گزارش یافتن بهترین Learning Rate برای مدل GPT-2

هدف این آزمایش، ارزیابی تأثیر نرخ‌های یادگیری مختلف بر عملکرد یک مدل زبانی مبتنی بر معماری GPT-2 و تعیین مناسب‌ترین مقدار Learning Rate برای دستیابی به کمترین میزان Loss در طول فرآیند آموزش است.

۱. تنظیمات مدل و داده‌ها

مدل مورد استفاده : GPT-2

```
    model=GPTConfig(
        vocab_size=10_000,
        max_seq_len=1024,
        inlayer=8,
        head=16,
        n_embd=128,
        f_expnd=4),
```

Optimizer: Addams

```
class OptimizerConfig:
    max_lr: float = 6e-4
    betas: tuple = (0.9, 0.95)
    weight_decay: float = 0.1
    fused: bool = True
    warmup_steps: int = 265
    alpha: float = 0.1
```

Loss Function: CrossEntropyLoss

دیتاست : TinyStories

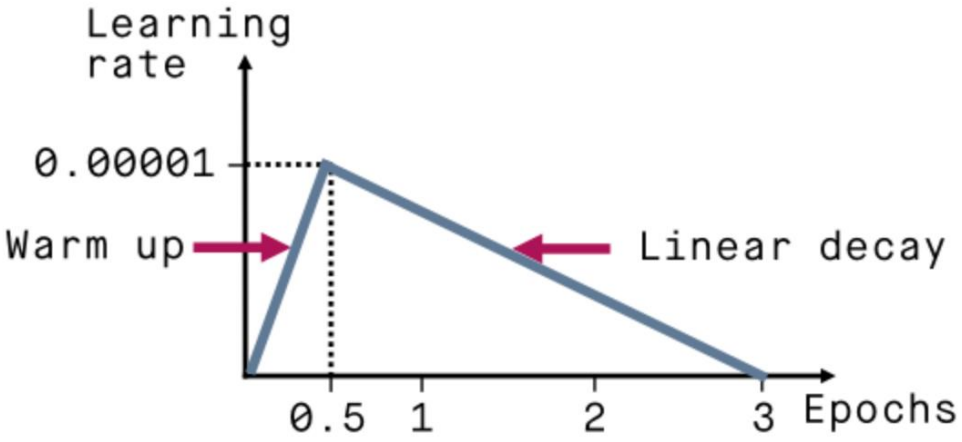
تعداد توکن: 65,000,000

Batch Size: 48

seq_len: 512

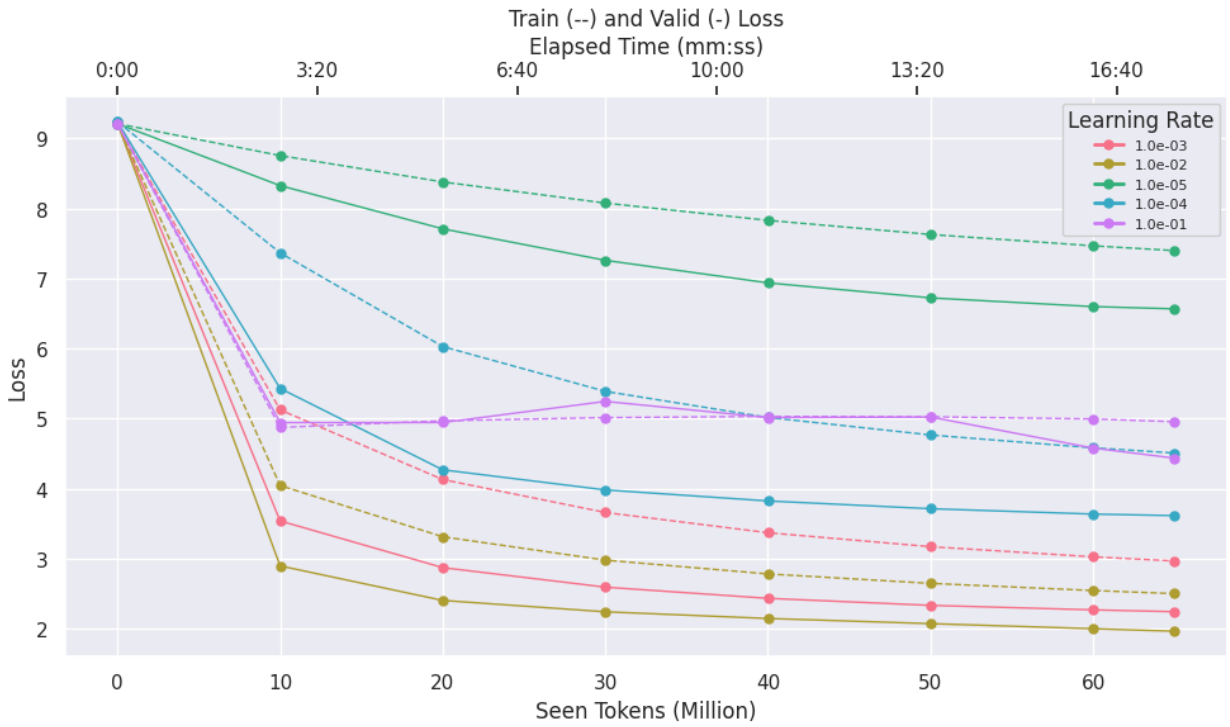
Learning Rate Scheduler: Linear Decay همراه با Warmup

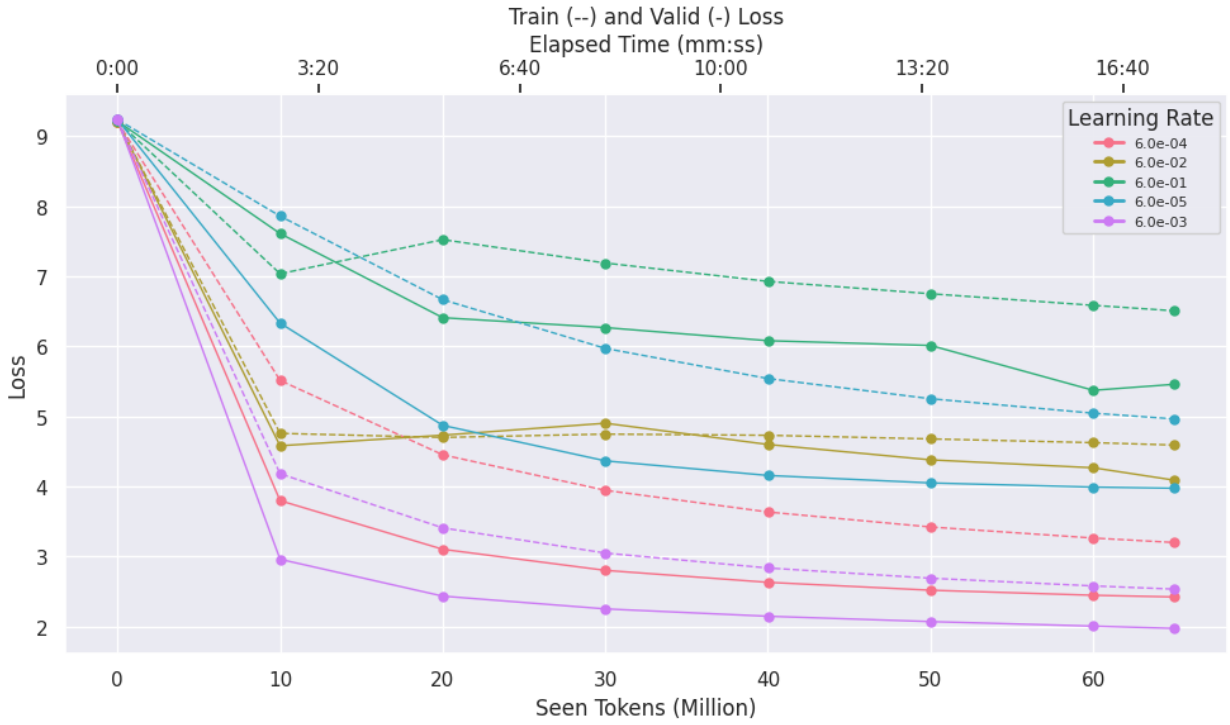
- Warmup: در مراحل ابتدایی آموزش نرخ یادگیری به‌صورت خطی از صفر تا مقدار اولیه افزایش یافته است.
- Linear Decay: پس از warmup، نرخ یادگیری به‌صورت خطی کاهش یافته

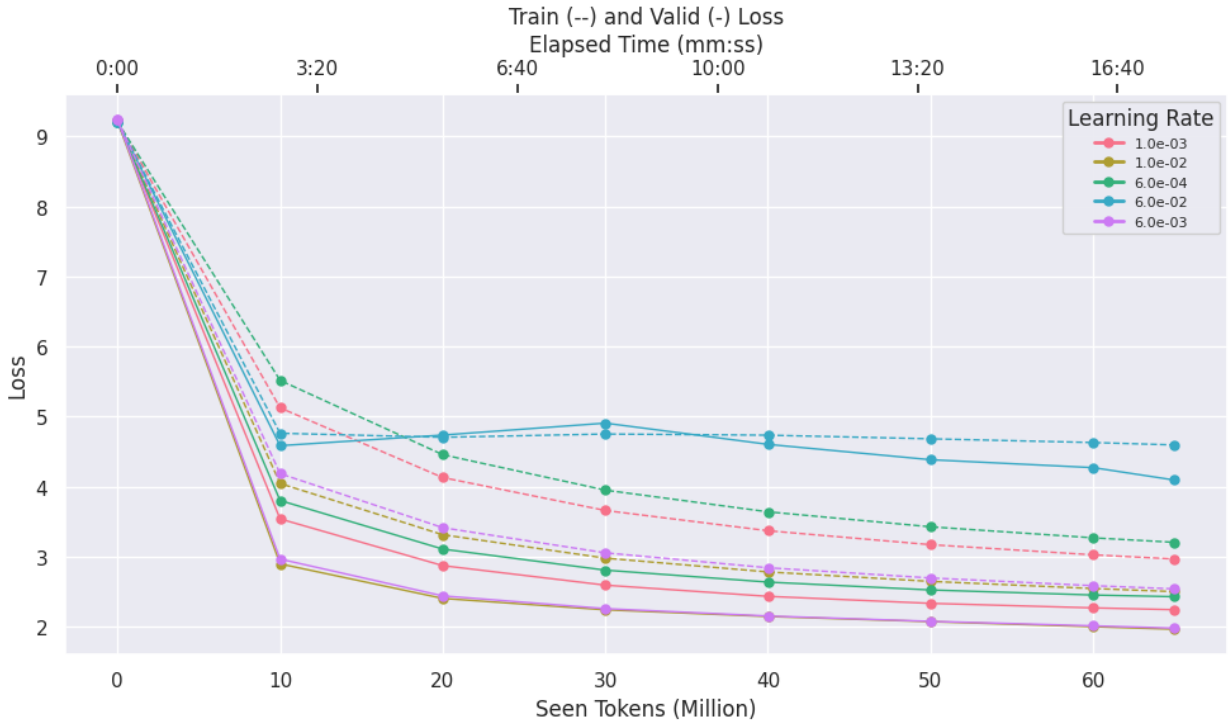


۲. مقایسه نرخ‌های یادگیری اولیه مختلف

در تصاویر زیر، عملکرد مدل با نرخ‌های یادگیری مختلف نشان داده شده است:

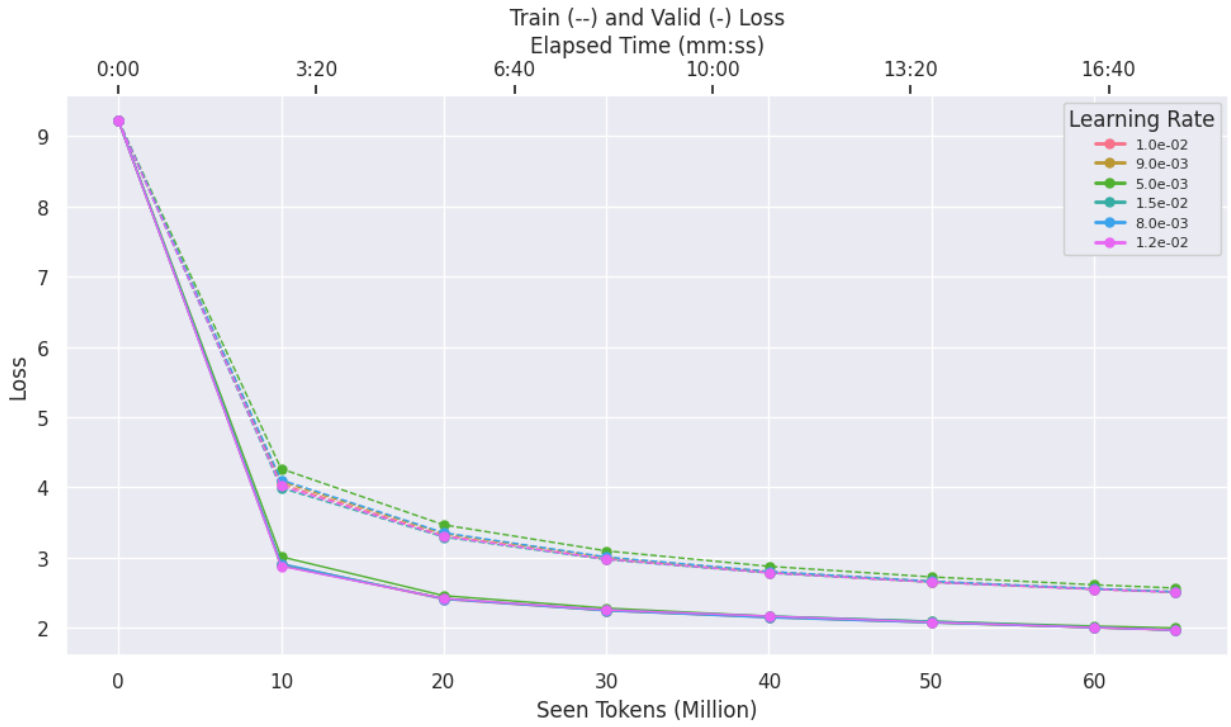






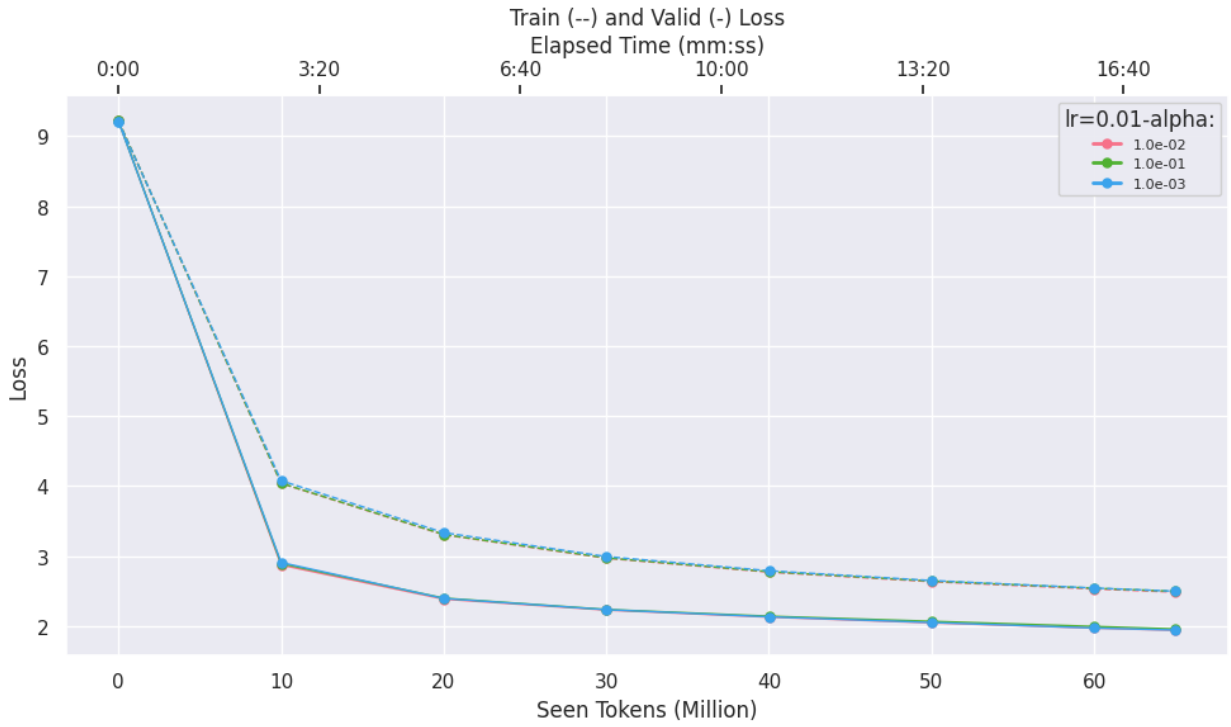
همان‌طور که در نمودارها مشاهده می‌شود، بهترین عملکرد مربوط به نرخ یادگیری اولیه **0.01** است. این مقدار منجر به کاهش سریع و پایدار loss در هر دو مجموعه آموزش و اعتبارسنجی شده است.

۳. بررسی نرخ یادگیری در حوالی 0.01



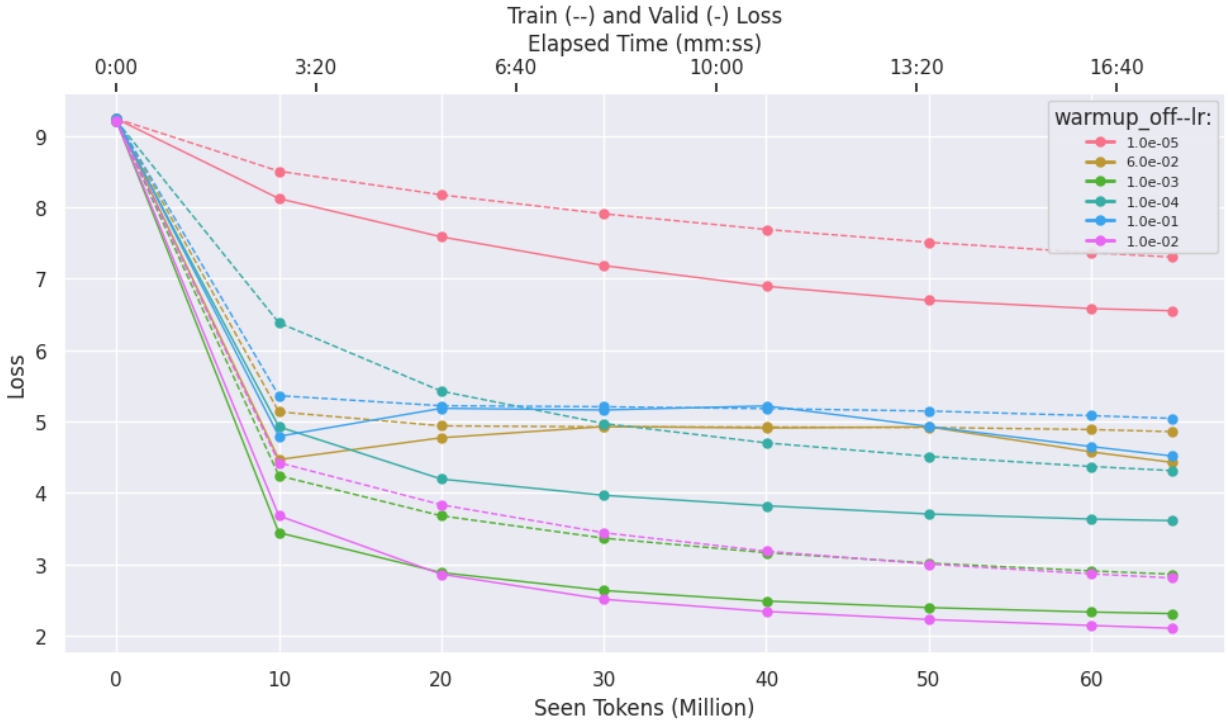
در آزمایش‌های جزئی‌تر با نرخ‌هایی در حوالی ۰,۰۱، مشاهده شد که نرخ‌هایی مانند ۰,۰۰۹، ۰,۰۱۲ و ۰,۰۰۸ نیز عملکردی مشابه و نزدیک به ۰,۰۱ داشتند، اما در مجموع ۰,۰۱ همچنان بهترین تعادل را بین سرعت همگرایی و دقت نهایی ایجاد کرده است.

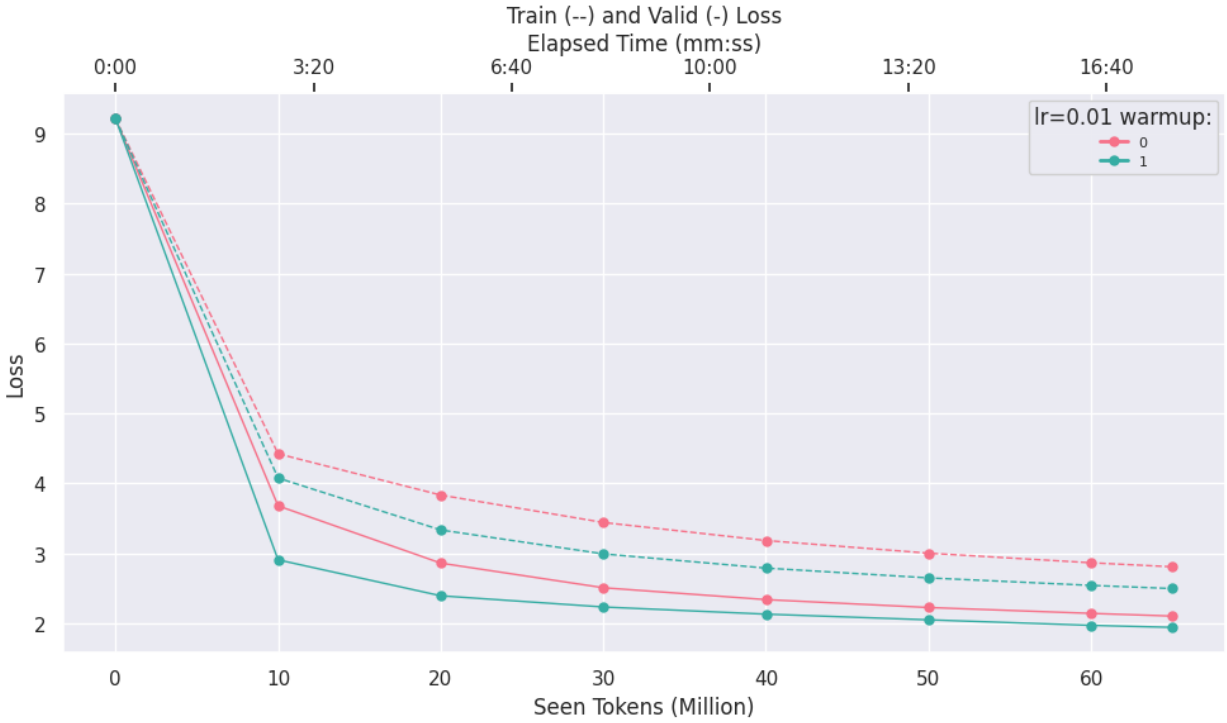
۴. بررسی تاثیر پارامتر α در Linear Learning Rate Decay



با توجه به اینکه مدل تنها بر روی ۶۵ میلیون توکن آموزش دیده است و کاهش loss در نمودار همچنان ادامه دارد، به نظر می‌رسد آموزش در این مرحله برای ارزیابی دقیق مقادیر مختلف آلفا کافی نباشد. بنابراین، ادامه فرایند آموزش با حجم بیشتری از توکن‌ها جهت دستیابی به تحلیل دقیق‌تر توصیه می‌شود.

۵. بررسی تاثیر خاموش‌بودن Warmup





- با Warmup غیرفعال (۰) مدل شروع خوبی ندارد.
- با Warmup فعال (۱) Loss خیلی سریع‌تر کاهش یافته (شیب تندتر) و مقدار نهایی Loss کمتر است.

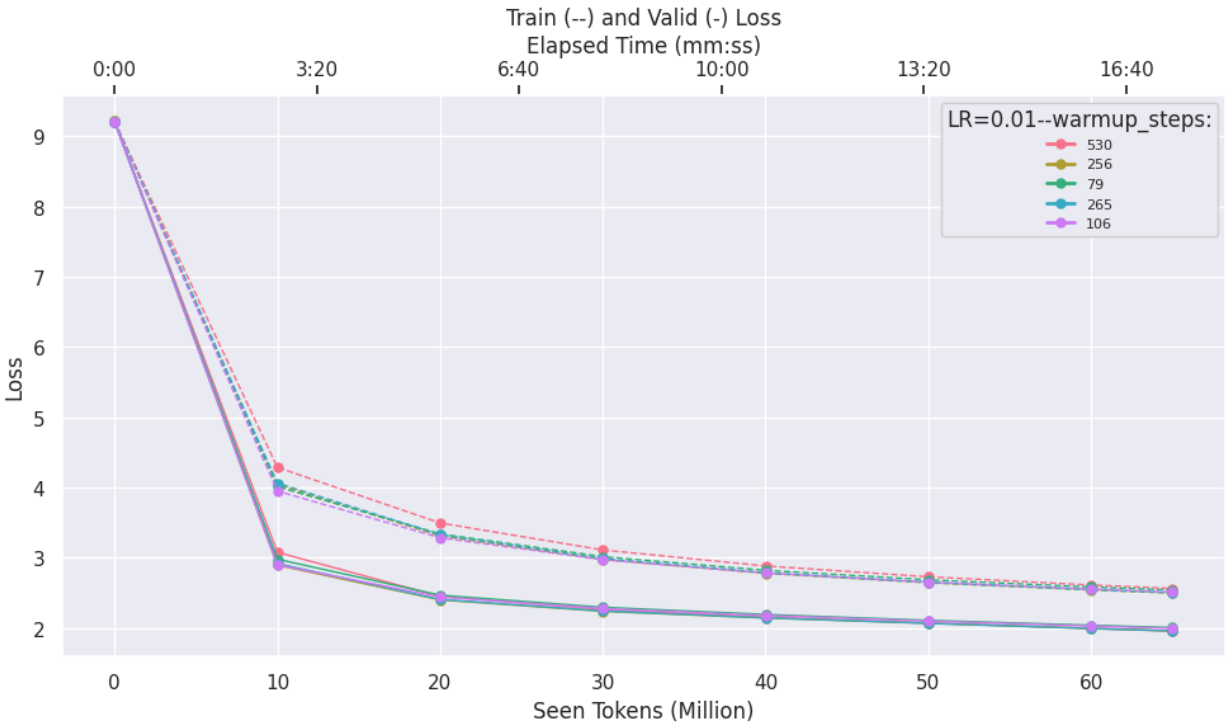
۶. تحلیل Warmup Steps

Warmup باعث می‌شود مدل در مراحل ابتدایی آموزش، با احتیاط یادگیری را آغاز کند. این رویکرد با افزایش تدریجی نرخ یادگیری، از ناپایداری و واپاشی جلوگیری کرده و به مدل اجازه می‌دهد پایه‌ای پایدار برای یادگیری ایجاد کند. پس از آن، مدل با نرخ یادگیری کامل، سریع‌تر و مؤثرتر آموزش خواهد دید.

به‌طور معمول، مقدار warmup steps بین **۳٪ تا ۱۰٪** از کل تعداد گام‌های آموزش (training steps) در نظر گرفته می‌شود.

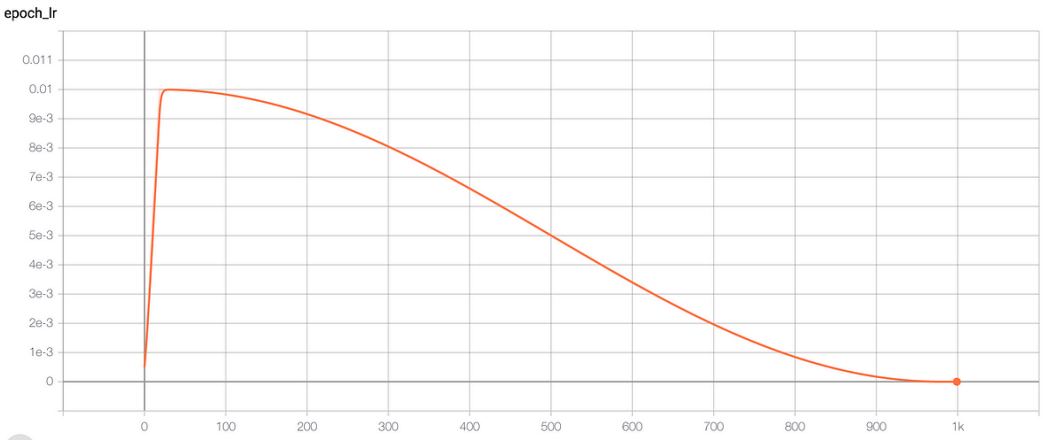
$$warmup\ steps = \frac{65{,}000{,}000}{512 \times 48} \times 10\% \approx 2645$$

3% $warmup\ steps$ = 79	4% $warmup\ steps$ = 106	9.7% $warmup\ steps$ = 256
10% $warmup\ steps$ =265	20% $warmup\ steps$ =530	

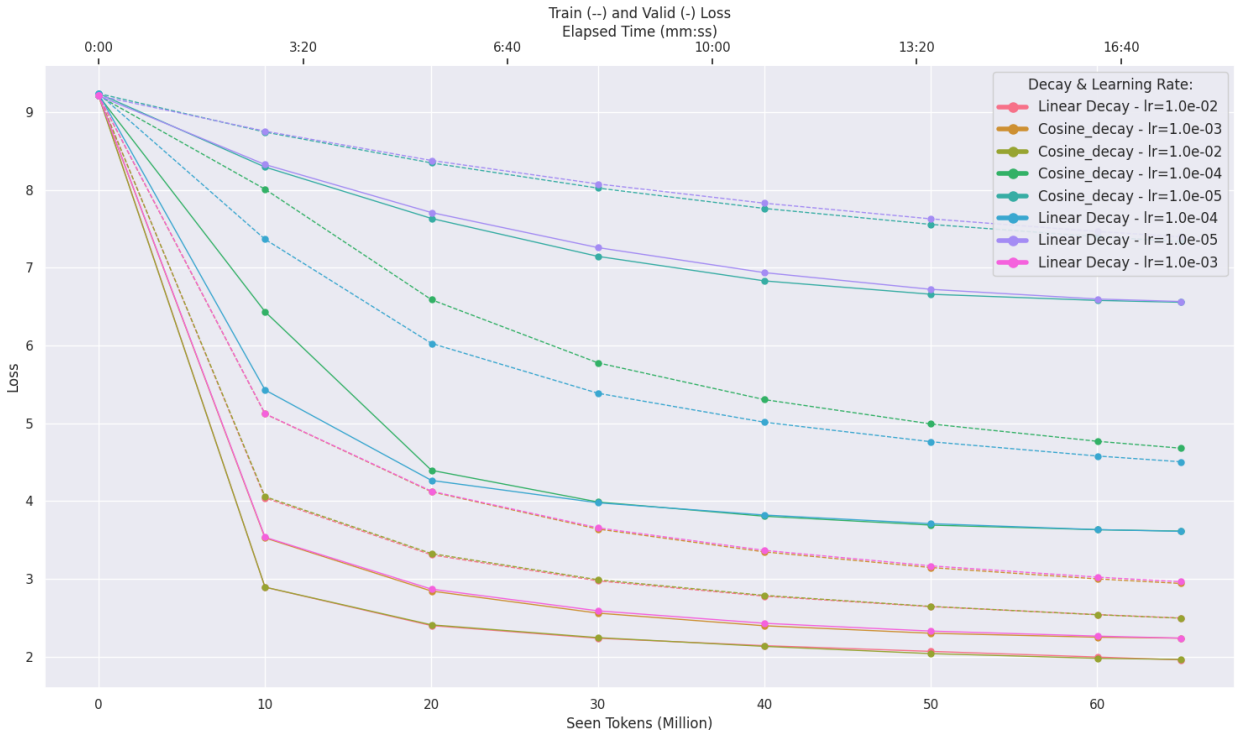


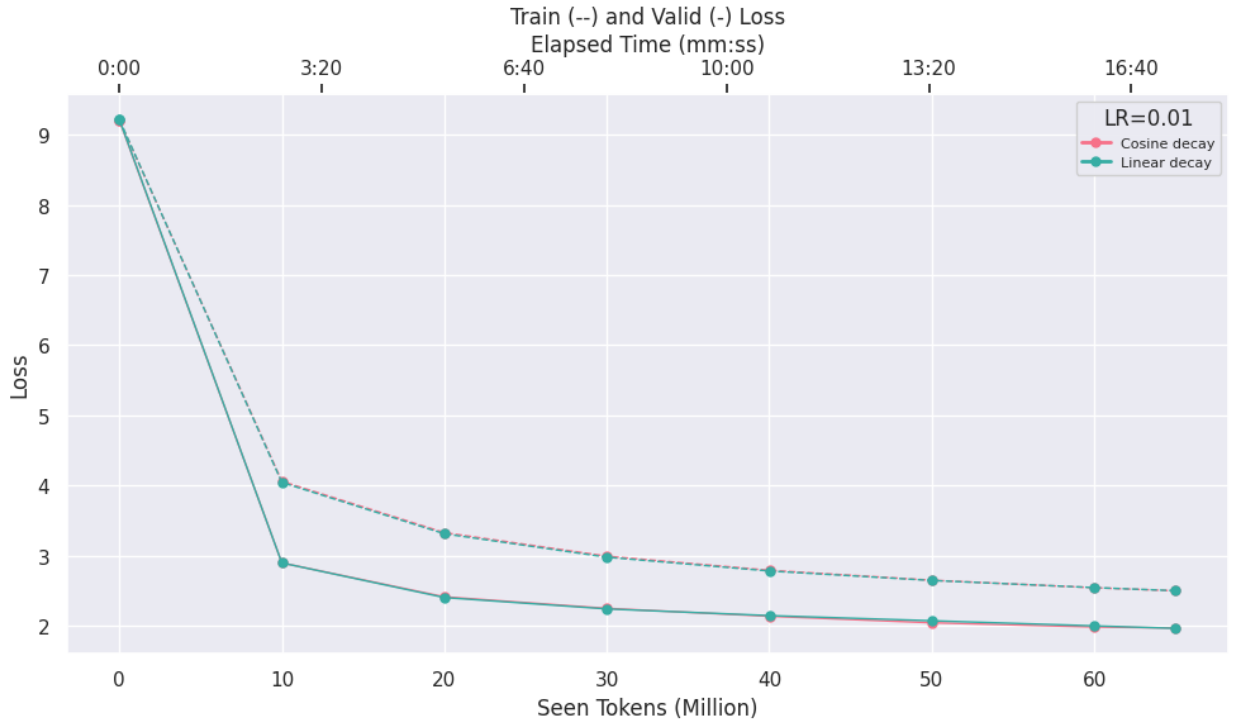
مقادیر warmup کمتر از ۲۶۵ (نظیر ۷۹، ۱۰۶، ۲۵۶ و ۲۶۵) عملکرد بهتری از خود نشان داده‌اند. در مقابل، مقدار ۵۳۰ موجب شده است که مدل با تأخیر بیشتری به نرخ یادگیری کامل برسد، که این موضوع منجر به کاهش سرعت هم‌گرایی و کندی در فرآیند آموزش شده است.

۷. بررسی استفاده از Cosine Decay به جای Linear Decay



نمودار cosine Decay





نتایج حاصل از نمودار نشان می‌دهد که هر دو روش کاهش نرخ یادگیری عملکرد مشابهی داشته‌اند، اما Cosine Decay در گام‌های پایانی آموزش توانسته است مقدار کمتری از loss را به‌دست آورد.