الدليل الشامل لمقاييس تقييم الذكاء الاصطناعي (RAG Evaluation Guide)

هذا الملف يشرح بالتفصيل كل مصطلح تقييم (Metric) استخدمناه في مشروعنا، الفائدة منه، والنتيجة التي استفدناها من تطبيقه على النظام الطبي الخاص بنا.

أولاً: لماذا لا نستخدم (FP, FN, TP, TN) الكلاسيكية؟

في أنظمة التصنيف العادية (مثل معرفة هل الصورة قطة أم كلب)، نستخدم مصطلحات الإيجابي والسلبي الكاذب (False Positives / False Negatives).
نحن نبحث في قاعدة بيانات تحتوي على 10,000 فقرة، **(Information Retrieval - RAG)** لكن في **أنظمة البحث والاسترجاع** والمريض يسأل سؤالاً له إجابة واحدة صحيحة فقط (Ground Truth).

- **إذا جلبنا 10 فقرات (Top-K=10):** سيكون لدينا 1 True Positive، 9 False Positives (لأنها فقرات أخرى مفيدة ربما لكنها ليست الإجابة النموذجية)، و 9,990 True Negatives.
- حساب ה- Precision (10 من أصل 1) 10% العادي هنا سيعطيك دائماً.
- لذلك، في عالم الـ RAG ومحركات البحث مثل جوجل، نستخدم مقاييس مخصصة مثل `@K` "التصنيف" وليس فقط "الترتيب" لتقييم.

ثانياً: مقاييس الاسترجاع (Retrieval Metrics)

1. معدل الإصابة (Hit Rate @K)

- **ما هو؟** هل ظهرت الفقرة الصحيحة (Ground Truth) ضمن أول K نتيجة تم جلبها من قاعدة البيانات؟ (نحن استخدمنا Top-10).
- **فائدته:** يخبرنا هل محرك البحث (Vector DB) يعمل أم أنه "أعمى" ولا يجد الإجابة. إذا كان الـ Hit Rate صفراً، فالذكاء الاصطناعي سيهلوس حتماً.
- **ماذا استفدنا؟** تأكدنا أن نظامنا قوي جداً بنسبة نجاح تفوق 92%، مما يعني أن المعلومة لا تضيع أبداً.

2. دقة البحث (Precision @K)

- **ما هو؟** من بين الـ 10 فقرات التي جلبناها، كم فقرة كانت صحيحة تماماً؟
- **فائدته:** يخبرنا بكمية "الضوضاء" (Noise). إذا جلبنا 10 فقرات وكانت واحدة فقط مفيدة، فالـ Precision 10 هو 10%.

- ليقوم بتنظيف هذه **Reranker** يجلب ضوضاء كثيرة، ولذلك قمنا بابتكار خطوة **الـ** Vector Search **ماذا استفدنا؟** أدركنا أن الـ الضوضاء.

3. متوسط الرتبة العكسية (MRR - Mean Reciprocal Rank)

- **ما هو؟** يقيس "أين ظهرت" الإجابة الصحيحة. إذا ظهرت في المركز الأول، تأخذ 1.0. إذا ظهرت في المركز الثاني، تأخذ 0.5. إذا ظهرت في المركز الثالث، تأخذ 0.33.
- سيضطر لقراءة نصوص كثيرة قبل أن يجد الإجابة، أم أنها في الصدارة؟ (LLM أو الـ) **فائدته:** يخبرنا هل المستخدم
- **ماذا استفدنا؟** تأكدنا أن الأدوية تُسترجع غالباً في المركز الأول (MRR 0.83).

4. جودة الترتيب (NDCG - Normalized Discounted Cumulative Gain)

- **ما هو؟** مقياس معقد قليلاً يعطي وزناً أكبر للنتائج الصحيحة التي تظهر في الأعلى، ويعاقب النظام إذا جلب نتائج صحيحة ولكن في ذيل القائمة.
- **فائدته:** التقييم المثالي للـ Reranker.
- **ماذا استفدنا؟** ساعدنا في اكتشاف أن الـ Reranker الأجنبي (MS-MARCO) كان يدمر ترتيب الأدوية المصرية، وبناءً عليه اتخذنا القرار المعماري بفصله (Bypass).

ثالثاً: مقاييس التوليد والأمان الطبي (Generation & Safety Metrics)

5. دقة الإجابة (Answer Relevance)

- **ما هو؟** يقيس هل الإجابة النهائية التي نطق بها الموديل ترد فعلاً على سؤال المريض؟ أم أن الموديل يتحدث في وادٍ آخر؟
- **فائدته:** يحمينا من الإجابات "المراوغة".
- **ماذا استفدنا؟** قمنا بتغيير طريقة قياسه من (Cosine Similarity) العقيمة، إلى استخدام (LLM-as-a-Judge) ليعمل كطبيب بشري يقيم الإجابة من 1 إلى 5. وحصل نظامنا على 4.8 / 5.0.

6. الأمان والموثوقية (Faithfulness / RAGAS)

- **ما هو؟** يقيس مدى ارتباط كلام الموديل بالفقرات التي جلبها. هل الموديل "اخترع" معلومة من رأسه (Hallucination) أم استند 100% للمرجع؟
- **فائدته: أهم مقياس طبي على الإطلاق!** لا مجال للتأليف في الطب.
- **ماذا استفدنا؟** حقق نظامنا 100% (1.0)، مما يثبت أن الـ System Prompt الخاص بنا مكتوب بعبقرية ويجبر الموديل على الالتزام الحرفي بالمصادر.

7. مناعة الضوضاء (Noise Robustness)

- نقوم بحقن النظام بفقرات كاذبة ومضللة (مثلاً: "البندول يعالج السرطان") ونرى هل سينخدع .(Adversarial) **ما هو؟** اختبار هجومي بها أم لا.
- **فائدته:** يضمن أن النظام لا يتبع أي هراء يُقدم له، بل يفلتر المعلومات.
- **ماذا استفدنا؟** حقق نظامنا 100% نجاح، الموديل تجاهل كل الضوضاء الكاذبة وركز فقط على الحقائق.

ملخص الإنجاز:

لم نقم فقط ببرمجة شات بوت، بل قمنا بـ ،(Google و OpenAI في العالم كـ AI التي تستخدمها كبرى شركات الـ) باستخدام هذه المقاييس **هندسة نظام طبي آمن، ومختبر، ومثبت كفاءته بالأرقام الرياضياتية الدقيقة**.