |
| ✓ | رسالة رفض عند عدم وجود دليل كافٍ. |
| ✓ | Feedback: Helpful / Not Helpful. |
| ✓ | Dashboard بسيط: عدد الملفات، عدد الأسئلة، الأسئلة غير المجابة، الملفات الأكثر استخدامًا. |

## خارج نطاق النسخة الأولى

| الميزة | سبب التأجيل |
|---|---|
| تكاملات Slack / Google Drive / Microsoft 365 | تزيد التعقيد، ويمكن إضافتها بعد إثبات قيمة المنتج. |
| صلاحيات متقدمة لكل مستند | مهمة لاحقًا، لكن النسخة الأولى يمكن أن تبدأ بصلاحيات بسيطة. |
| OCR عربي كامل للملفات المصورة | صعب وقد يأخذ وقتًا؛ يمكن دعمه كتحسين لاحق. |
| Long-term memory متقدمة | قد تسبب تعقيدًا في الخصوصية والدقة، ويمكن الاكتفاء بسجل المحادثات. |
| Learning من feedback بشكل تلقائي | في الـ MVP نسجل feedback فقط بدون إعادة تدريب. |

# 8. المكونات التقنية المقترحة

| الطبقة | التقنية المقترحة | الدور في المشروع |
|---|---|---|
| Frontend | Next.js | واجهة الأدمن، واجهة الموظف، chat UI، dashboard. |
| Backend | FastAPI | APIs للملفات، المستخدمين، المحادثات، والتحليلات. |
| Database | PostgreSQL | تخزين المستخدمين، المستندات، الرسائل، feedback، والـ analytics. |
| Vector DB | Qdrant | تخزين embeddings والبحث الدلالي داخل chunks. |
| AI Workflow | LangGraph | تنظيم خطوات ingestion, retrieval, generation, citation verification. |
| LLM / Embeddings | OpenAI API | توليد الإجابات وتحويل النصوص إلى embeddings. |
| Document Processing | PyMuPDF / Unstructured | استخراج النص من PDF وDOCX وتقسيمه. |
| Deployment | Docker + Railway / Azure | تشغيل الخدمات ونشر demo قابل للتجربة. |

## Workflow مبسط

```
Admin Uploads Document
        ↓
Document Parser extracts text + metadata
        ↓
Chunking + Embeddings
        ↓
Store chunks in Qdrant + metadata in PostgreSQL
        ↓
Employee asks a question
        ↓
Hybrid Retrieval gets relevant evidence
        ↓
Answer Generation creates grounded response
        ↓
Citation Verification validates support
        ↓
Answer + Sources + Feedback + Analytics
```

# 9. قاعدة البيانات والـ Modules

## Modules الأساسية

### Document Module

رفع، عرض، حذف، ومعرفة حالة معالجة كل مستند.

### Auth Module

تسجيل دخول، أدوار بسيطة: Admin وEmployee.

### Chat Module

حفظ المحادثات والرسائل وربط الإجابة بالمصادر.

### RAG Module

استرجاع المقاطع المناسبة وتوليد إجابات من المصادر.

### Analytics Module

عرض أكثر الأسئلة، الأسئلة غير المجابة، والملفات المستخدمة.

### Feedback Module

تسجيل تقييم المستخدم للإجابة.

## جداول مقترحة في PostgreSQL

| الجدول | أهم الحقول | الغرض |
|---|---|---|
| users | id, name, email, password_hash, role | إدارة المستخدمين والأدوار. |
| documents | id, title, file_path, status, uploaded_by, created_at | إدارة الملفات المرفوعة. |
| document_chunks | id, document_id, qdrant_point_id, page, section | ربط أجزاء المستند بالـ vector store. |
| conversations | id, user_id, created_at | تجميع رسائل كل محادثة. |
| messages | id, conversation_id, role, content, citations | تخزين الأسئلة والإجابات والمصادر. |
| feedback | id, message_id, rating, comment | قياس جودة الإجابات. |
| knowledge_gaps | id, question, topic, count, status | تسجيل الأسئلة التي لم يجد لها النظام دليلًا. |

# 10. عوامل الثقة وتقليل الهلاوس

أهم خطر في أي نظام يعتمد على LLM هو أن يولد إجابة تبدو مقنعة لكنها غير مدعومة بمصدر. لذلك سيعتمد المشروع على مبدأ Evidence-first.

## آليات مقترحة

### Citation Required

لا يتم عرض إجابة نهائية بدون مصدر واضح من المستندات.

### Grounded Prompting

يُطلب من النموذج الإجابة فقط من المقاطع المسترجعة، وعدم استخدام معرفة خارجية.

### Citation Verification

خطوة تحقق تقارن claims في الإجابة بالمقاطع المسترجعة.

### Confidence Threshold

إذا كانت نتائج البحث ضعيفة، يرفض النظام الإجابة بدل التخمين.

**رسالة الرفض المقترحة:**

"لم أجد معلومات كافية داخل مستندات الشركة المرفوعة للإجابة على هذا السؤال بثقة. يمكن رفع مستند أو سياسة توضح هذا الموضوع."

## مقاييس تقييم بسيطة

- **Answer Accuracy:** هل الإجابة صحيحة بناءً على المستند؟
- **Citation Precision:** هل المصدر المعروض يدعم الإجابة فعلًا؟
- **Refusal Quality:** هل النظام يرفض الإجابة عندما لا يوجد دليل؟
- **User Feedback:** نسبة helpful vs not helpful.

# 11. خطة التنفيذ المقترحة

| المرحلة | المدة | المخرجات |
|---|---|---|
| تحليل المتطلبات والتصميم | أسبوع | ERD، user flows، شاشة مبدئية، وتحديد ملفات demo. |
| Backend + Auth + DB | أسبوعين | APIs للمستخدمين، المستندات، المحادثات، والـ analytics. |
| Document Processing + Qdrant | أسبوعين | رفع الملفات، استخراج النص، chunking، embeddings، وتخزين vectors. |
| RAG Chat + Citations | أسبوعين | واجهة سؤال وإجابة موثقة بالمصادر. |
| Dashboard + Feedback | أسبوع | إحصائيات، أسئلة غير مجابة، وfeedback. |
| Testing + Demo + Documentation | أسبوع | اختبارات، تحسين UI، تجهيز العرض النهائي. |

## تقسيم فريق مقترح

- **Frontend 1:** Admin dashboard وdocument management.
- **Frontend 2:** Employee chat وRTL UI.
- **Backend:** FastAPI, Auth, PostgreSQL APIs.
- **AI/RAG:** Qdrant, embeddings, retrieval, prompt design.
- **Document Processing:** PDF/DOCX parsing, chunking, metadata.
- **Testing/Deployment:** Docker, deployment, QA, presentation.

# 12. المخاطر وخطة التعامل معها

| الخطر | التأثير | خطة التعامل |
|---|---|---|
| ضعف جودة استخراج النص من PDF | إجابات غير دقيقة | اختيار ملفات demo واضحة، وإظهار حالة processing/failure. |
| الهلاوس في إجابات LLM | فقدان الثقة | Evidence-first prompts + citation verification + refusal mode. |
| تعقيد multi-agent الكامل | تأخير التنفيذ | استخدام workflow بسيط بدل agents معقدة في النسخة الأولى. |
| دعم OCR العربي | صعوبة تقنية | تأجيل OCR الكامل إلى Phase 2، والتركيز على PDFs نصية وDOCX. |
| اتساع نطاق المشروع | عدم اكتمال المشروع | الالتزام الصارم بميزات MVP فقط. |
| تكلفة APIs | تكلفة تشغيلية | استخدام dataset صغيرة للـ demo وتخزين embeddings مرة واحدة. |

## معايير نجاح المشروع

### Accuracy Success

معظم أسئلة الـ demo تحصل على إجابة صحيحة من المصدر.

### Functional Success

النظام يرفع ملفات، يعالجها، ويجيب بإجابات موثقة.

### Academic Success

المشروع يوضح مفاهيم RAG، vector DB، citations، وanalytics بوضوح.

### UX Success

المستخدم غير التقني يستطيع استخدام النظام بسهولة.

# 13. المطلوب من الدكاترة

**طلب الموافقة:**
نطلب الموافقة المبدئية على تنفيذ مشروع "AI Company Knowledge Hub" كمشروع تخرج، بشرط أن تكون النسخة الأولى MVP مركزة على رفع المستندات، البحث الذكي، الإجابات الموثقة، ودashboard بسيط للتحليلات.

## حدود ما سيتم تسليمه

- تطبيق Web يعمل من خلال Admin وEmployee roles.
- رفع ملفات PDF/DOCX وتحويلها إلى Knowledge Base قابلة للبحث.
- Chat assistant يجيب من المستندات فقط.
- إجابات تحتوي على مصادر واضحة.
- Dashboard لعرض الاستخدام والأسئلة غير المجابة.
- Documentation يشرح architecture والـ RAG workflow.

## الصياغة المختصرة للموافقة

المشروع المقترح هو منصة AI Knowledge Assistant للشركات الصغيرة والمتوسطة، تعتمد على Hybrid RAG للإجابة على أسئلة الموظفين من مستندات الشركة الداخلية مع إظهار المصادر. يتميز المشروع بدعم عربي/إنجليزي، Knowledge Gap Detection، Verified Citations، وواجهة بسيطة مناسبة للـ SMEs. النسخة الأولى ستكون MVP مركزة وقابلة للتنفيذ خلال مدة مشروع التخرج.

# 14. المراجع الرسمية

تم استخدام صفحات رسمية للمنتجات المنافسة لتحديد الفكرة العامة لكل منافس ونقطة المقارنة. الروابط التالية مخصصة للمراجعة وليست جزءًا من تنفيذ المشروع.


1. Glean - Work AI Platform: https://www.glean.com/
2. Microsoft Support - Supported languages for Microsoft 365 Copilot: https://support.microsoft.com/en-us/microsoft-365-copilot/supported-languages-for-microsoft-365-copilot
3. OpenAI Help Center - Company knowledge in ChatGPT Business, Enterprise, and Edu: https://help.openai.com/en/articles/12628342-company-knowledge-in-chatgpt-business-enterprise-and-edu
4. Sana Agents Help Center - Language Support: https://support.sana.ai/en/articles/154426-language-support
5. Guru - The Governed Knowledge Layer for Enterprise AI: https://www.getguru.com/
6. Coveo - Relevance Generative Answering: https://www.coveo.com/en/platform/generative-ai
7. Hebbia - Official Product Website: https://www.hebbia.com/
8. Notion Help Center - Enterprise Search: https://www.notion.com/help/enterprise-search


تم إعداد هذا الملف كنسخة مبدئية قابلة للتعديل قبل تقديمها بشكل رسمي للجنة المشرفة.