Enerji İstatistik Notu 77: Yapay Zeka Ne Kadar Enerji Tüketir

Tek cümle: "*Bireysel olarak arama-okuma ve sonuçları bulma ile 4 dakika süren bir işlev, eğer üretken yapay zeka ile 1 dakika sürüyor ise o zaman üretken yapay zeka daha verimli olmaktadır*"

Barış Sanlı, barissanli2@gmail.com

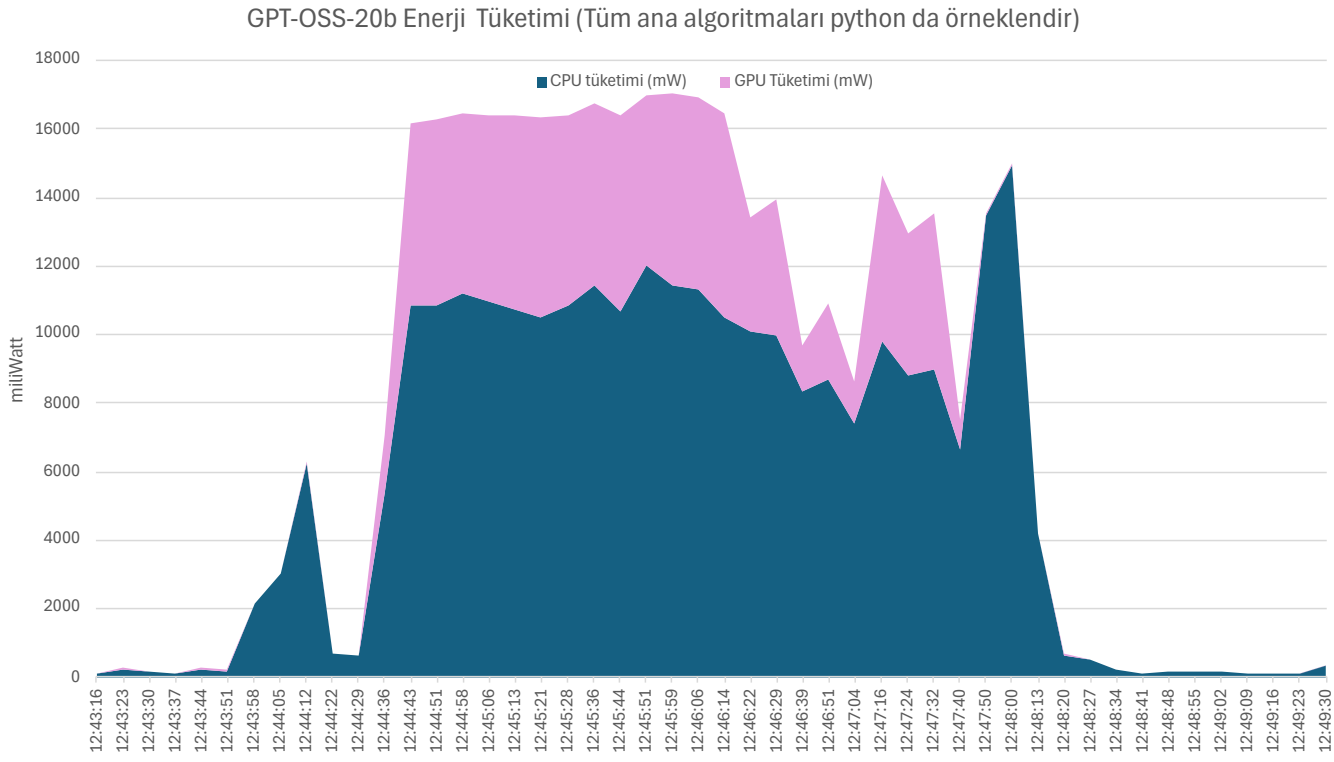
Excel: http://github.com/barissanli/ein

Yapay zekanın ne kadar enerji tükettiği ve bunun farklı cihazlardaki ayakizinin ne kadar olduğu bu notun konusudur. Tüm veriler kendi ölçümlerim olup, ana mesaj "arama başına değil ama görev başına yapay zekanın daha verimli olduğu ve bu verimin daha çok kullanım getireceği"dir. Görev olarak ise Python kodlamada ana algoritmalara örnek verilmesi talep edilmiştir.

Genel olarak laptop pili 75 Wh, cep telefonu pili de 20-25 Wh olarak kabul edilebilir. Laptop

OpenAI'ın gpt-oss:20b (20 milyar parametreli) modelinin yayınlanması ile birlikte, daha tanıdık bir modelin sıradan bilgisayar üzerinde çalıştırılması mümkün olmuştur. Yazılım olarak LM Studio ile yerel olarak çalıştırılan modelde, GPU (grafik işlemci)'de mümkün olan en üst seviyede çalıştırılacak şekilde ayarlanmıştır. GPU tarafı daha çok paralel matriks işlemlerinde uzmanlaştığı için tüketimi sınırlı olabilmektedir. Fakat son dönemde bilgisayarlardaki ana "duvar", veriyi işlemci ve grafik işlemci ile hafıza arasında hareket ettirme hızıdır.

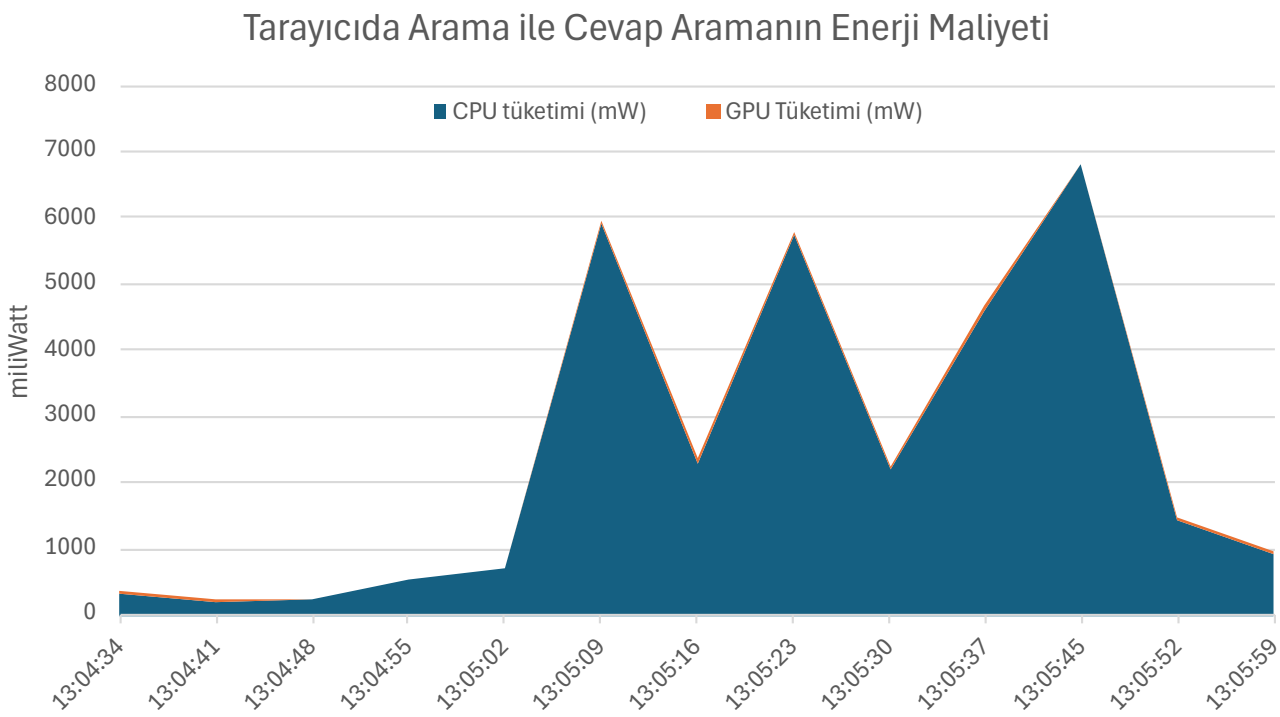
Verilen görevi yerine getirirken, bir hazırlık ve modelin hafızaya yüklenmesi, sonrasında işlem ve sonrasında da görevin sonlandırılıp hafızanın temizlenmesi işlemleri vardır. GPU tüketimi modelin aktif olarak çalıştığı dönemleri göstermektedir.



Temelde yerel çalışan bir GPT, nispeten küçük olmasına rağmen, 16 Watt'lık anlık enerji talebini getirmektedir. Yerel çalıştığı için ortalama 50-60 parça/saniyenin çok altında 12 parça/saniye seviyelerinde bir üretim olduğundan da 4 dakikaya yakın bir zamanda işlemi tamamlamaktadır.

Görev olarak verilen Python kodlama, doğru ve eksiksiz çalışmakla birlikte, 0.86Wh(watt-saat) enerji talebi ile sonuçlanmıştır. Eğer tüm hazırlama-temizleme de hesaba katılırsa 1.2 Wh enerji tüketimi olmuştur (16 Watt * 4 dakika/60 dakika = 1.06 Watt gibi).

Aynı bilgisayarda aynı işlem için arama yapıldığında ise, aramanın Google'daki enerji maliyet hariç, bilgisayarda tarayıcı ile gezinme ekran açma, okuma gibi sebeplerden yine 6 Watt-2 Watt arası enerji tüketimi olmaktadır. Bu 6.7 Watt'a kadar çıkarken, ortalamada 4 Watt gibi bir enerji tüketimine sebep olmaktadır. Kullanılan tarayıcı bu bilgisayardaki en verimli tarayıcıdır.



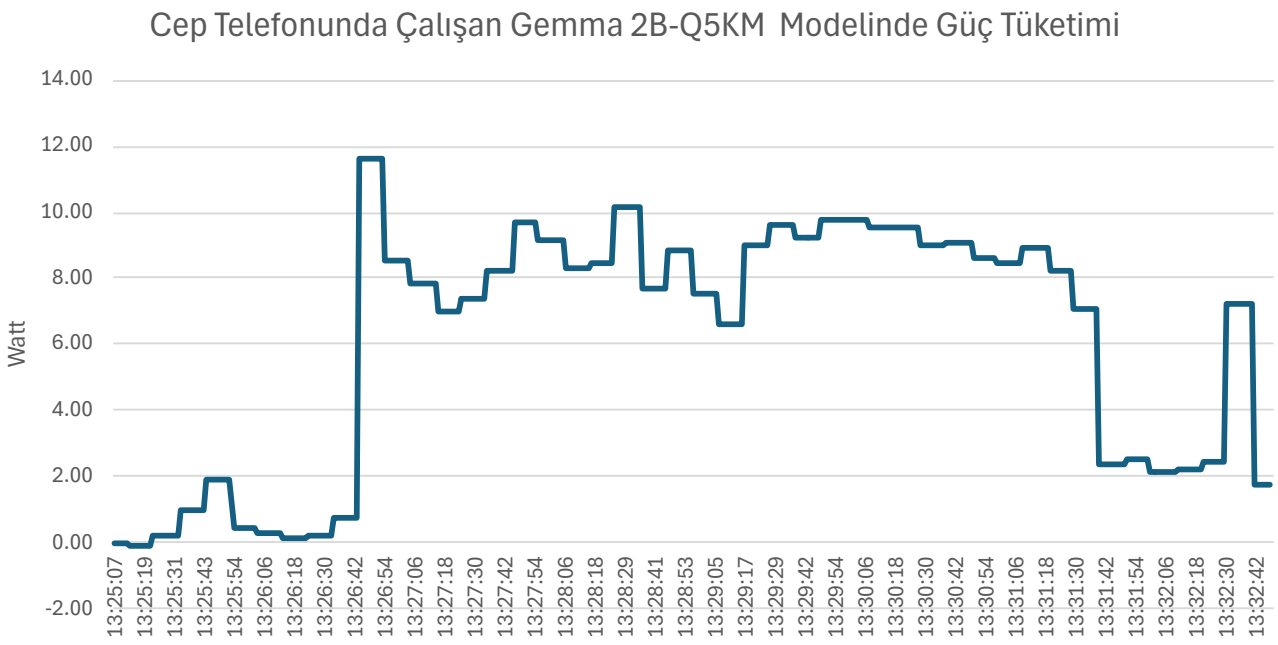
Temelde bir işi:

- Tarayıcıda arama ile yapmak ortalama 4 Watt ile arama süresi
- Tek seferde yapay zekada yapmak ise 16 Watt anlık güç tüketimi ile genelde 3-4 dakika sürmektedir.

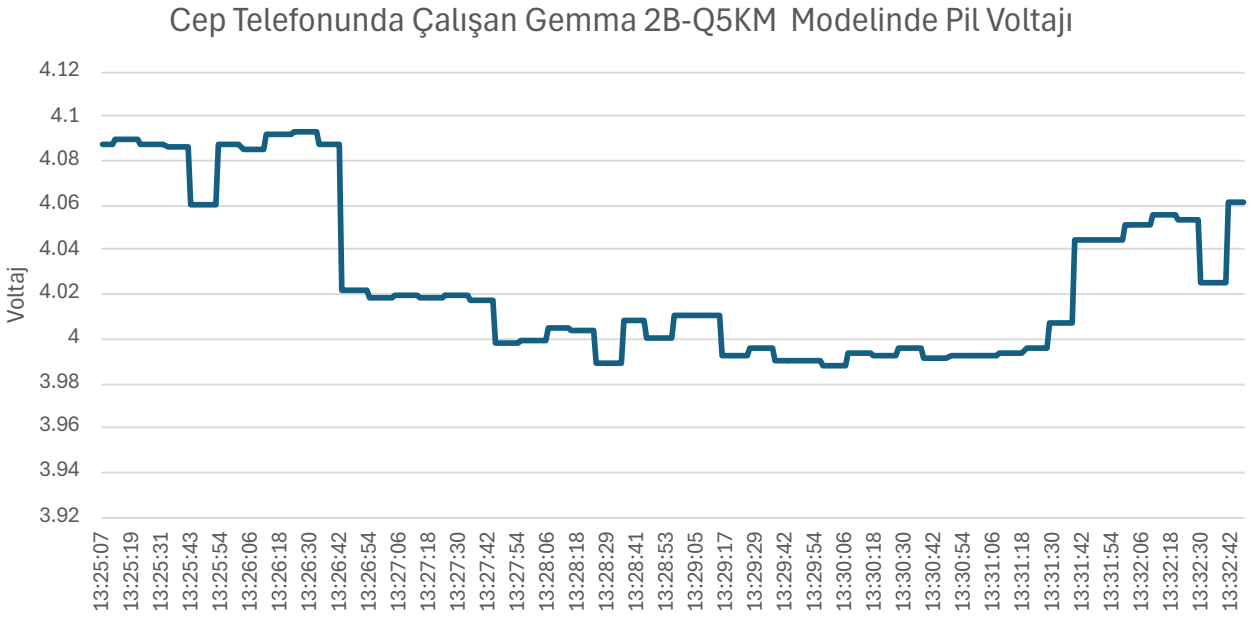
Yani enerji tüketimindeki ana belirsizlik bir görevin tamamlanması için gereken zamandır.

Son kısımda da cep telefonu üzerinde bir GPT daha doğrusu LLM çalıştırılmasındaki enerji tüketimi ölçülmüştür. Bu ölçümün daha doğru olması için USB üzerinden özel bir fonksiyon ile yapılmıştır. Benzer sonuçlar cihaz üstü ölçüm için kullanılan programlarla da teyit edilmiştir.

Cep telefonu pili 20-25 Wh denilmişti, bunun 4 Volt ve 5Ah(Amper saat) civarında olduğu kabul edilebilir. Anlık olarak cep telefonu 0.1-0.2 Ah boşta tüketmektedir. Ekran açıkken bu 0.8 Wh altına pek de sık düşmemektedir. Kısaca 0.4-0.8 Wh bir tüketim her hâlükârda vardır.



Model çalışması sırasında enerji talebi 12 sonra 10 Watt civarında pik yapmıştır. Burada telefonun ısınmasından dolayı bir limitleme (throttling) de olmuş olabilir. Bunun için de detaylı frekans verilerine bakmak gerekir. Fakat ortalamada 9 Watt ile benzer bir görev tamamlanabilmiştir. Bu sırada internette gezinmek ise 1-1.5 Watt enerji tüketmektedir.



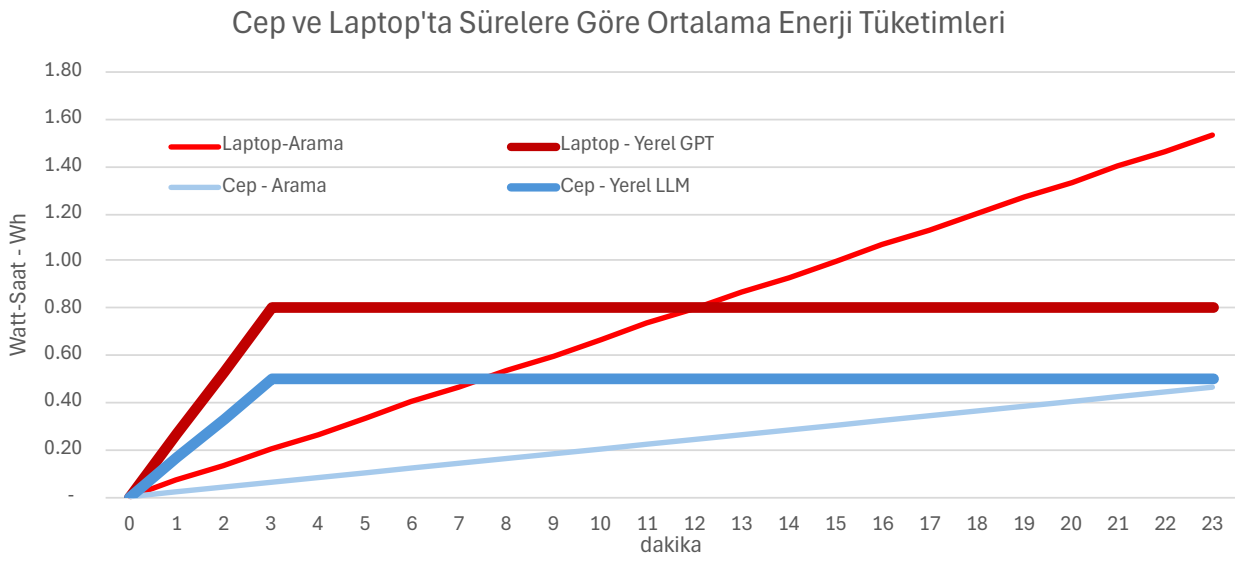
İlgilileri için yukarıdaki grafikte, güç çekimi ile birlikte pildeki voltajın nasıl düştüğünü göstermektedir. Aynı işlem için gereken güç için voltaj düştükçe de daha çok akım gerekmektedir. Voltaj aynı zamanda pil seviyesi bazı programlarda kalan %kapasiteyi gösterir.

Sonuçları bir araya getirirsek aşağıdaki tablo ile karşılaşırız. Tüm ölçümler de Excel tablosunda bulunabilir.

Method	Pik Güç Talebi (watt)	Ortalama Güç talebi (watt)	Süre
Laptop – Yerel GPT	17	16	4 min
Laptop -Arama	6.7	4	1 min
Mobile – Yerel LLM	12	9	3 min
Mobile – Arama	2	1.2	1 min

Farklı ortamlarda farklı yapay zekaların kullanımının maliyetleri ve sonuçları farklıdır. Tabii ki sonuç kalitesi de değişmektedir. Fakat ana nokta bir görevin nerede ne kadar sürede tamamlanabildiğidir. Bireysel olarak arama-okuma ve sonuçları bulma ile 4 dakika süren bir işlev, eğer üretken yapay zeka ile 1 dakika sürüyor ise o zaman üretken yapay zeka daha verimli olmaktadır. Çünkü daha önceki çalışmaların hesaba katmadığı ana konu, tarayıcının açık kaldığı ve arama yapıldığı her an 4 Watt civarında da kullanıcı tarafında enerji tüketimi olmaktadır.

Aşağıdaki grafik de bunu göstermektedir.



Önümüzdeki dönemde kullanıcılar kendi cihazlarında daha fazla yerel büyük dil modeli çalıştıracaklar. Arama, forumlara bakma, deneme, tekrar forumlara gitme gibi bir zinciri takip eden arama işlevlerinde çok vakit kaybedilebilmektedir. Ama GPT gibi uygulamalar "zaman paradır" bakışı ile inanılmaz bir zamansal verimlilik sağlamaktadır.

1865'te kömürdeki verimli kullanım ile talebin artacağını söyleyen Jevons'un paradoksu[1] gibi yapay zekanın temel "zaman kazandırma" işlevi, gelişen algoritma, çip ve model yapıları ile daha çok kullanılması beklenir. Çünkü bu modeller kesinlikle "zaman verimli", ayrıca arama ile dakikalarca sürecek işlemleri daha kısa sürede halledebildiği için de enerji verimli. Ama toplam tüketimde bu verim artan talep olarak geri dönecektir.

[1] https://en.wikipedia.org/wiki/Jevons_paradox