R2 RAGAS Evaluation

Final Project - BLM5109 Collective Learning

1st PhD Student Önder GÖRMEZ
Computer Engineering
Faculty of Electrical and Electronics
Yıldız Technical University
İstanbul, Türkiye
ondergormez@gmail.com

2th Prof. Dr. M. Fatih AMASYALI
Computer Engineering
Faculty of Electrical and Electronics
Yıldız Technical University
İstanbul, Türkiye
amasyali@yildiz.edu.tr

***Abstract*—Bu makalede Türkçe veri setleri üzerinde eğitilmiş 2 modelin performanslarının karşılaştırılması için çalışmalar yapılmış ve sonuçları raporlanmıştır.**

***Index Terms*—RAGAS Scores, RAGAS Evaluation, Answer Correctness, Answer Relevancy, Faithfulness**

I. Introduction

Makalede 2. bölümde Retrieval Augmented Generation Evaluation Metrics konusuna kısaca değinilecek ve kullanılan metrikler verilecek, 3. bölümde Veri Seti tanıtılacak, 4. bölümde Context Uzunluğunun LLM'in Cevabına Etkisi üzerine yapılan çalışmalar aktarılacak, 5. bölümde Context İçerisinde Doğru Cevabın Yerinin LLM'in Cevabına Etkisi üzerine yapılan çalışmalar aktarılacak ve son olarak 6. bölümde Sonuçlar ve Gelecek Çalışmalar özetlenecektir.

II. Retrieval Augmented Generation Evaluation Metrics

Bu çalışma kapsamında RAGAS metriklerinden 3 tanesi kullanılacaktır.

- Faithfulness
- Answer Correctness
- Answer Relevancy

Yukarıda verilen 3 metriğin ne olduğu aşağıda açıklanmıştır.

A. Faithfulness

Faithfulness metriği, llm in oluşturduğu yanıtın db den retrival edilen (çekilen) context ile ne kadar tutarlı olduğunu ölçer. 0 ile 1 arasında değişir ve daha yüksek puanlar daha iyi bir sonucu işaret eder. LLM tarafından üretilen yanıtta ortaya atılan iddialar, db den çekilen context ile desteklenebiliyorsa tutarlı olarak kabul edilir [12].

Bunu hesaplamak için:

- Yanıttaki ortaya atılan tüm iddiaları belirlenir.
- Her iddianın, context ten (bağlamdan) çıkarılıp çıkarılamayacağını görmek için kontrol edilir.
- Ve aşağıda belirtilen tutarlılık puanı hesaplanır: [12]

$$\text{Faithfulness Score} = \frac{\text{LLM yanıtında ve context'te eşleşen iddaalar}}{\text{LLM yanıtındaki toplam iddia sayısı}}$$

B. Answer Correctness

Cevap doğruluğunun değerlendirilmesi, llm tarafında oluşturulan cevabın ground truth ile karşılaştırılması ile ölçülür.Puanlar 0 ile 1 arasında değişir. Daha yüksek bir puan, oluşturulan cevap ile ground truth arasında daha yakın bir hizalanma olduğunu ve cevabın daha iyi olduğunu ifade eder [13].

Aşağıdaki şekilde bir örnek verilebilir:

- Ground truth: Einstein was born in 1879 in Germany.
- High answer correctness: In 1879, Einstein was born in Germany.
- Low answer correctness: Einstein was born in Spain in 1879 [13].

TP (True Positive): Hem ground truth hem de llm tarafından oluşturulan cevapta bulunan gerçekler veya ifadeler.

FP (False Positive): LLM tarafından oluşturulan cevapta bulunan ancak ground truth ta bulunmayan ifadeler veya iddialar.

FN (False Negative): LLM tarafından oluşturulan cevapta bulunmayan ancak ground truth ta bulunan ifadeler veya iddialar.

Yukarıdaki bilgilere aşağıdaki şekilde bir örnek verilebilir:

- TP: [Einstein was born in 1879]
- FP: [Einstein was born in Spain]
- FN: [Einstein was born in Germany] [13].

$$\text{F1 Score} = \frac{|\text{TP}|}{|\text{TP}| + 0.5 \times (|\text{FP}| + |\text{FN}|)}$$

C. Answer Relevancy

Answer Relevancy metriği ile, llm tarafından üretilen bir yanıtın kullanıcı girdisiyle ne kadar alakalı olduğu ölçülür. Daha yüksek puanlar kullanıcı girdisiyle daha iyi uyumu gösterirken, yanıt eksikse veya gereksiz/fazla bilgiler içeriyorsa daha düşük puanlar verilir [11].

Bu metrik, kullanıcı tarafından girilen soru ve llm'in yanıtı kullanılarak aşağıdaki şekilde hesaplanır:

- LLM yanıtına dayalı yapay bir soru kümesi oluşturulur.
- Bu sorular, LLM'in verdiği yanıtın içeriğini yansıtacak şekilde tasarlanmıştır.

- Kullanıcı tarafından girilen soru için word embedding yöntemi ile embedding oluşturulur.
- Yapay soru kümesinin de embedding leri oluşturulur.
- Yapay soru kümesi için oluşturulan embedding ler ile kullanıcının girdiği sorular arasında kosinüs benzerliğine dayalı (cosine similarity) bir benzerlik hesabı yapılır.
- Tüm sorular için oluşan benzerlik değerinin ortalaması answer relevancy değeri olarak kabul edilir [11].

$$\text{Answer Relevancy} = \frac{1}{N}\sum_{i=1}^{N}\text{cosine similarity}(E_{g_i}, E_o)$$

$$\text{Answer Relevancy} = \frac{1}{N}\sum_{i=1}^{N}\frac{E_{g_i}\cdot E_o}{\|E_{g_i}\|\|E_o\|}$$

Where:

- E_{g_i}: Embedding of the i^{th} generated question.
- E_o: Embedding of the user input.
- N: Number of generated questions (default is 3) [11].

III. Veri Seti

Bu çalışma kapsamında Hugging Face üzerinde yayınlanmış ve Türkçe dilinde olan LLM'lerin RAGAS metriklerinin ölçümü için hazırlanmış Metin/WikiRAG-TR veri seti kullanılmıştır [1].

A. Veri Setinin Özellikleri

Dataset Özellikleri:

- Örnek Sayısı: 5999 (5725 sentetik olarak oluşturulmuş soru-cevap çifti, 274 tane context içerisinde sorunun cevabı olmayan negatif örnek)
- Veri Seti Boyutu: 20,5 MB
- Dil: Türkçe

Veri Setinin Kolonları:

- id: Her soru için unique (eşsiz) olan bilgi
- question: Kullanıcı tarafından girilen soru
- answer: Ground truth dediğimiz aslında LLM tarafından üretilmesi beklenen doğru cevap
- context: İçerisinde bir veya birden fazla paragraf barındıran ve LLM'in cevap üretmek için kullandığı bağlam
- `is_negative_response`: Context içerisinde bulunan cevaplar arasında doğru cevap olup olmadığını gösteren flag. Bu değer veri setindeki çoğu data için 0 değerinde.
- `number_of_articles`: Context içerisinde kaç tane paragraf olduğu bilgisi
- `ctx_split_points`: Paragrafların hangi context in içerisinde hangi index ten başladığı bilgisi
- `correct_intro_idx`: Context içerisinde sorunun cevabı olan paragrafın başladığı index bilgisi. Negatif yani context içerisinde cevabın olmadığı durumda -1 değerini alır.

IV. Context Uzunluğunun LLM'in Cevabına Etkisi

Bu bölümde yukarıda detayları verilen veri seti üzerinde 2 LLM modeli için context uzunluğunun oluşturulan cevaba etkisi incelenecektir. Bu aşamada veri setin bulunan random 100 soru için RAGAS evaluation yapılacaktır. Context uzunlukları sırasıyla 1, 5, 10 ve 15 olarak seçilecek ve bu context lerin içerisinde doğru cevabın konumu tam ortada olacaktır. Context hazırlanırken doğru cevaplar sırasıyla 1, 3, 5 ve 8. index lerde olacaktır.

A. Cosmos - Mean Faithfulness

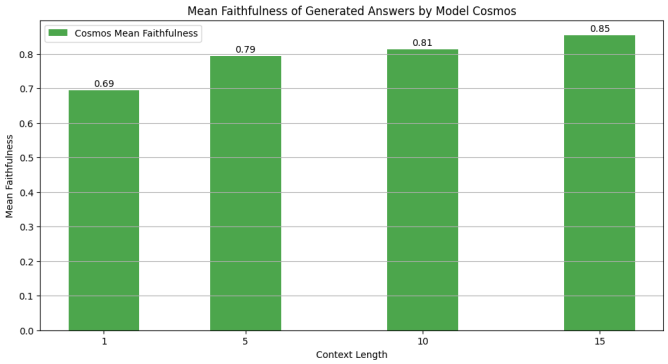

Fig. 1. Cosmos - Mean Faithfulness

Fig. 1 den anlaşılacağı üzere Cosmos LLM modeli için faithfulness değerinin context büyüdükçe arttığı görülmektedir.

B. Gemma - Mean Faithfulness

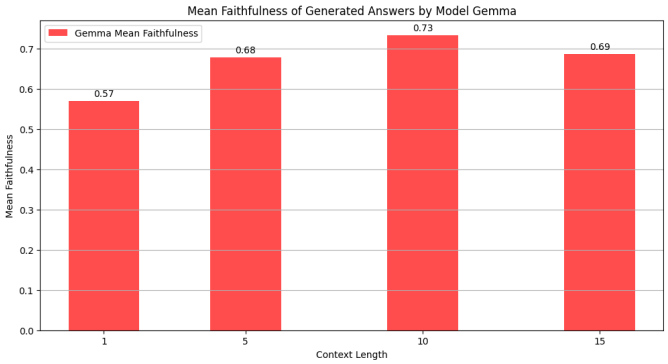

Fig. 2. Gemma - Mean Faithfulness

Fig. 2 den anlaşılacağı üzere Gemma LLM modeli için faithfulness değerinin context büyüdükçe arttığı fakat context boyut 10 dan sonra 15 olduğunda azaldığı görülmektedir.

C. Cosmos vs Gemma - Mean Faithfulness

Fig. 3 den anlaşılacağı üzere Cosmos modeli Gemma modeline göre daha iyi bir faithfulness performansı sergilemektedir.

D. Cosmos - Answer Correctness

Fig. 4 den anlaşılacağı üzere Cosmos LLM modeli için answer correctness değerinin context büyüdükçe azaldığı görülmektedir.

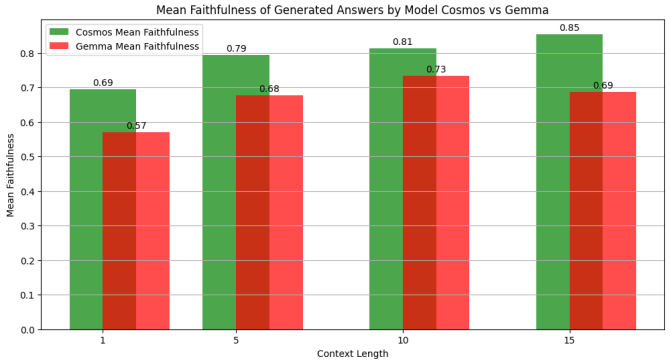

Fig. 3. Cosmos vs Gemma - Mean Faithfulness

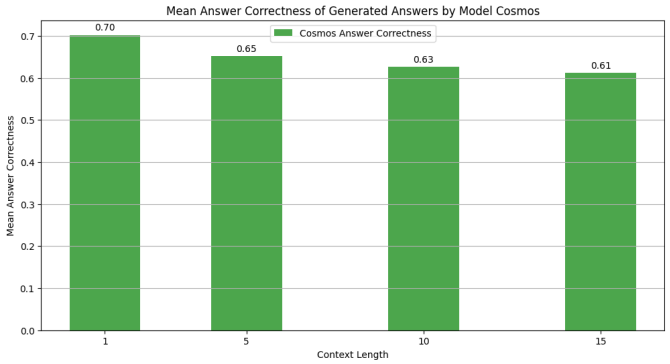

Fig. 4. Cosmos - Mean Answer Correctness

E. Gemma - Mean Answer Correctness

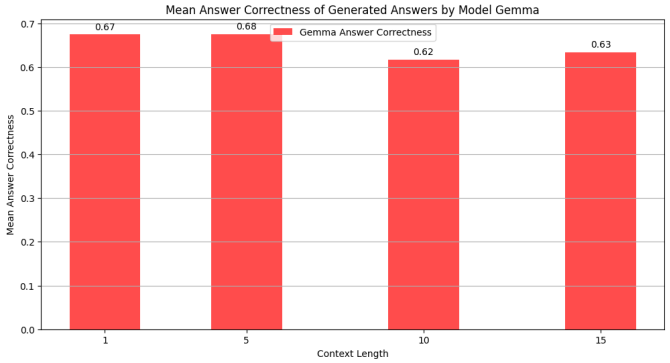

Fig. 5. Gemma - Mean Answer Correctness

Fig. 5 den anlaşılacağı üzere Gemma LLM modeli için için answer correctness değerinin genel olarak context büyüdükçe azaldığı görülmektedir.

F. Cosmos vs Gemma - Mean Answer Correctness

Fig. 6 den anlaşılacağı üzere Cosmos ve Gemma modeli answer correctness metriğinde birbirine yakın bir performans sergilemektedir.

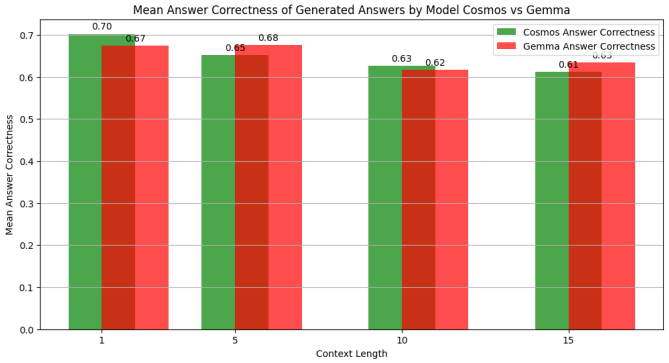

Fig. 6. Cosmos vs Gemma - Mean Answer Correctness

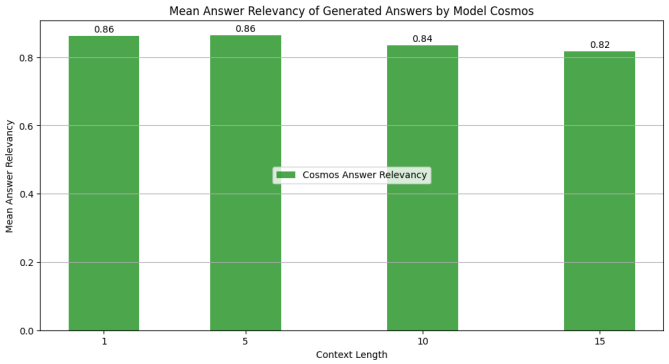

Fig. 7. Cosmos - Mean Answer Relevancy

G. Cosmos - Answer Relevancy

Fig. 7 den anlaşılacağı üzere Cosmos LLM modeli için answer relevancy değerinin context büyüdükçe azaldığı görülmektedir.

H. Gemma - Mean Answer Relevancy

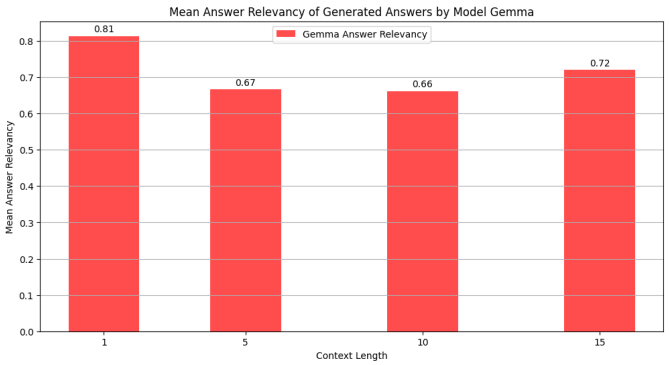

Fig. 8. Gemma - Mean Answer Relevancy

Fig. 8 den anlaşılacağı üzere Gemma LLM modeli için için answer relevancy değerinin genel olarak context büyüdükçe azaldığı görülmektedir.

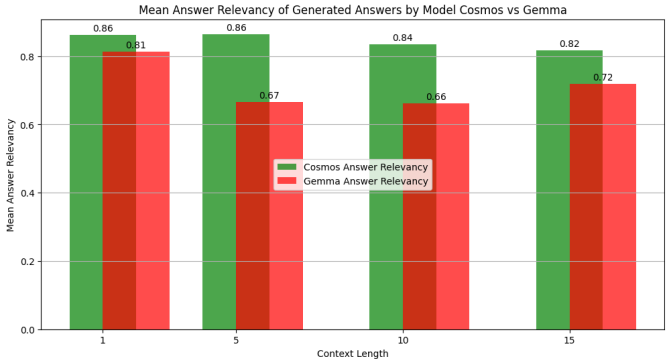

Fig. 9. Cosmos vs Gemma - Mean Answer Relevancy

I. *Cosmos vs Gemma - Mean Answer Relevancy*

Fig. 9 den anlaşılacağı üzere Cosmos modeli Gemma modeline göre answer relevancy metriğinde çok daha iyi bir performans sergilemektedir.

V. Context İçerisinde Doğru Cevabın Yerinin LLM'in Cevabına Etkisi

Bu bölümde veri seti üzerinde 2 LLM modeli için context içerisinde doğru cevabın bulunduğu yerin LLM'lerin oluşturduğu cevaba etkisi incelenecektir. Bu aşamada veri setinde bulunan random 50 soru için RAGAS evaluation yapılacaktır.

15 tane paragraftan oluşan sabit bir context seçilmiştir. Bu context içerisinde doğru cevabın bulunduğu index sırasıyla 1, 8 ve 15 olacak şekilde değiştirilerek performans ölçümü yapılacaktır.

A. *Cosmos - Mean Faithfulness*

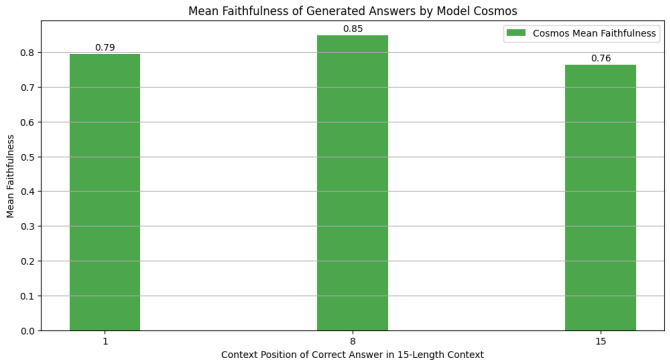

Fig. 10. Cosmos - Mean Faithfulness

Fig. 10 den anlaşılacağı üzere Cosmos LLM modeli için faithfulness değerinin doğru cevabın tam ortada olması durumunda daha yüksek çıktığı görülmektedir.

B. *Gemma - Mean Faithfulness*

Fig. 11 den anlaşılacağı üzere Gemma LLM modeli için faithfulness değerinin doğru cevabın tam ortada olması durumunda daha yüksek çıktığı görülmektedir.

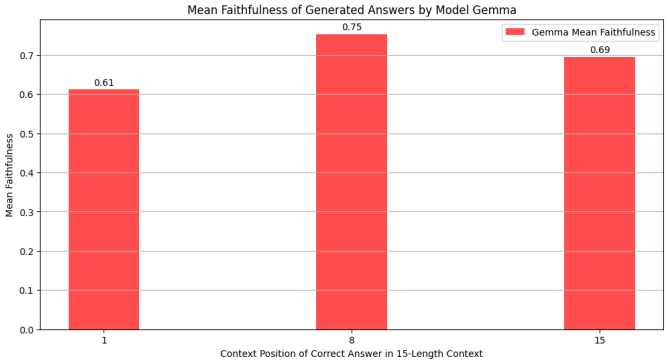

Fig. 11. Gemma - Mean Faithfulness

C. *Cosmos vs Gemma - Mean Faithfulness*

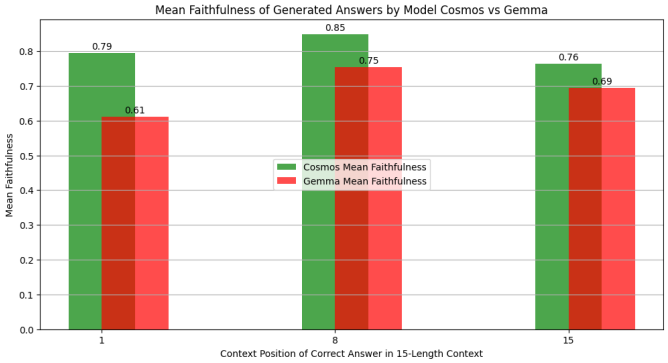

Fig. 12. Cosmos vs Gemma - Mean Faithfulness

Fig. 12 den anlaşılacağı üzere Cosmos modeli Gemma modeline göre daha iyi bir faithfulness performansı sergilemektedir.

D. *Cosmos - Answer Correctness*

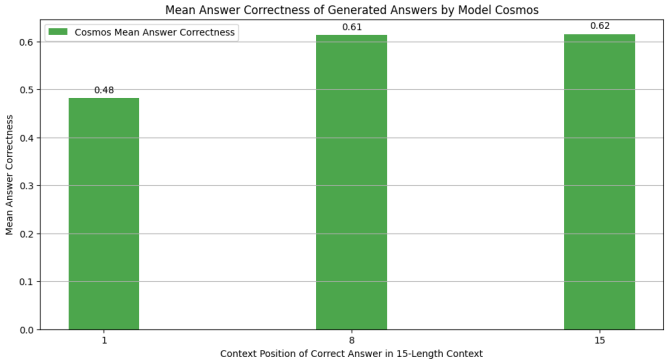

Fig. 13. Cosmos - Mean Answer Correctness

Fig. 13 den anlaşılacağı üzere Cosmos LLM modeli için answer correctness değerinin doğru cevabın ortaya doğru kayması durumunda arttığı görülmektedir. Fakat ortadan sona doğru gitmesi çok fazla bir performans artışına neden olmamaktadır.

E. Gemma - Mean Answer Correctness

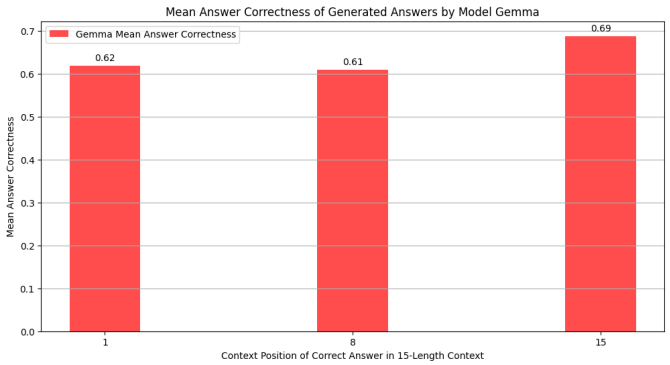

Fig. 14. Gemma - Mean Answer Correctness

Fig. 14 den anlaşılacağı üzere Gemma LLM modeli için için answer correctness değerinin doğru cevabın sonda olduğu context için daha iyi çıktığı görülmektedir.

F. Cosmos vs Gemma - Mean Answer Correctness

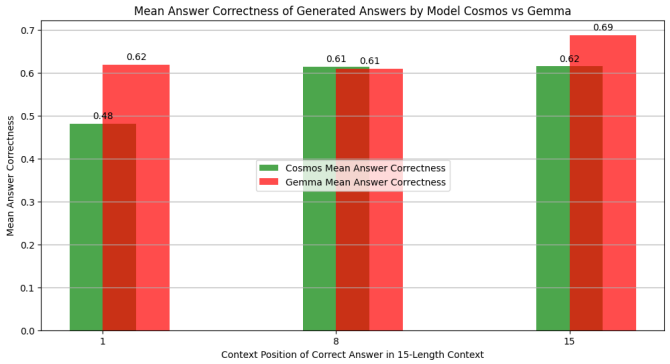

Fig. 15. Cosmos vs Gemma - Mean Answer Correctness

Fig. 15 den anlaşılacağı üzere Gemma modeli answer correctness metriğinde Cosmos modeline göre daha iyi bir performans sergilemektedir.

G. Cosmos - Answer Relevancy

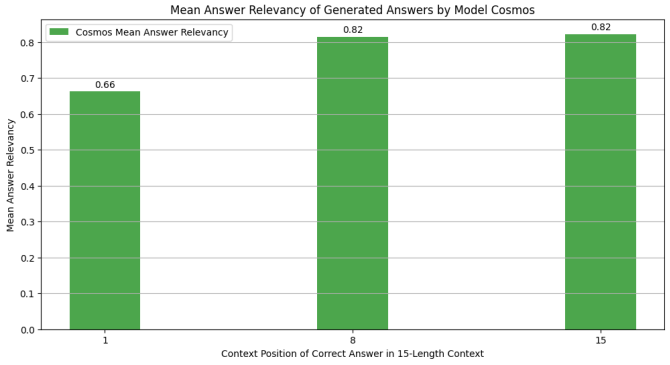

Fig. 16. Cosmos - Mean Answer Relevancy

Fig. 16 den anlaşılacağı üzere Cosmos LLM modeli için answer relevancy değerinin doğru cevabın ortaya doğru kayması durumunda arttığı görülmektedir. Fakat ortadan sona doğru gitmesi çok fazla bir performans artışına neden olmamaktadır.

H. Gemma - Mean Answer Relevancy

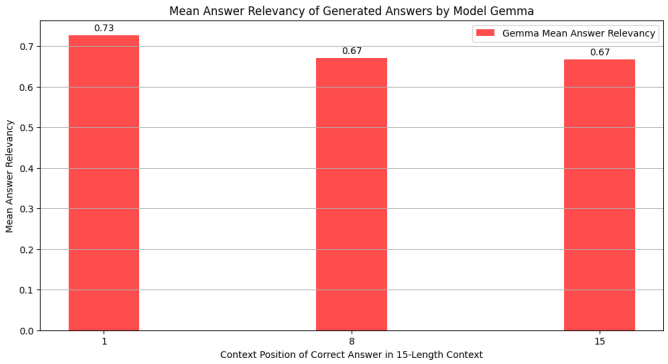

Fig. 17. Gemma - Mean Answer Relevancy

Fig. 17 den anlaşılacağı üzere Gemma LLM modeli için için answer relevancy değerinin genel olarak context içerisinde doğru cevap ortaya doğru gittiğinde azaldığı görülmektedir. Fakat ortadan sona doğru gitmesi sonucu çok değiştirmemiştir.

I. Cosmos vs Gemma - Mean Answer Relevancy

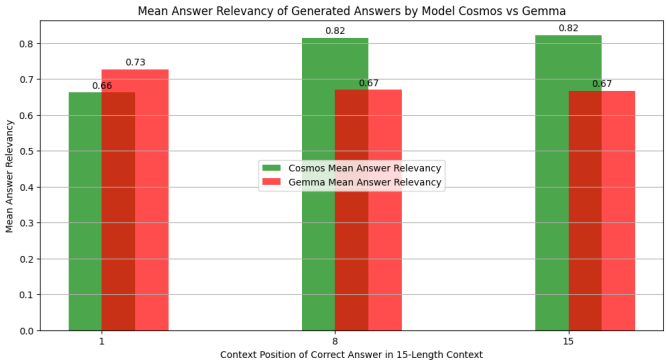

Fig. 18. Cosmos vs Gemma - Mean Answer Relevancy

Fig. 18 den anlaşılacağı üzere Cosmos modeli ortadan ve sonda daha başarılı iken Gemma modeli doğru cevap context in başında iken daha iyi bir performans sergilemektedir.

VI. Sonuçlar ve Gelecek Çalışmalar

Farklı context uzunlukları için:

- Cosmos modeli Gemma modeline göre daha iyi bir faithfulness performansı sergilemektedir.
- Cosmos modeli Gemma modeline göre answer relevancy metriğinde çok daha iyi bir performans sergilemektedir.

15 uzunluklu context içerisinde doğru cevapların farklı konumda olması durumunda:

- Cosmos modeli Gemma modeline göre daha iyi bir faithfulness performansı sergilemektedir.

- Cosmos modeli Gemma modeline göre answer relevancy metriğinde daha iyi bir performans sergilemektedir.

Gelecekte yapılabilecek çalışmalar aşağıdaki gibidir;

- Makalede kullanılan 3 RAGAS metriği dışında daha farklı RAGAS metrikleri kullanılabilir.
- Ölçümlemelerin yapıldığı soru sayısı arttırılabilir.
- Context uzunluğu daha büyük değerlerde denemeler ve performan ölçümleri yapılabilir.

Bu çalışma ile ilgili yapılan geliştirmeler Kollektif Öğrenme dersine ait github reposunda paylaşılmıştır. Gelecekte yapılacak olan çalışmalarda kullanılabilmesi için detaylı bir şekilde kodun dokümantasyonu yapılmıştır [4].

ACKNOWLEDGMENT

Bu makalenin oluşturulmasında ve yayına hazır hale getirilmesinde bizi teşvik eden, geri bildirimlerini ve desteklerini esirgemeyen sayın hocamız Prof. Dr. M. Fatih AMASYALI'ya teşekkürü bir borç bilirim.

REFERENCES

[1] "Datasets: Metin/WikiRAG-TR" https://huggingface.co [Online]. Available: https://huggingface.co/datasets/Metin/WikiRAG-TR [Accessed: 20-Jan-2025].

[2] "ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1" https://huggingface.co [Online]. Available: ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1 [Accessed: 20-Jan-2025].

[3] "Metin/Gemma-2-9b-it-TR-DPO-V1" https://huggingface.co [Online]. Available: https://huggingface.co/Metin/Gemma-2-9b-it-TR-DPO-V1 [Accessed: 20-Jan-2025].

[4] "Github - Önder Görmez" https://github.com [Online]. Available: https://github.com/ondergormez/BLM5109_Collective_Learning/tree/main/03-Project [Accessed: 20-Jan-2025].

[5] "Overview of Metrics" https://docs.ragas.io [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/overview/ [Accessed: 20-Jan-2025].

[6] "List of available metrics" https://docs.ragas.io [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/ [Accessed: 20-Jan-2025].

[7] "Context Precision" https://docs.ragas.io [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/context_precision/ [Accessed: 20-Jan-2025].

[8] "Context Recall" https://docs.ragas.io [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/context_recall/ [Accessed: 20-Jan-2025].

[9] "Context Entities Recall" https://docs.ragas.io [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/context_entities_recall/ [Accessed: 20-Jan-2025].

[10] "Noise Sensitivity" https://docs.ragas.io [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/noise_sensitivity/ [Accessed: 20-Jan-2025].

[11] "Response Relevancy or Answer Relevancy" https://docs.ragas.io [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/answer_relevance/ [Accessed: 20-Jan-2025].

[12] "Faithfulness" https://docs.ragas.io [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/faithfulness/ [Accessed: 20-Jan-2025].

[13] "Answer Correctness" https://docs.ragas.io [Online]. Available: https://docs.ragas.io/en/v0.1.21/concepts/metrics/answer_correctness.html [Accessed: 20-Jan-2025].

[14] "Multi modal faithfulness" https://docs.ragas.io [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/multi_modal_faithfulness/ [Accessed: 20-Jan-2025].

[15] "Multi modal relevance" https://docs.ragas.io [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/multi_modal_relevance/ [Accessed: 20-Jan-2025].

[16] "Chat Templates" https://huggingface.co [Online]. Available: https://huggingface.co/docs/transformers/main/en/chat_templating [Accessed: 20-Jan-2025].