แน่นอนครับ เป็นคำถามที่ดีมากครับ การเลือกใช้ Metric เหล่านี้เพื่อวัดประสิทธิภาพของ RAG Output เป็นวิธีที่ถูกต้องและเป็นมาตรฐานครับ แต่ต้องเข้าใจก่อนว่าแต่ละตัวมี "มุมมอง" ในการให้คะแนนที่แตกต่างกัน
เหมาะสมควรเป็นเท่าไหร่ครับ

ภาพรวม: ทำไมต้องใช้ทั้ง 3 ตัว?

ลองนึกภาพว่าเรามี "คำตอบต้นแบบ" (Reference/Ground Truth) ที่ถูกต้อง 100% แล้วเราให้ AI ของเราตอบคำถามเดียวกันออกมา (Generated Text)

- **BLEU จะเป็นเหมือน "ผู้ตรวจที่เข้มงวด"**: สนใจว่า AI ใช้คำศัพท์และวลีที่ ตรงกับต้นแบบเป๊ะๆ แค่ไหน
    - **ROUGE จะเป็นเหมือน "ผู้ตรวจที่เน้นความครบถ้วน"**: สนใจว่า AI เก็บ ใจความสำคัญทั้งหมดจากต้นแบบมาได้ครบหรือไม่
    เป๊ะๆ แต่สนว่าประโยคที่ AI สร้างขึ้นมานั้น มีความหมายเหมือนกับต้นแบบ หรือไม่

การใช้ทั้ง 3 ตัวจะทำให้เราเห็นภาพรวมประสิทธิภาพของ AI ได้ครบทุกมิติ ทั้งความแม่นยำในการใช้คำ, ความครบถ้วนของข้อมูล, และความสามารถในการเข้าใจและสื่อสารความหมาย

1. BLEU (Bilingual Evaluation Understudy)

- ทำไมถึงใช้?
  BLEU ถูกสร้างมาเพื่องานแปลภาษาโดยเฉพาะ มันจึงเก่งในการวัด ความแม่นยำ (Precision) และความสละสลวยของประโยคที่สร้างขึ้นใหม่เทียบกับต้นฉบับ
  วัดว่ากลุ่มคำ (n-grams) ที่ AI สร้างขึ้นมานั้น ปรากฏอยู่ใน "คำตอบต้นแบบ" มากน้อยแค่ไหน ยิ่งมีกลุ่มคำที่ซ้ำกันเยอะ (เช่น คำ, วลี 2 คำ, วลี 3 คำ) คะแนนก็จะยิ่งสูง และ BLEU จะมีตัวลงโทษ (Brevity Penalty) หากประโยคที่ AI สร้างนั้นสั้นเกินไป
    - จุดเด่นในบริบท RAG:
  เหมาะมากสำหรับระบบที่ต้องการความถูกต้องของ "คำศัพท์เฉพาะทาง" เช่น ในระบบ Smart Tax Assistant ของคุณ การที่ AI ใช้คำว่า "ลดหย่อนภาษี" หรือ "กองทุน RMF" ได้อย่างถูกต้องและตรงกับต้นแบบ จะทำให้ BLEU ให้คะแนนสูง ซึ่งสำคัญมากสำหรับความน่าเชื่อถือ

2. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- ทำไมถึงใช้?
  ROUGE ถูกสร้างมาเพื่องานสรุปความ (Summarization) ดังนั้นมันจึงเก่งในการวัด ความครอบคลุมของเนื้อหา (Recall)
    - วัดอะไร?

  วัดว่าคำต่างๆ ใน "คำตอบต้นแบบ" ถูกนำมาใส่ไว้ในคำตอบของ AI ครบถ้วนมากน้อยแค่ไหน มันไม่ค่อยสนใจว่า AI จะเรียบเรียงประโยคใหม่ยังไง ตราบใดที่เนื้อหาสำคัญจากต้นฉบับยังอยู่ครบ ROUGE ก็จะให้คะแนนสูง

    - **ROUGE-1, ROUGE-2**: วัดการซ้ำกันของคำเดี่ยวๆ และวลี 2 คำ
    - **ROUGE-L**: วัดลำดับคำที่เหมือนกันที่ยาวที่สุด (Longest Common Subsequence) ซึ่งช่วยประเมินความต่อเนื่องของใจความได้ดี

- จุดเด่นในบริบท RAG:
  สำคัญมากในการเช็คว่า AI ไม่ได้ลืมให้ข้อมูลที่จำเป็น เช่น คำตอบต้นแบบบอกว่า "RMF ลดหย่อนได้ไม่เกิน 500,000 บาท และต้องถือครอง 5 ปี" ROUGE จะช่วยตรวจสอบว่า AI ได้พูดถึงทั้ง "วงเงิน" และ "เงื่อนไขเวลา" ครบถ้วนหรือไม่

3. BERTScore

- ทำไมถึงใช้?
  BLEU และ ROUGE มีจุดอ่อนคือไม่เข้าใจ ความหมาย (Semantics) มันมองว่าคำว่า "ราชา" กับ "กษัตริย์" เป็นคนละคำกัน BERTScore ถูกสร้างขึ้นมาเพื่อแก้ปัญหานี้โดยเฉพาะ
  แทนที่จะนับคำที่ซ้ำกัน BERTScore จะแปลงคำในประโยคของ AI และประโยคต้นแบบให้เป็น Vector (ชุดตัวเลขที่แทนความหมาย) ด้วยโมเดล BERT แล้วนำ Vector นั้นมาเทียบกันว่ามีความหมายใกล้เคียงกันแค่ไหน
    จุดเด่นในบริบท RAG:
  นี่คือตัววัดที่ทรงพลังที่สุดสำหรับแอปพลิเคชันยุคใหม่ หากคำตอบต้นแบบคือ "การลงทุนนี้ให้ผลตอบแทนสูงสุด" และ AI ตอบว่า "การลงทุนแบบนี้จะทำให้คุณได้กำไรมากที่สุด" BLEU/ROUGE จะให้คะแนนต่ำ แต่ BERTScore จะรู้ว่าทั้งสองประโยคมีความหมายเหมือนกัน และจะให้คะแนนสูงมาก

ค่าประมาณเท่าไหร่ถึงจะเรียกว่า "แม่นยำ"?

คำเตือนที่สำคัญที่สุด: ไม่มี "ตัวเลขมหัศจรรย์" ที่ใช้ได้กับทุกงาน ค่าคะแนนจะแปรผันสูงมากตามความซับซ้อนของงานและคุณภาพของ "คำตอบต้นแบบ"
อย่างไรก็ตาม นี่คือเกณฑ์โดยประมาณที่คุณสามารถใช้เป็นจุดอ้างอิงได้:

Metric	😟 **แย่ (Needs**	🙂 **พอใช้ (Acceptable)**	😀 **ดี (Good)**	😁 **ดีมาก (Excellent)**

	Work)			
BLEU	< 0.3	0.3 - 0.4	0.4 - 0.6	> 0.6
ROUGE-L	< 0.3	0.3 - 0.5	0.5 - 0.7	> 0.7
BERTScore	< 0.85	0.85 - 0.88	0.88 - 0.91	> 0.91

ข้อสังเกต:

- BERTScore มักจะมีค่าสูงกว่าตัวอื่น เพราะมันวัดความหมาย ไม่ใช่แค่การสะกดคำ
    - **สิ่งสำคัญกว่าการยึดติดกับตัวเลข คือการใช้คะแนนเหล่านี้เพื่อ "เปรียบเทียบ"** เช่น โมเดล A ได้ ROUGE-L ที่ 0.55 แต่พอเราปรับ Prompt ใหม่ในโมเดล B แล้วคะแนน ROUGE-L เพิ่มเป็น 0.62 **นี่คือสัญญาณที่ชัดเจนว่าการปรับปรุงของคุณได้ผล**

สรุป: สำหรับระบบของคุณ การใช้ **BERTScore** เป็นตัวชี้วัดหลักเพื่อดูความถูกต้องทางความหมาย และใช้ **ROUGE-L** เป็นตัวเสริมเพื่อเช็คว่า AI ไม่ได้ลืมเงื่อนไขหรือข้อมูลสำคัญ จะเป็นแนวทางที่ทรงพลังและเหมาะสมที่สุดครับ