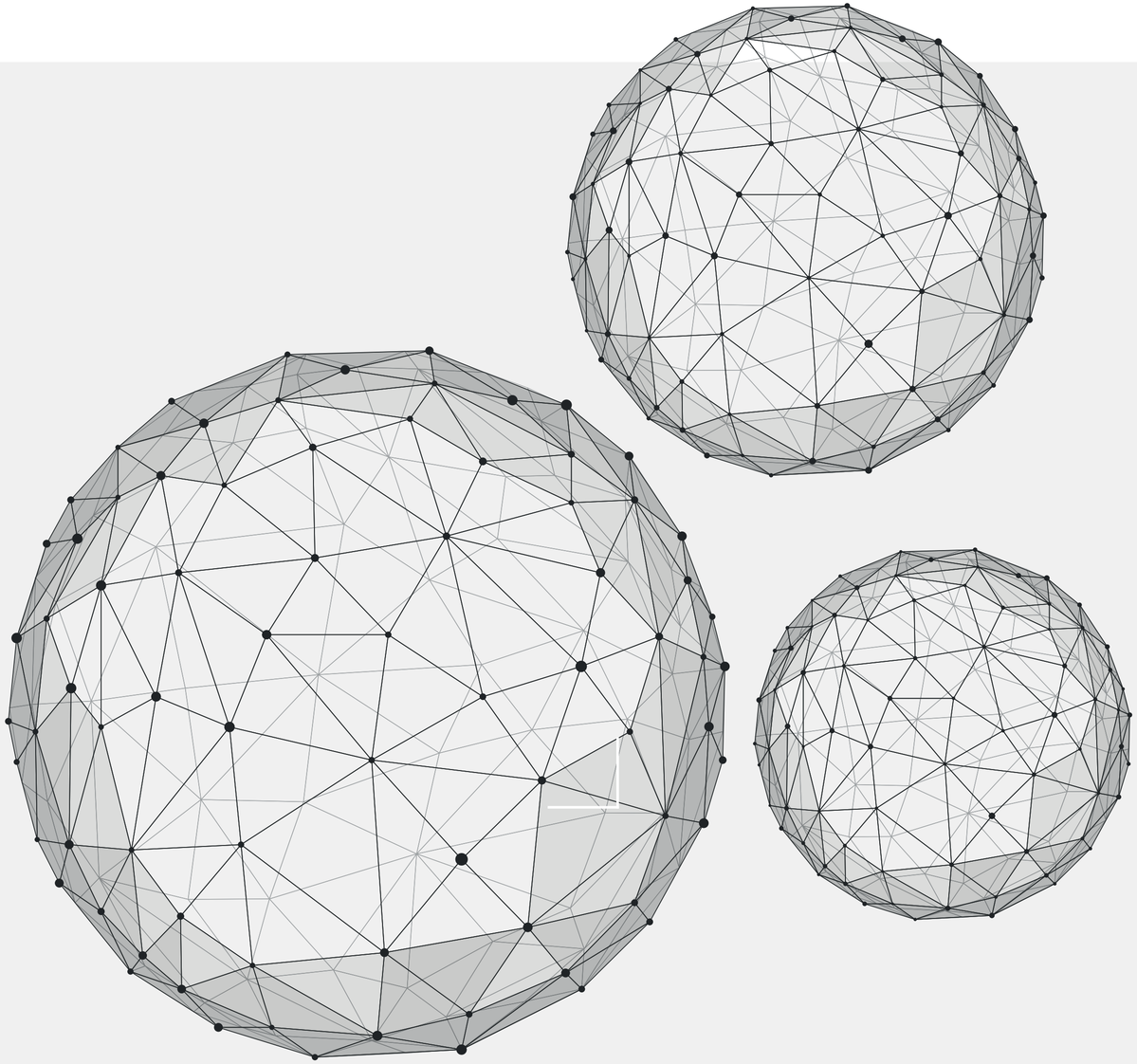
Trabajo Final de Máster

Comparativa de arquitecturas post-transformers para la traducción automática de texto

Felipe Ramírez Herrera



VNIVERSITAT DE VALÈNCIA

Agenda

Contribución

¿En qué medida las arquitecturas innovadoras desarrolladas tras el modelo Transformer original mejoran el rendimiento, la eficiencia computacional y la capacidad de generalización en tareas de traducción automática?

1. Comparar el desempeño de precisión, pérdida, exactitud y perplejidad entre el Transformer original y sus variantes recientes en diferentes conjuntos de datos (estructurados e informales).

2. Evaluar la eficiencia computacional de los modelos en términos de tiempo de entrenamiento, validación e inferencia, así como consumo de memoria y número de parámetros entrenables.

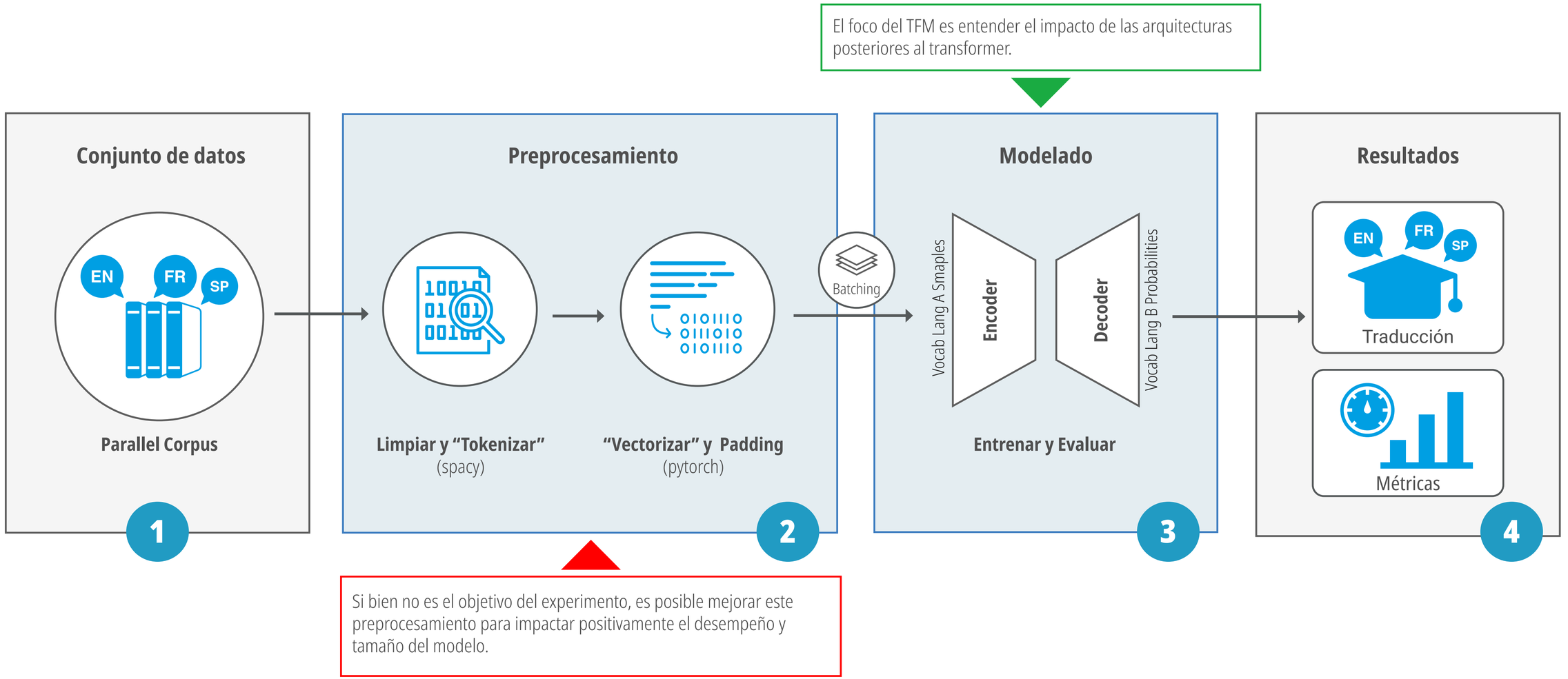
3. Determinar la capacidad de generalización de las arquitecturas mediante el análisis de sus resultados en validación, observando métricas como exactitud y perplejidad.

4. Identificar cuál de las arquitecturas modernas representa una mejor alternativa práctica al Transformer original, considerando el equilibrio entre rendimiento, estabilidad y consumo de recursos.

Nota: Ya casi se cumplen 8 años de la publicación Attention is All you need (Vaswani et al., 2017) y 10 años de la publicación de Neural Machine Translation by Jointly Learning to Align and Translate (Bahdanau et al, 2015).

Tarea: Neural Machine Translation (NMT)

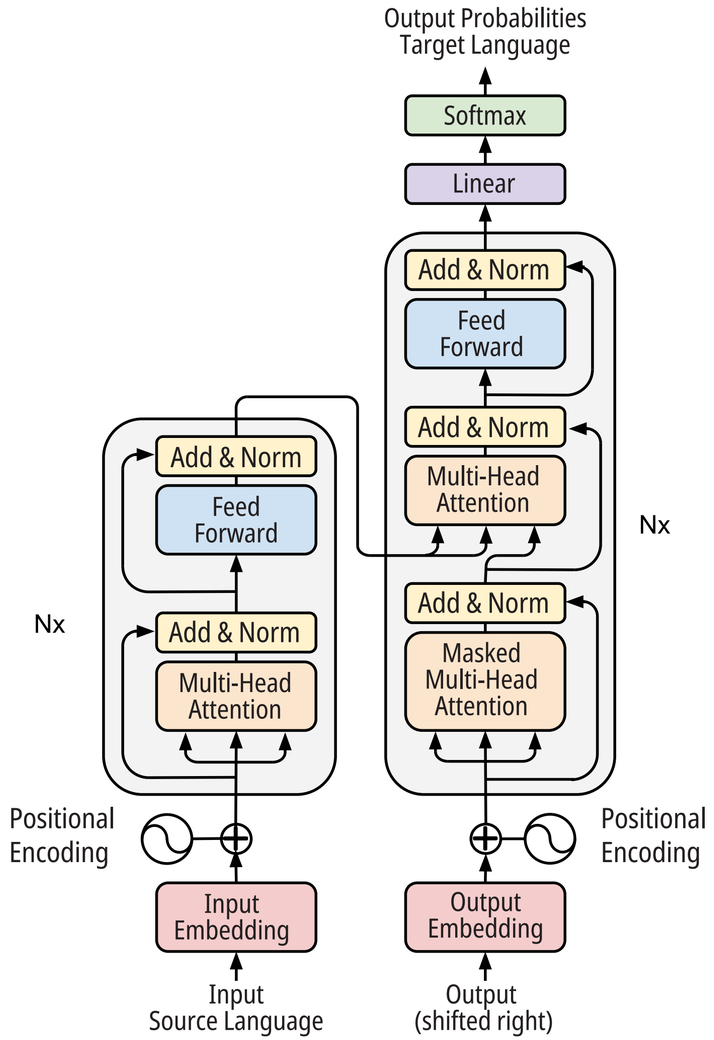
Está diseñada originalmente como una tarea de aprendizaje de extremo a extremo. Procesa directamente una secuencia de origen a una secuencia de destino (seq-to-seq).



Nota: El objetivo de aprendizaje es encontrar la secuencia de destino correcta dada la secuencia de origen, lo cual puede verse como un problema de clasificación de alta dimensionalidad que intenta mapear las dos oraciones en el espacio semántico. En todos los principales modelos modernos de NMT, este proceso se puede dividir en un paso de codificación y un paso de decodificación, y así separar funcionalmente todo el modelo. (Yang, Wang, & Chu, 2020)

Transformers: Self-Attentive Networks

Variantes de la arquitectura clásica para soportar tareas de traducción automática.



Modelo	Principales Diferencias
Universal Transformer (UT) (Dehghani et al., 2019)	Añade recurrencia en profundidad (repetidas refinaciones por posición), combinando paralelismo con inductivo recurrente; puede detener dinámicamente (dynamic halting) el refinamiento por token, mientras que Vainilla aplica una pila fija de capas.
Gated State Space (GSS) (Mehta et al., 2022)	Sustituye atención global por modelos de espacio de estado más eficientes (O(L logL) en lugar de $O(L^2)$), con gating para reducir la dimensionalidad y acelerar FFTs; más eficiente en secuencias largas.
Mamba (Gu & Dao, 2024)	Elimina la atención y MLPs, usando modelos de espacio de estado selectivos basados en el input; logra razonamiento dependiente del contenido y es lineal en longitud; Transformer depende de self-attention fijo sin selección dinámica.
Mega (Ma et al., 2022)	Reemplaza multi-head attention por una atención de una sola cabeza equipada con promedio móvil exponencial (EMA), introduciendo bias posicional local; ofrece variante de complejidad lineal mediante chunking, mientras Transformer usa self-attention con costos cuadráticos.
ST-MoE (Zoph et al., 2022)	Introduce Mixture-of-Experts (MoE) para activar dinámicamente solo partes del modelo, reduciendo cómputo efectivo mientras escala en tamaño total de parámetros; Transformer usa todos sus parámetros en cada paso. Además, aplica técnicas de estabilización y fine-tuning especializadas.

Notas: Tomado de Vaswani et al. (2017).

Conjunto de datos

Un corpus paralelo contiene traducciones del mismo documento en dos o más idiomas, alineadas al menos a nivel de oración. Estos tienden a ser más raros que los corpora menos comparables[1]. Es muy poco frecuente su uso en literatura o ejemplos.

Conjuntos de datos

Conjunto de datos	Espacio requerido en SSD	Número total de líneas	Vocabulario EN (Tokens)	Vocabulario ES (Tokens)
UN-Parallel Corpus[1]	~3,81 GB	25 227 004	577 671 824	669 899 929
OpenSubtitles V2018	~3.80 GB	61 434 251	391 976 667	364 279 696

Entorno de prueba

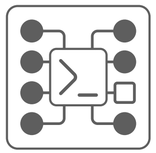
Elemento	Descripción
CPU	AMD® Ryzen ® 9 5950X 4.9 Ghz con 16 Cores / 32 Hilos, Caché L3: 64 MB.
GPU	NVIDIA ® GeForce ® RTX 3060 12 GB de memoria GDDR6 y 3584 CUDA Cores.
Memoria	64 GB de Memoria DDR4 (4 módulos)
Almacenamiento	Kingston® NV1 NVMe M.2 PCIe Gen3, 1TB.
Sistema operativo	WSL2 sobre Microsoft Windows ® 11 Professional (Ubuntu® 22.04.2 LST)
Ambiente	Miniconda. Python 3.10.13, Pytorch 2.2.0 / NVidia CUDA 12.1.105

Nota: (1) La versión actual del fully-aligned corpus contiene contenido que fue producido y traducido manualmente entre 1990 y 2014, incluyendo alineaciones a nivel de oración. Véase https://conferences.unite.un.org/uncorpus

Particionamiento de los datos y otras configuraciones

Datos de entrenamiento

Conjunto de datos	Número de ejemplos	Ejemplos para entrenamiento	Ejemplos para validación	Ejemplos para inferencia
UN-Parallel Corpus	250 000 (2,19%)	175 000 (~1.54%)	50 250 (~0.44%)	24 750 (~0,22%)
OpenSubtitles V2018	1 000 000 (1,62%)	700 000 (~1,14%)	201 000 (~0.33%)	99 000 (~0.16%)

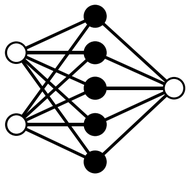


Hiperparámetros y configuración

Conjunto de datos	min_seq_length	max_seq_length	EN-Vocab Size	ES-Vocab Size
UN-Parallel Corpus	3	450 (941)	50 288	68 931
OpenSubtitles V2018	3	80	70 729	94 982

Optimizadores	Scheduler	Early Stopping	Clipping	Loss Measure (ST)	Batch Size
1 x ADAM LR: 0.001	LROnPlateau	5	15	Cross-Entropy (CE)	UNPC: 8 OSPC: 16

Configuración de las arquitecturas

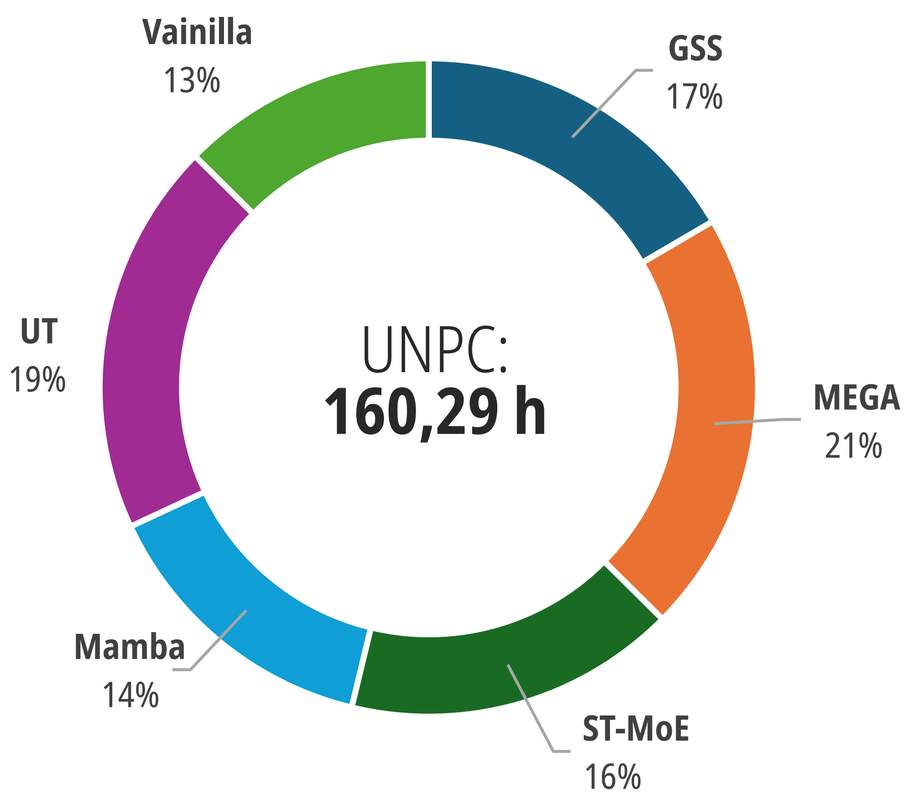


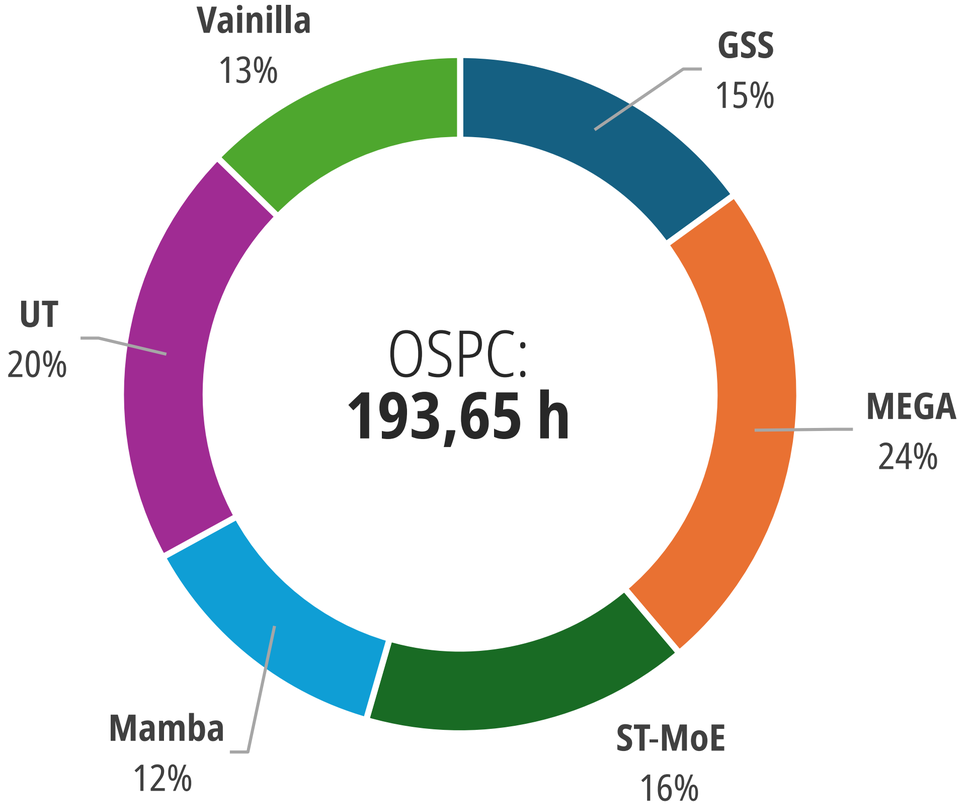
Arquitetura

Modelo	Parámetros	Modelo	Parámetro
Vainilla Transformer	-num_encoder_layers:6 -num_decoder_layers:6 -emb_size(d_model):512 -nhead:8 -ff_dim:512 -dropout:0.1 - lr: 0.0001	**ST-MoE Transformer**	- d_model: 512 - n_heads: 8 - num_experts: 8 - gating_top_n: 2 - dropout: 0.1 - max_seq_length: variable - lr: 0.0001
Universal Transformer (UT)	- emb_dim: 512 - hidden_dim: 1024 - pff_dim: 1024 - n_layers: 4 - n_heads: 8 - dropout_ratio: 0.1 - max_len: variable - pad_id: 0 - lr: 0.0001	**Gated State Spaces (GSS)**	- d_model: 512 - n_heads: 8 - depth: 3 - dss_kernel_H: 512 - dss_kernel_N: 256 - dim_expansion_factor: 4 - dropout: 0.1 - max_seq_length: variable - lr: 0.0001
Mamba	- d_model: 512 - n: 256 - block_count: 4 - dropout: 0.3 - max_length: variable - lr: 0.0001 Mamba Interno: - d_state: 16 - d_conv: 4 - expand: 2 Attention: - n_heads: 8	**Moving Average Gated Attention (MEGA)**	- d_model: 512 - n_heads: 8 - depth: 3 - ema_heads: 16 - attn_dim_qk: 64 - attn_dim_value: 256 - dropout: 0.1 - max_seq_length: variable - lr: 0.0001

Tiempo de entrenamiento y validación

Menor tiempo conduce a un uso más eficiente de los recursos computacionales y a la construcción rápida de modelos efectivos.

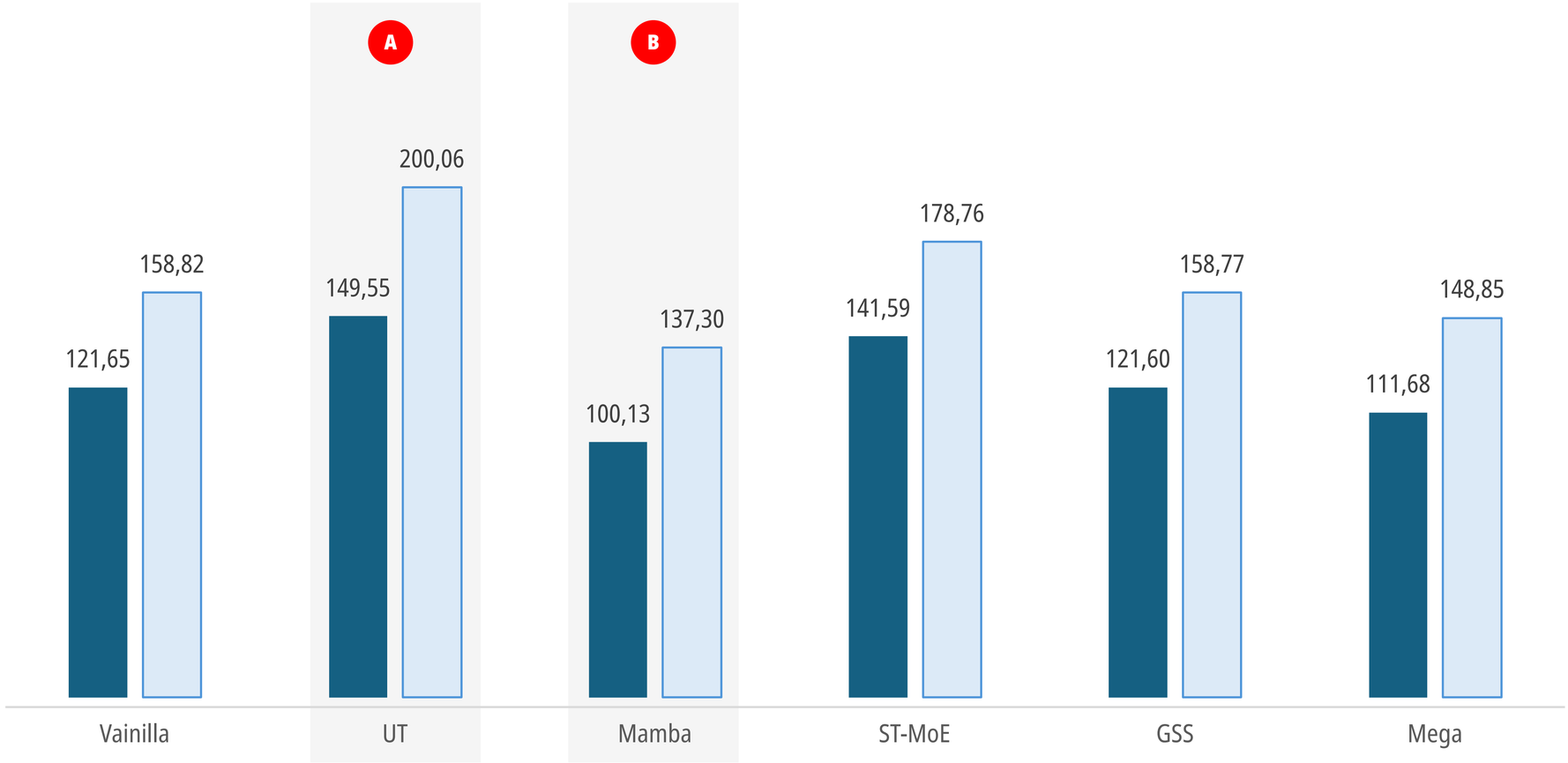




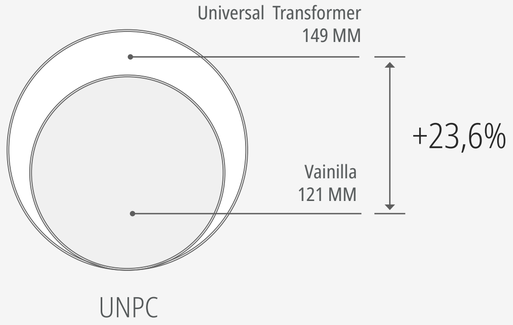
Nota: Tiempo promedio dado en horas tras efectuar 20 epochs. Se utiliza UNPC para designar al conjunto de datos Corpus Paralelo de las Naciones Unidas (Ziemski et al., 2016) y OSPC para designar a OpenSubtitles v2018 (Lison & Tiedemann, 2016).
Fuente: Elaboración propia del estudiante.

Cantidad de parámetros

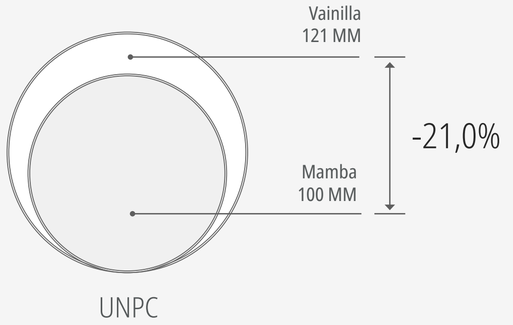
Efecto de las técnicas aplicadas sobre número de parámetros entrenables de los modelos propuestos.



A Algunos de los enfoques impactan negativamente en el número de parámetros:



B Mamba y Mega reducen la cantidad de parámetros necesarios para hacer el trabajo:

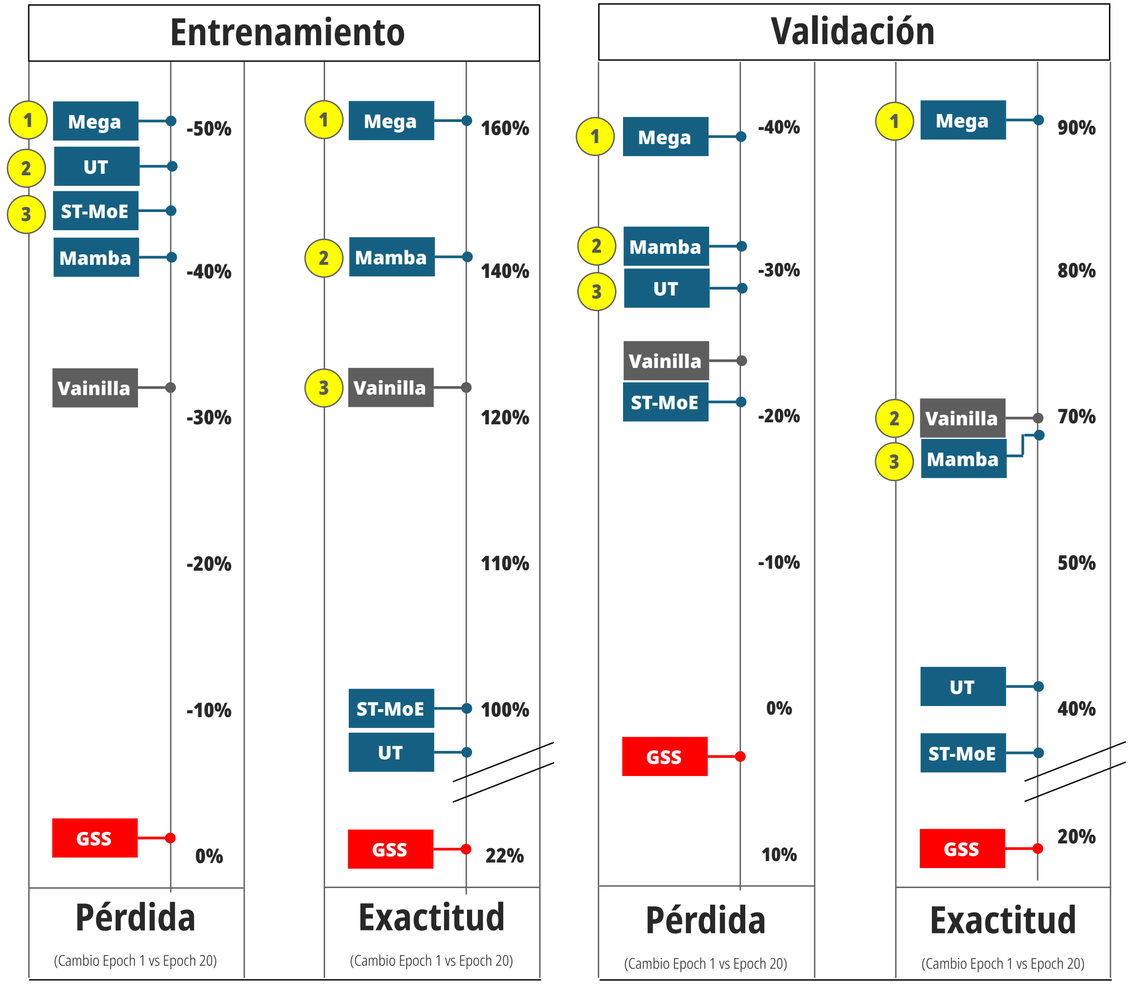


Fuente: Elaboración propia del estudiante. Cifras absolutas dadas en millones de parámetros. Se utiliza UNPC para designar al conjunto de datos Corpus Paralelo de las Naciones Unidas (Ziemski et al., 2016) y OSPC para designar a OpenSubtitles v2018 (Lison & Tiedemann, 2016).

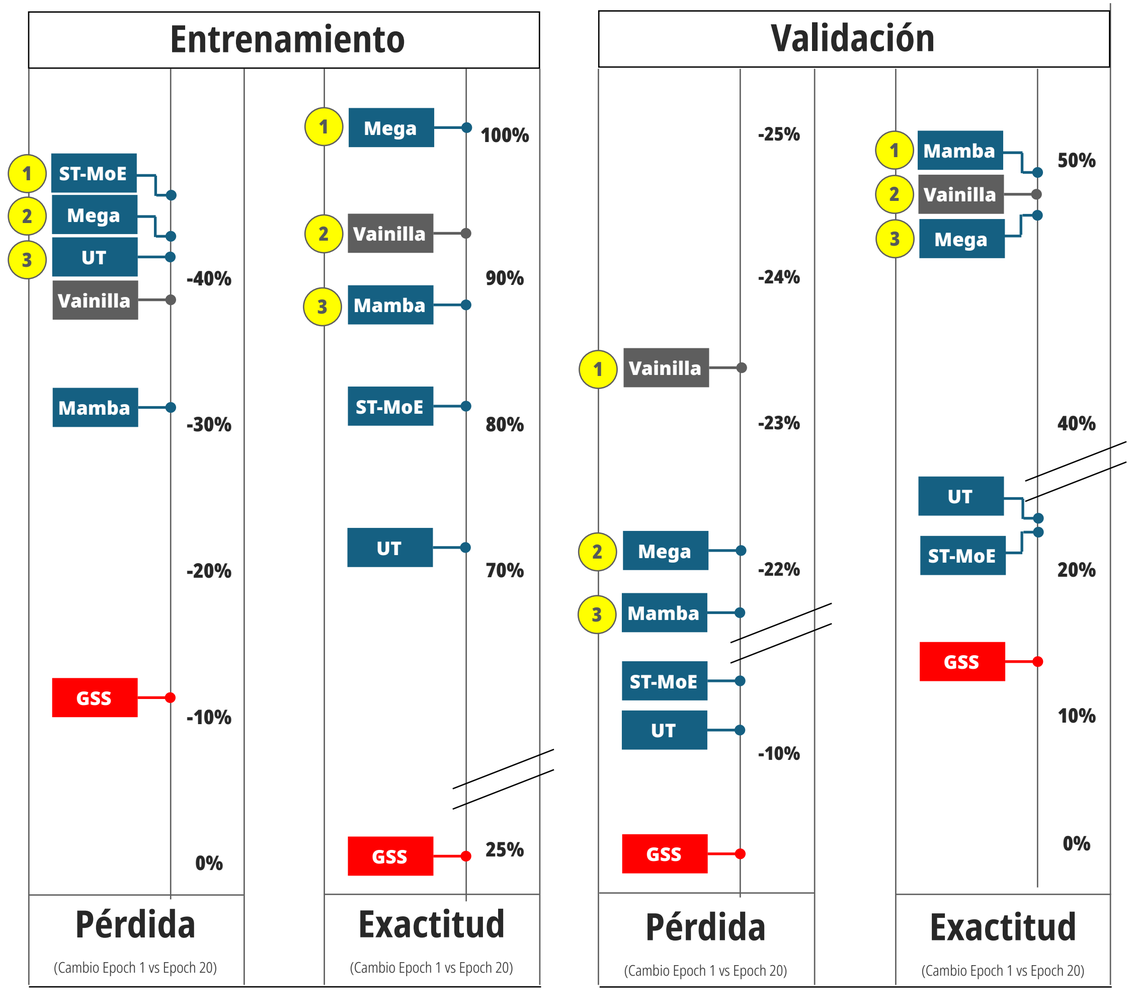
La escala de un modelo es uno de los temas más importantes y muchas veces condiciona la "capacidad" del modelo para generalizar. Dado un presupuesto computacional fijo, entrenar un modelo más grande durante menos pasos es mejor que entrenar un modelo más pequeño durante más pasos.

Resumen

Se muestran los principales resultados.



UN-Parallel Corpus V1

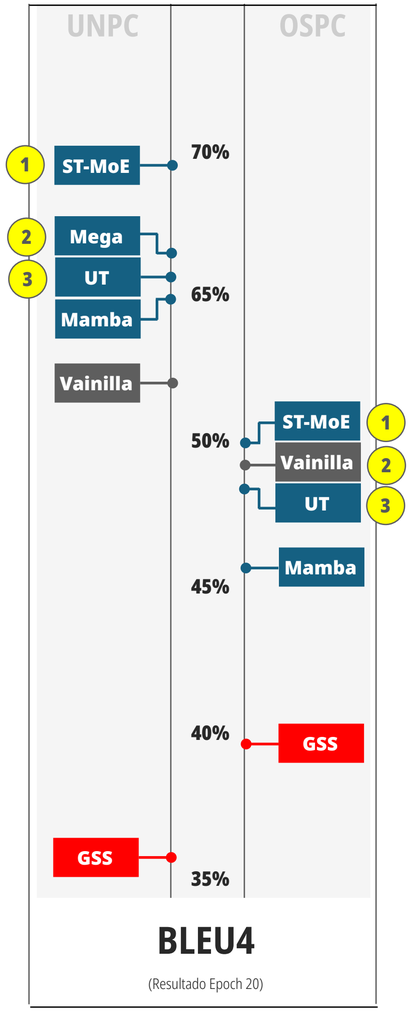


OpenSubtitles V2018

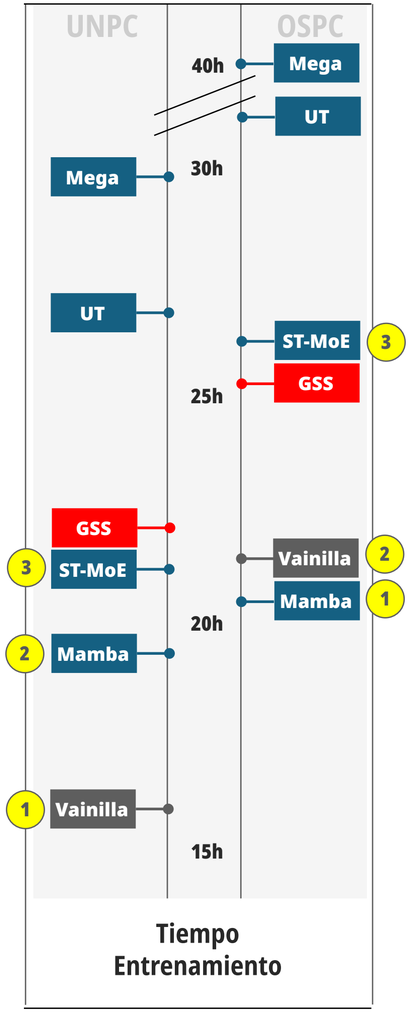
Resumen (continuación)

Se muestran los principales resultados.

- En general, todos los modelos muestran un desempeño superior en UNPC, lo cual es esperable dada la mayor regularidad y estructura del corpus (formado por frases más formales y menos ambiguas).
- Los modelos que obtienen los mejores valores absolutos en BLEU-4 en el conjunto UNPC son ST-MoE (0,6760), UT (0,6569), MEGA (0,6593) y Mamba (0,6482).
- El rendimiento sobre OSPC, todos los modelos sufren una caída, pero algunos logran conservar valores relativamente altos. En particular, UT mantiene un valor de 0,4844, ST-MoE y MEGA empatan con 0,4960 y Vainilla alcanza 0,4938, superando incluso a Mamba en este corpus.



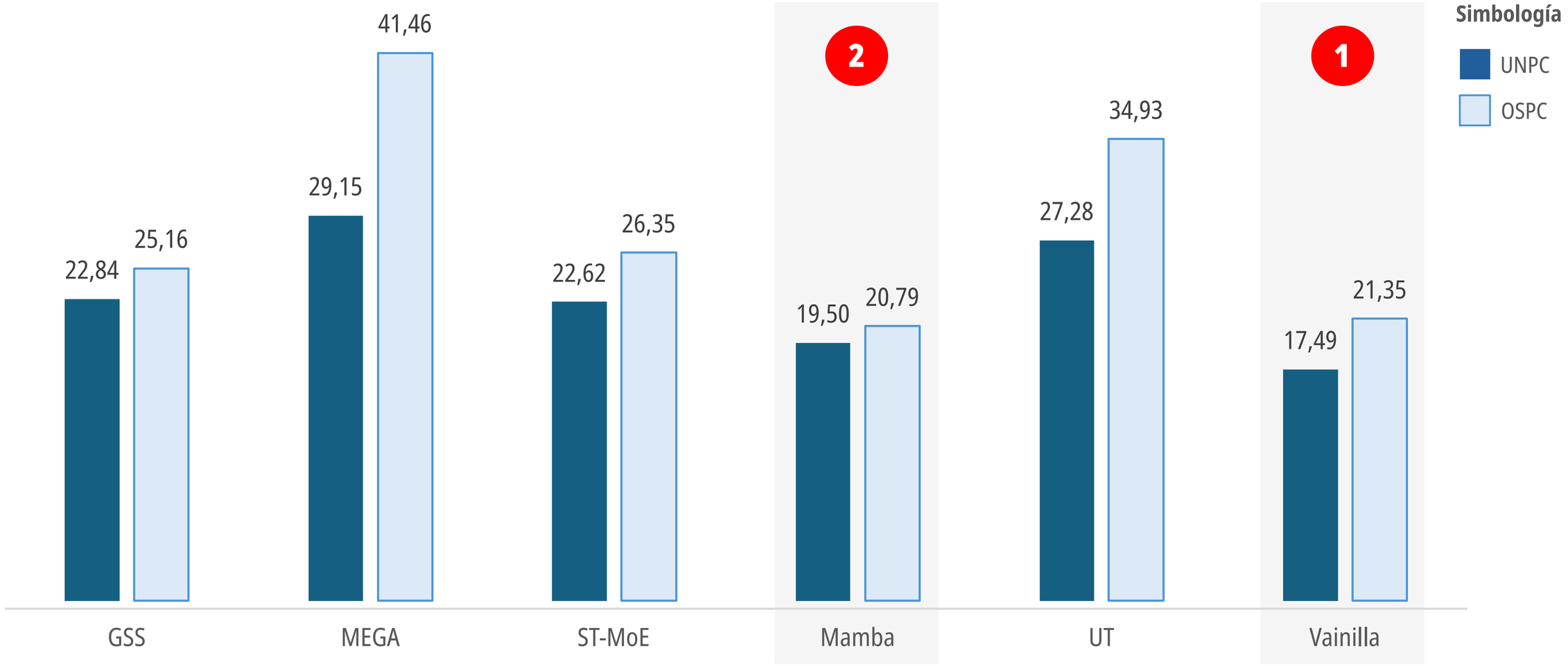
- El modelo Vainilla, es consistentemente el más rápido en ambas condiciones, con valores de 17,49 h (UNPC) y 21,35 h (OSPC), lo que refuerza su utilidad como baseline eficiente, especialmente en etapas tempranas de experimentación.
- Mamba destaca por ser la arquitectura optimizada más eficiente en esta métrica, con tiempos de entrenamiento cercanos al Vainilla (19,50 h en UNPC y 20,79 h en OSPC), incluso superándolo ligeramente en eficiencia en el corpus OSPC.
- UT y MEGA son las que presentan los mayores tiempos de entrenamiento.



Fuente: Elaboración propia del estudiante. Cifras absolutas dadas en millones de parámetros. Se utiliza UNPC para designar al conjunto de datos Corpus Paralelo de las Naciones Unidas (Ziemski et al., 2016) y OSPC para designar a OpenSubtitles v2018 (Lison & Tiedemann, 2016).

Tiempo de entrenamiento

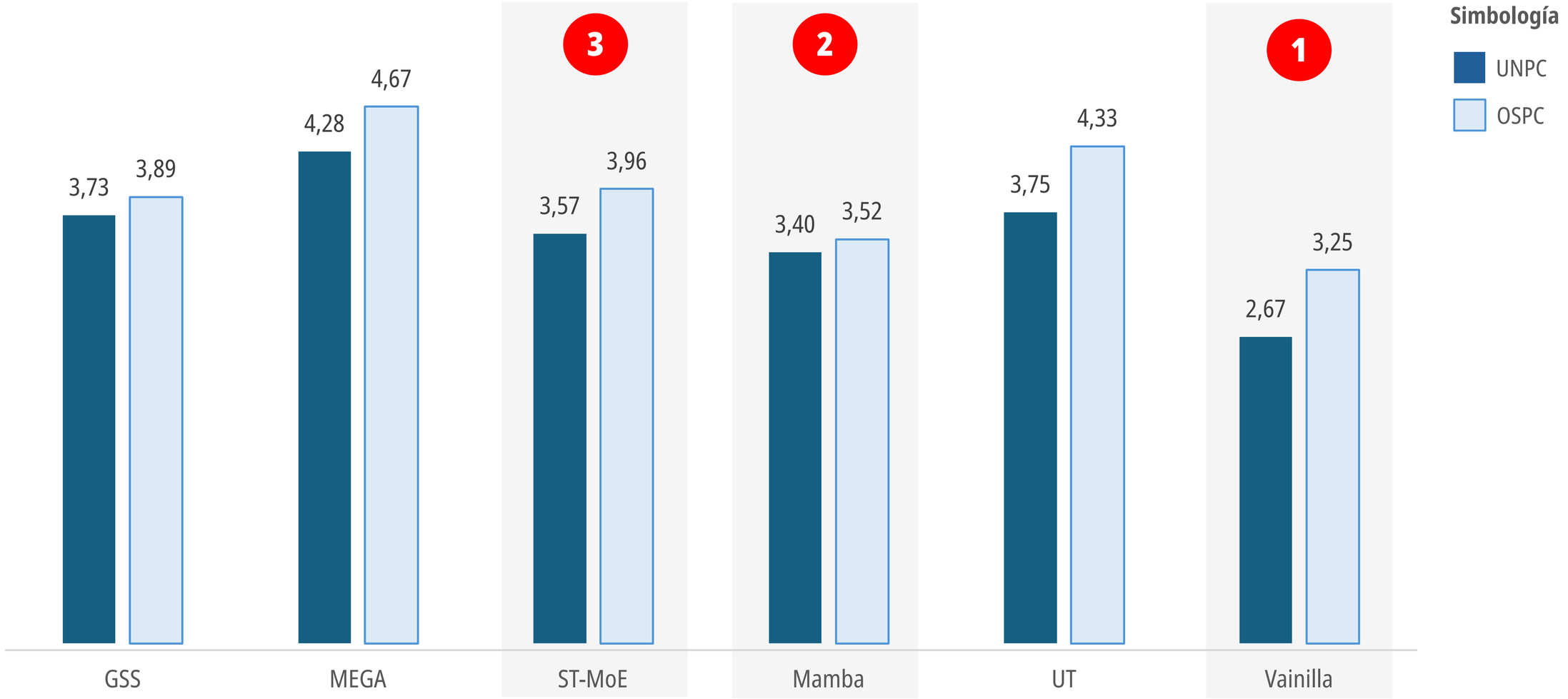
Menor tiempo conduce a un uso más eficiente de los recursos computacionales y a la construcción rápida de modelos efectivos.



Nota: Tiempo promedio dado en horas tras efectuar 20 epochs. Se utiliza UNPC para designar al conjunto de datos Corpus Paralelo de las Naciones Unidas (Ziemski et al., 2016) y OSPC para designar a OpenSubtitles v2018 (Lison & Tiedemann, 2016).
Fuente: Elaboración propia del estudiante.

Tiempo de validación

Menor tiempo conduce a un uso más eficiente de los recursos computacionales y a la construcción rápida de modelos efectivos.

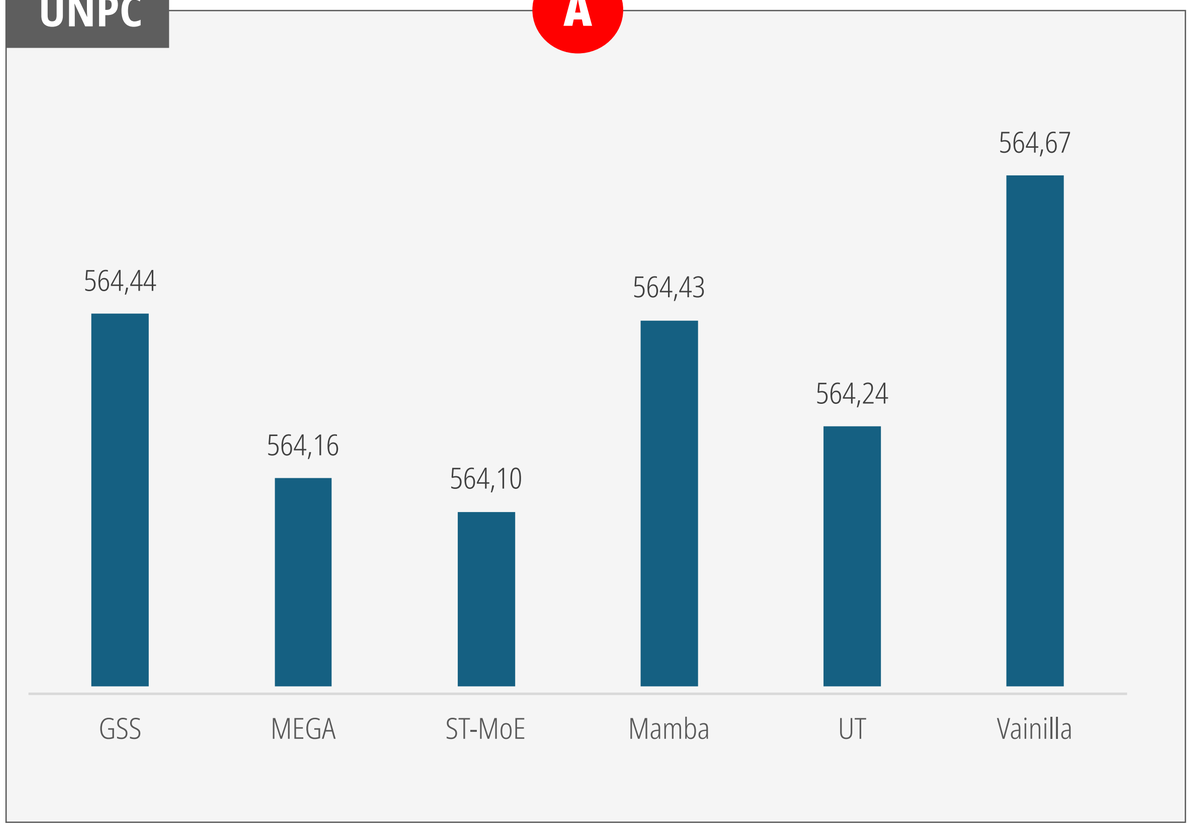


Nota: Tiempo promedio dado en horas tras efectuar 20 epochs. Se utiliza UNPC para designar al conjunto de datos Corpus Paralelo de las Naciones Unidas (Ziemski et al., 2016) y OSPC para designar a OpenSubtitles v2018 (Lison & Tiedemann, 2016).
Fuente: Elaboración propia del estudiante.

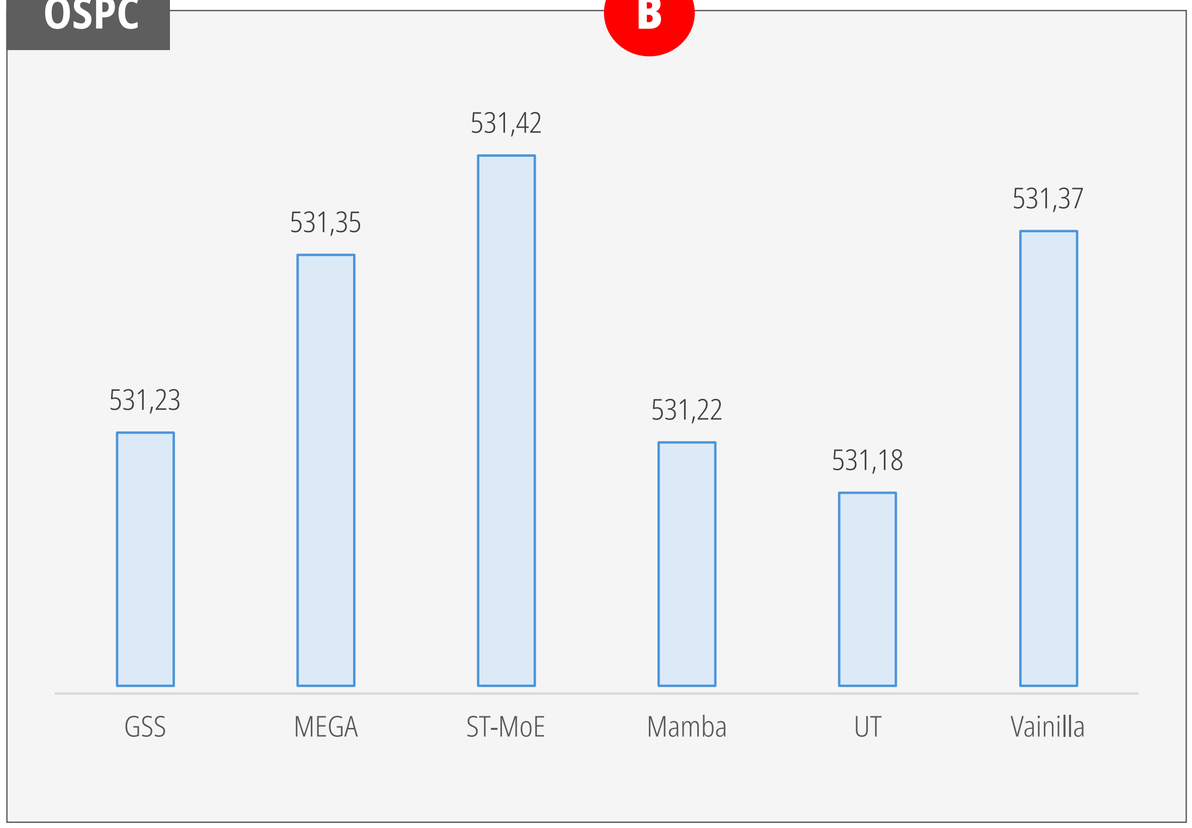
Consumo de memoria

La cantidad de memoria disponible en la unidad de procesamiento (CPU o GPU) puede limitar el tamaño del modelo y el tamaño de los lotes de datos que se pueden procesar simultáneamente.

UNPC **A**



OSPC **B**

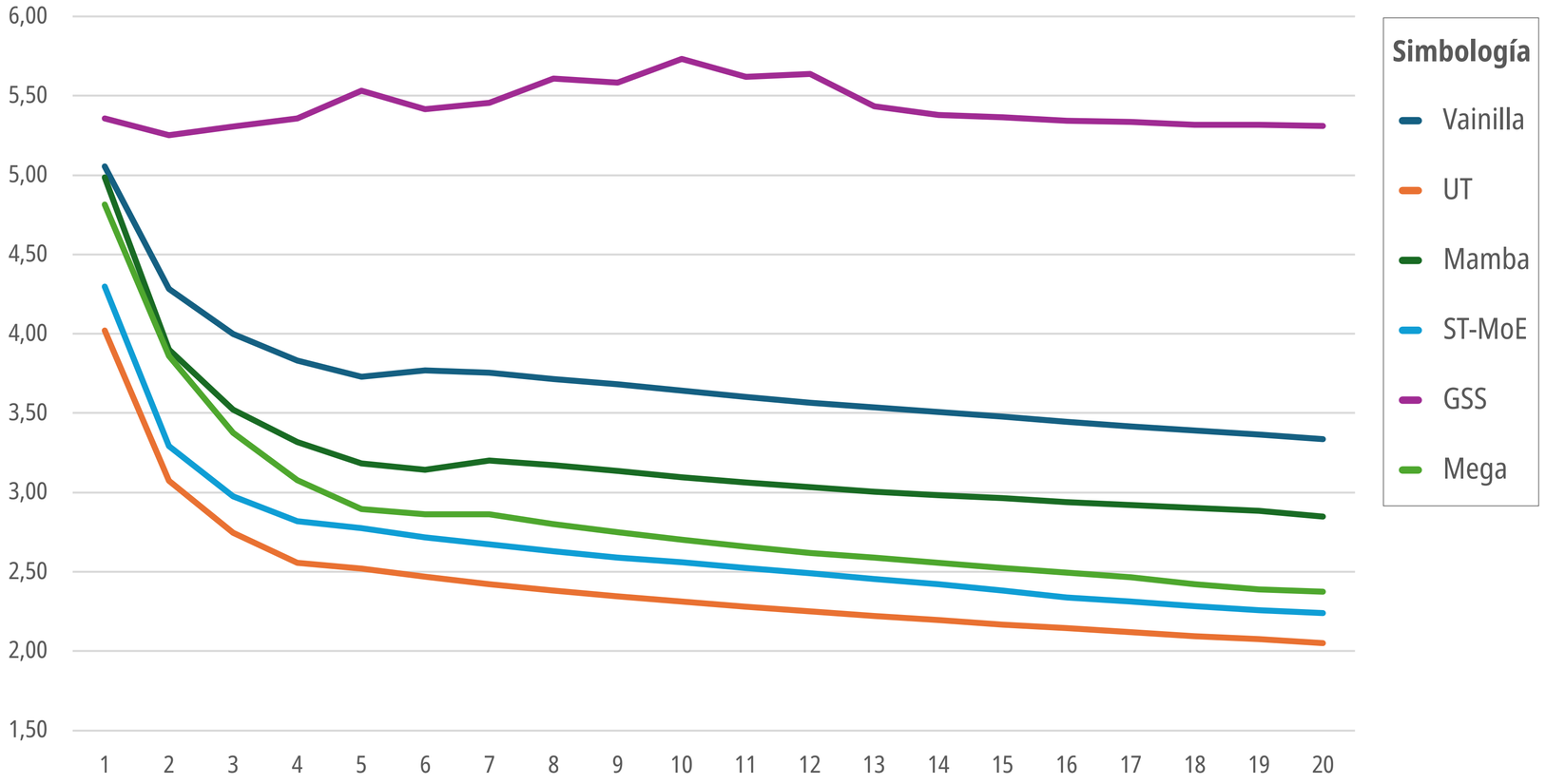


No hay una marcada diferencia entre los diferentes modelos.

Notas: Consumo de memoria promedio por batch dado en MB tras efectuar 20 ejecuciones.
Fuente: Elaboración propia del estudiante.

Curvas de pérdida - Entrenamiento (UNPC)

Comportamiento de la pérdida de los modelos durante el entrenamiento.

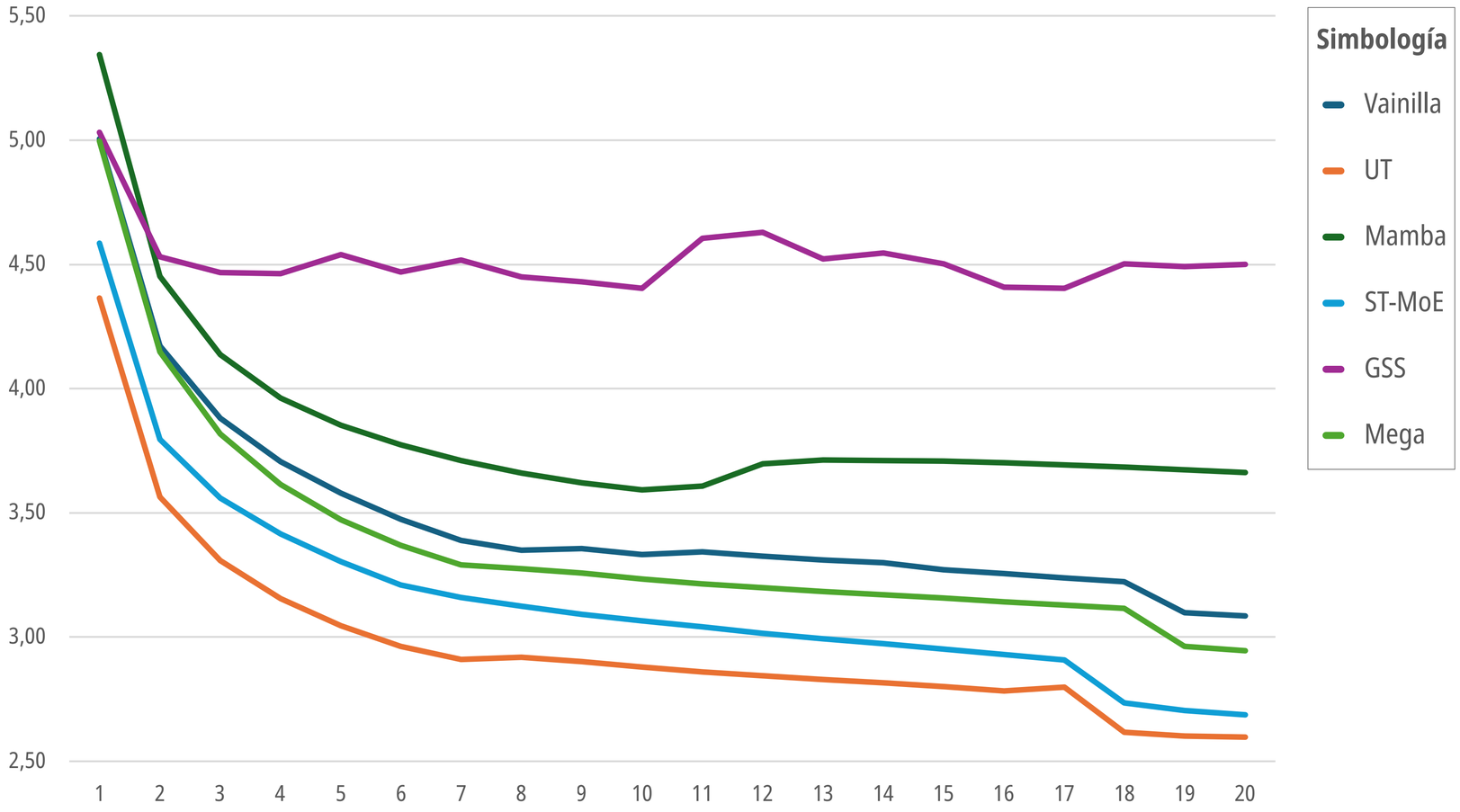


Fuente: Elaboración propia del estudiante.

- Modelos como UT, Mamba y MEGA logran un descenso rápido y sostenido en la pérdida, lo que evidencia una mayor eficacia en el proceso de aprendizaje desde etapas tempranas.
- Vainilla muestra una reducción más lenta, con una curva más plana a partir de la mitad del entrenamiento, lo que sugiere una convergencia limitada.
- La curva de GSS se mantiene elevada y presenta incluso oscilaciones leves a lo largo de las épocas, reflejando dificultades de aprendizaje o inestabilidad en el proceso de optimización.

Curvas de pérdida - Entrenamiento (OSPC)

Comportamiento de la pérdida de los modelos durante el entrenamiento.

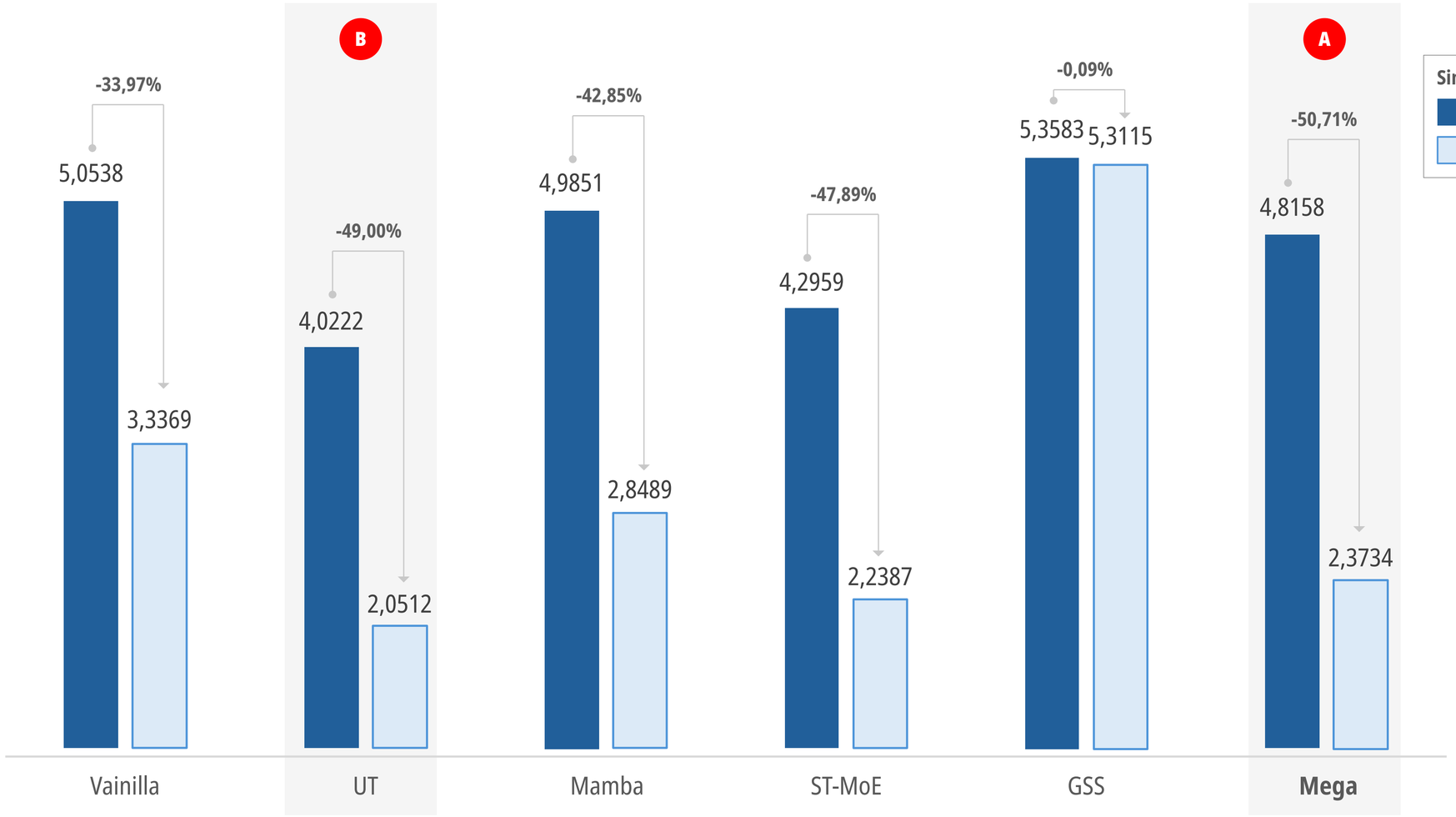


- UT muestra una caída abrupta de la pérdida, lo cual indica una rápida capacidad de aprendizaje inicial. A partir de la época 10, su curva se estabiliza, es la arquitectura con la menor pérdida final de entrenamiento,
- MoE también exhibe un comportamiento favorable, con una convergencia relativamente rápida y una pérdida final ligeramente superior a la de UT. Mamba y Vainilla, con descensos progresivos y estables.
- Mega, aunque comienza con una pérdida alta, logra reducirla progresivamente y mantiene una trayectoria regular, aunque con una convergencia más lenta.

Fuente: Elaboración propia del estudiante.

Pérdida - Entrenamiento (UNPC)

Comparación de las épocas 1 y 20 durante la fase de entrenamiento de los modelos.

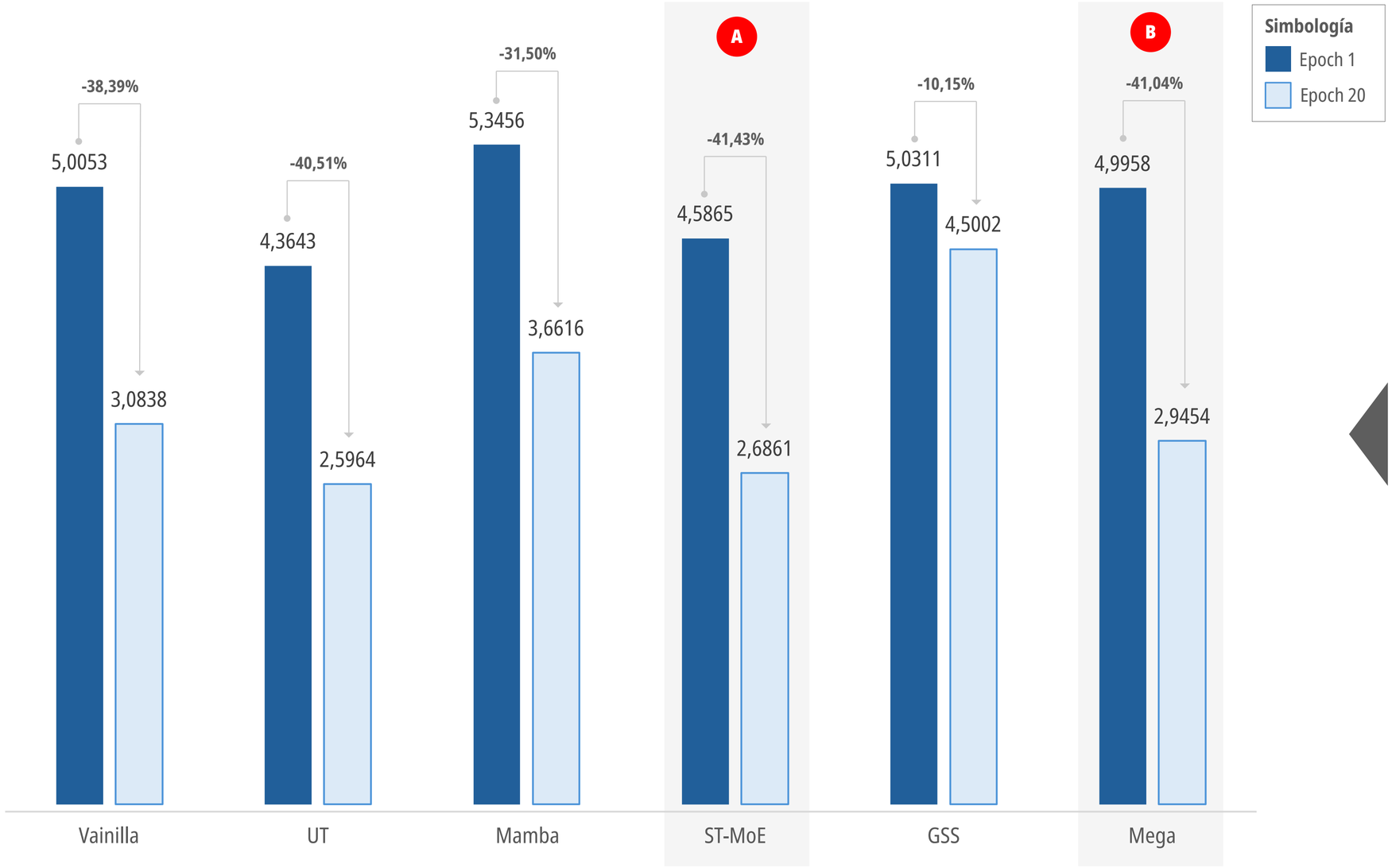


- El modelo MEGA destaca por lograr la mayor reducción de pérdida, con un -50,72 %, seguido de cerca por UT (-49,00 %) y ST-MoE (-47,89 %).
- Mamba, aunque no alcanza el descenso más pronunciado, muestra una mejora sustancial con un -42,85 %, confirmando su solidez como arquitectura eficiente tanto en desempeño como en velocidad de convergencia.

Fuente: Elaboración propia del estudiante.

Pérdida - Entrenamiento (OSPC)

Comparación de las épocas 1 y 20 durante la fase de entrenamiento de los modelos.

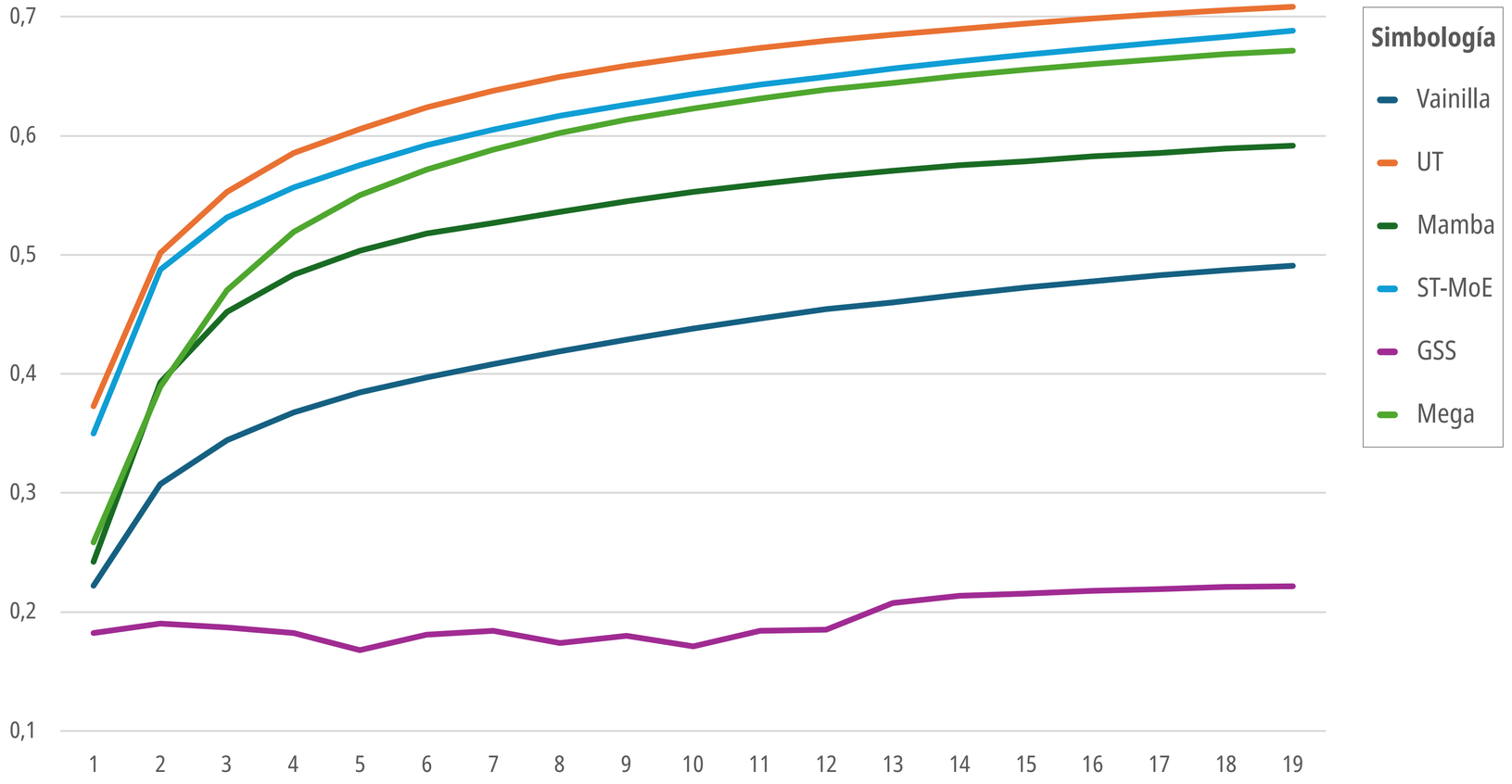


- ST-MoE (A), Mega (B) y UT alcanzan las reducciones más significativas de pérdida (más de un 40%).
- Vainilla también presenta una mejora sustancial, con una disminución de un 38,39%, aun cuando la arquitectura es la más simple y tiene un menor número de parámetros.

Fuente: Elaboración propia del estudiante.

Curvas de exactitud - Entrenamiento (UNPC)

Comportamiento de la medida de exactitud de los modelos durante el entrenamiento.

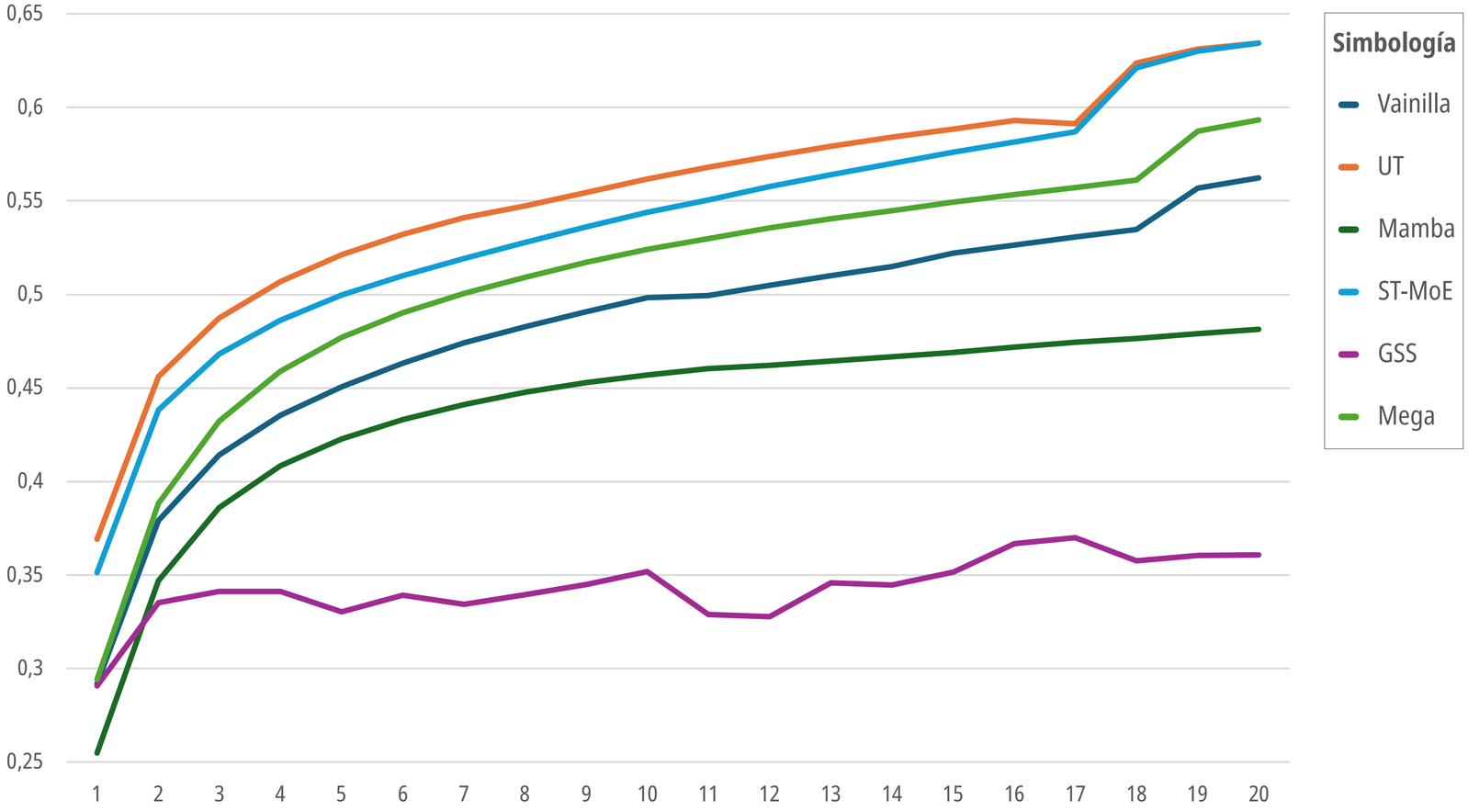


- Se observa un crecimiento acelerado en los modelos UT, MoE, Mamba y MEGA, todos los cuales superan rápidamente el umbral del 0,50 de exactitud.
- UT lidera consistentemente con una curva ascendente y sostenida, alcanzando el nivel más alto de exactitud al finalizar el entrenamiento.
- El modelo GSS presenta un desempeño claramente inferior.

Fuente: Elaboración propia del estudiante.

Curvas de exactitud - Entrenamiento (OSPC)

Comportamiento de la medida de exactitud de los modelos durante el entrenamiento.

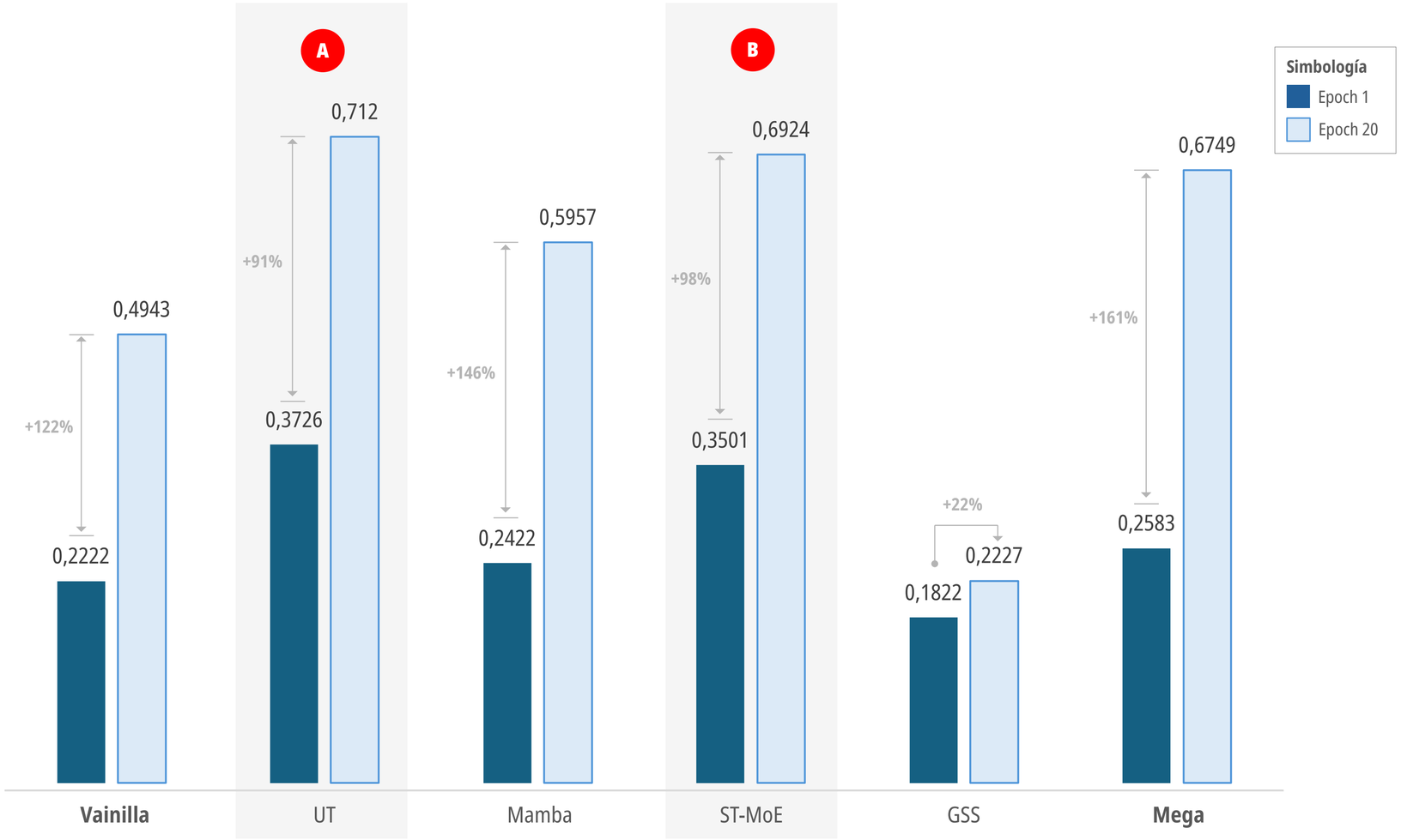


- MoE y UT se destacan por su rápido ascenso en la medida de exactitud, manteniéndose como los modelos con mejor desempeño general al alcanzar valores cercanos al 0,65.
- Mamba y MEGA también muestran un crecimiento constante, aunque con curvas ligeramente más moderadas.
- El modelo GSS y Vainilla presentan un desempeño claramente inferior.

Fuente: Elaboración propia del estudiante.

Exactitud - Entrenamiento (UNPC)

Comparación de las épocas 1 y 20 durante la fase de entrenamiento de los modelos.

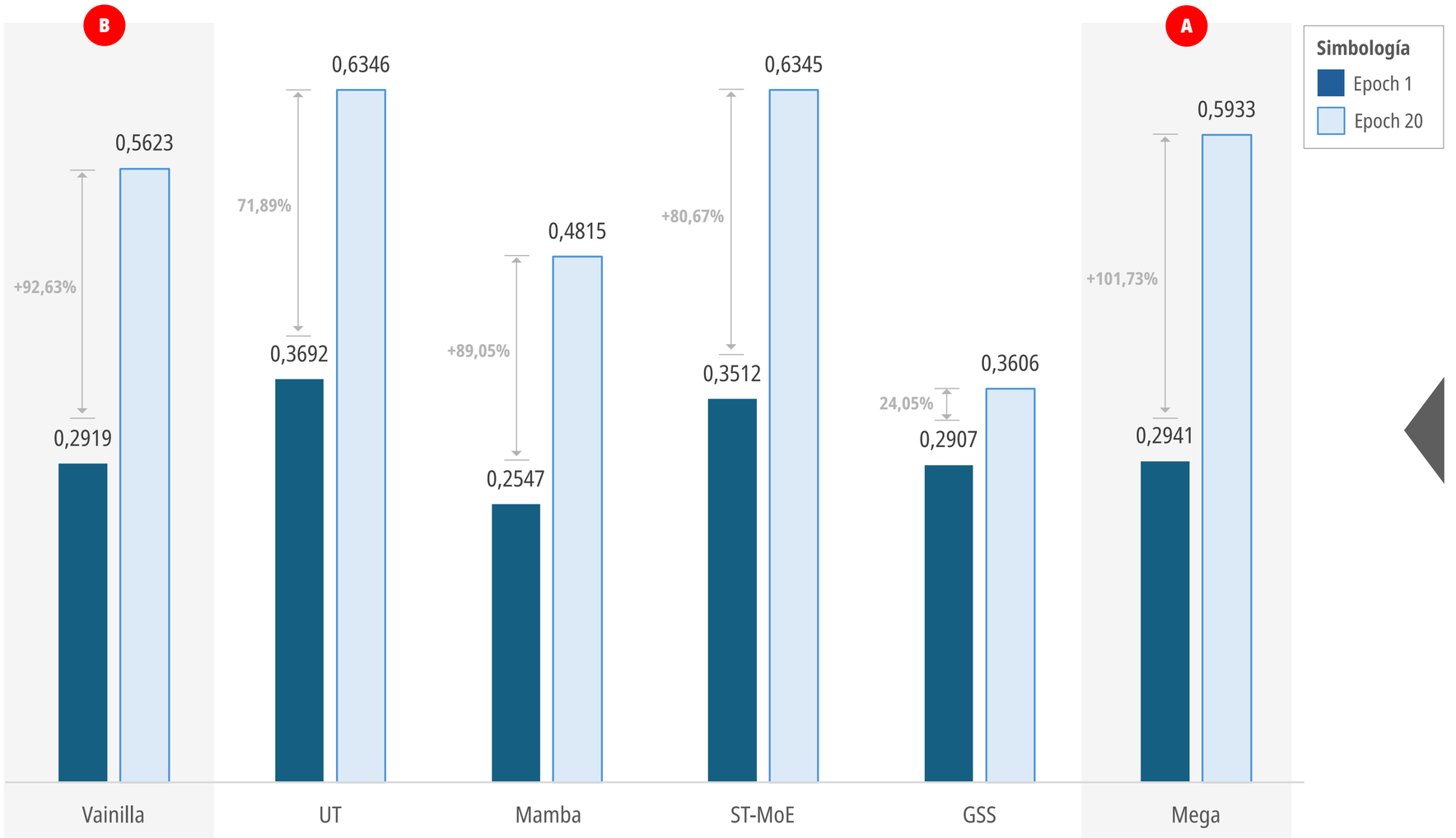


- MEGA y Mamba son los que presentan las mayores mejoras porcentuales, con un aumento del 161,29 % y 145,95 %
- MoE y UT también muestran mejoras significativas (97,77 % y 91,09 %).
- El modelo Vainilla mejora un 122,46 %, una cifra sorprendente considerando su simplicidad estructural, aunque partiendo desde una precisión inicial muy baja.

Fuente: Elaboración propia del estudiante.

Exactitud - Entrenamiento (OSPC)

Comparar la exactitud entre modelos proporciona una medida objetiva del rendimiento relativo de cada uno.

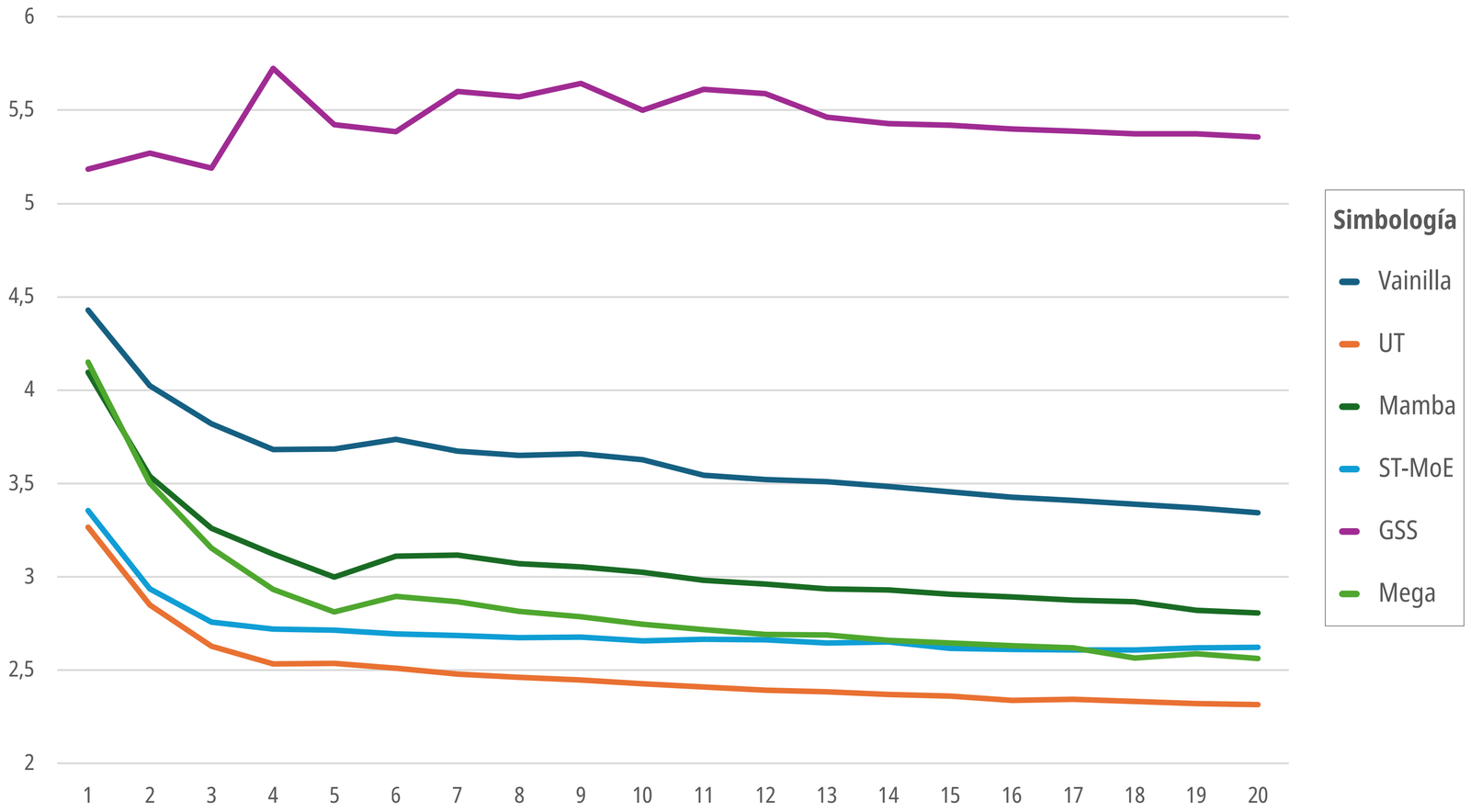


- El modelo Mega destaca como el que mayor incremento de exactitud presenta, pasando de 0,2941 a 0,5933, lo que representa un aumento del +101,73 %, seguido muy de cerca por Vainilla (+92,63 %) y Mamba (+89,05 %).
- Esto evidencia que incluso modelos con arquitecturas relativamente simples (como Vainilla) pueden lograr grandes avances durante el entrenamiento, cuando están bien optimizados.

Fuente: Elaboración propia del estudiante.

Curvas de Pérdida - Validación (UNPC)

Comportamiento de la medida de pérdida de los modelos durante la validación.

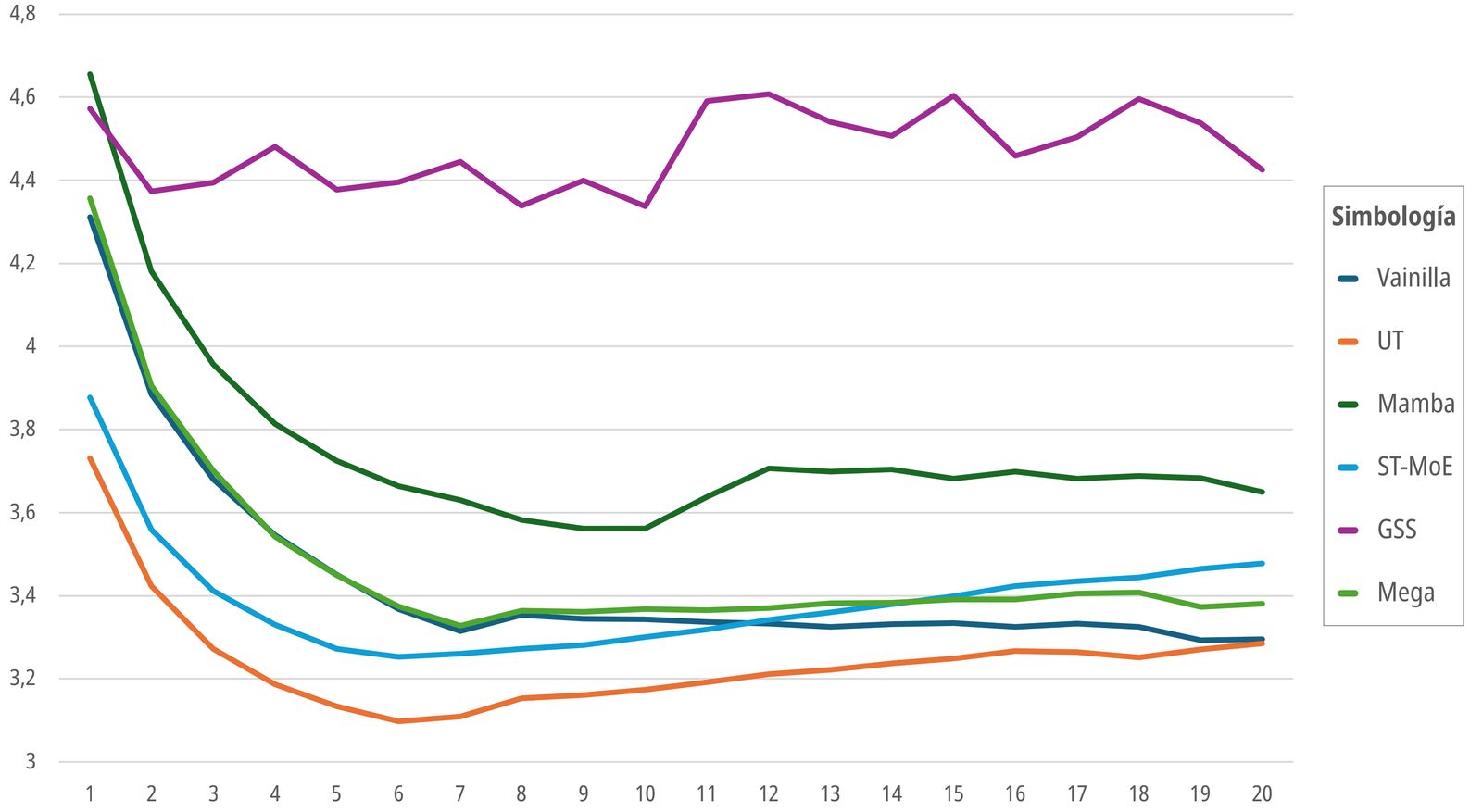


Nota: Una menor pérdida significa que el modelo está haciendo predicciones más precisas o está más cerca de la solución óptima.
Fuente: Elaboración propia del estudiante.

- UT, MoE y MEGA mantienen las curvas más bajas y estables a lo largo de las 20 épocas, lo que indica un desempeño sólido y consistente ante datos nuevos. UT, en particular, destaca por lograr la menor pérdida en validación.
- Vainilla muestra una pérdida claramente más elevada y una curva menos eficiente en su descenso, lo que evidencia una menor habilidad para generalizar.
- La curva de GSS se mantiene alta, con fluctuaciones significativas y sin una reducción clara, lo que sugiere inestabilidad durante la validación y dificultades para aprender patrones generalizables del corpus.

Curvas de Pérdida - Validación (OSPC)

Comportamiento de la medida de pérdida de los modelos durante la validación.

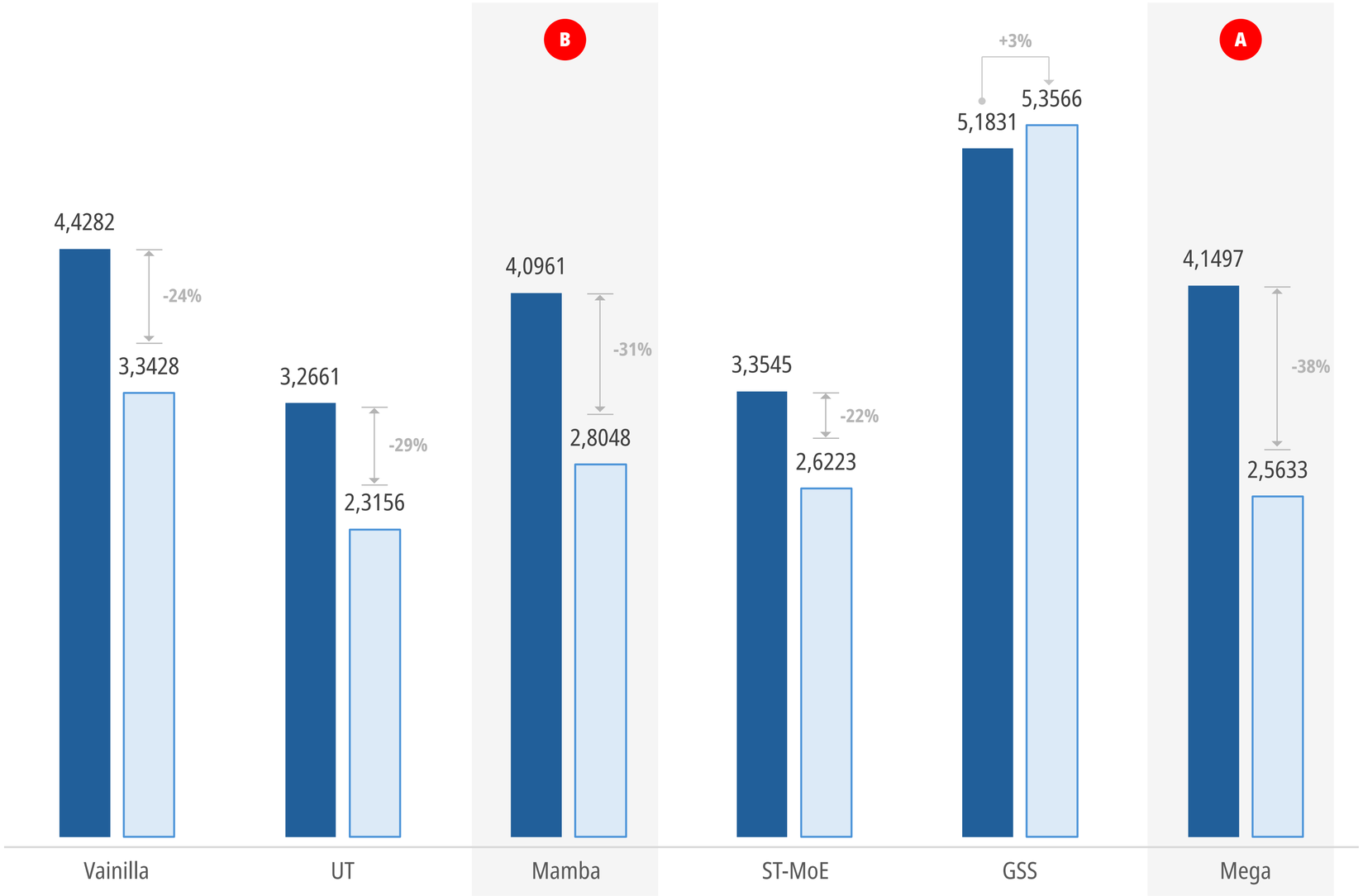


- MoE, UT y Mamba mantienen un desempeño favorable, con curvas de pérdida que tienden a descender de manera sostenida y alcanzan valores bajos en las últimas etapas.
- MEGA y Vainilla son más irregulares y mantienen valores de pérdida más elevados.
- GSS presenta nuevamente el comportamiento más problemático. Su curva de pérdida es elevada, con fuertes oscilaciones y picos abruptos entre las épocas 5 y 13.

Nota: Una menor pérdida significa que el modelo está haciendo predicciones más precisas o está más cerca de la solución óptima.
Fuente: Elaboración propia del estudiante.

Pérdida - Validación (UNPC)

Comparación de pérdida en las épocas 1 y 20 de la fase de validación de los modelos.

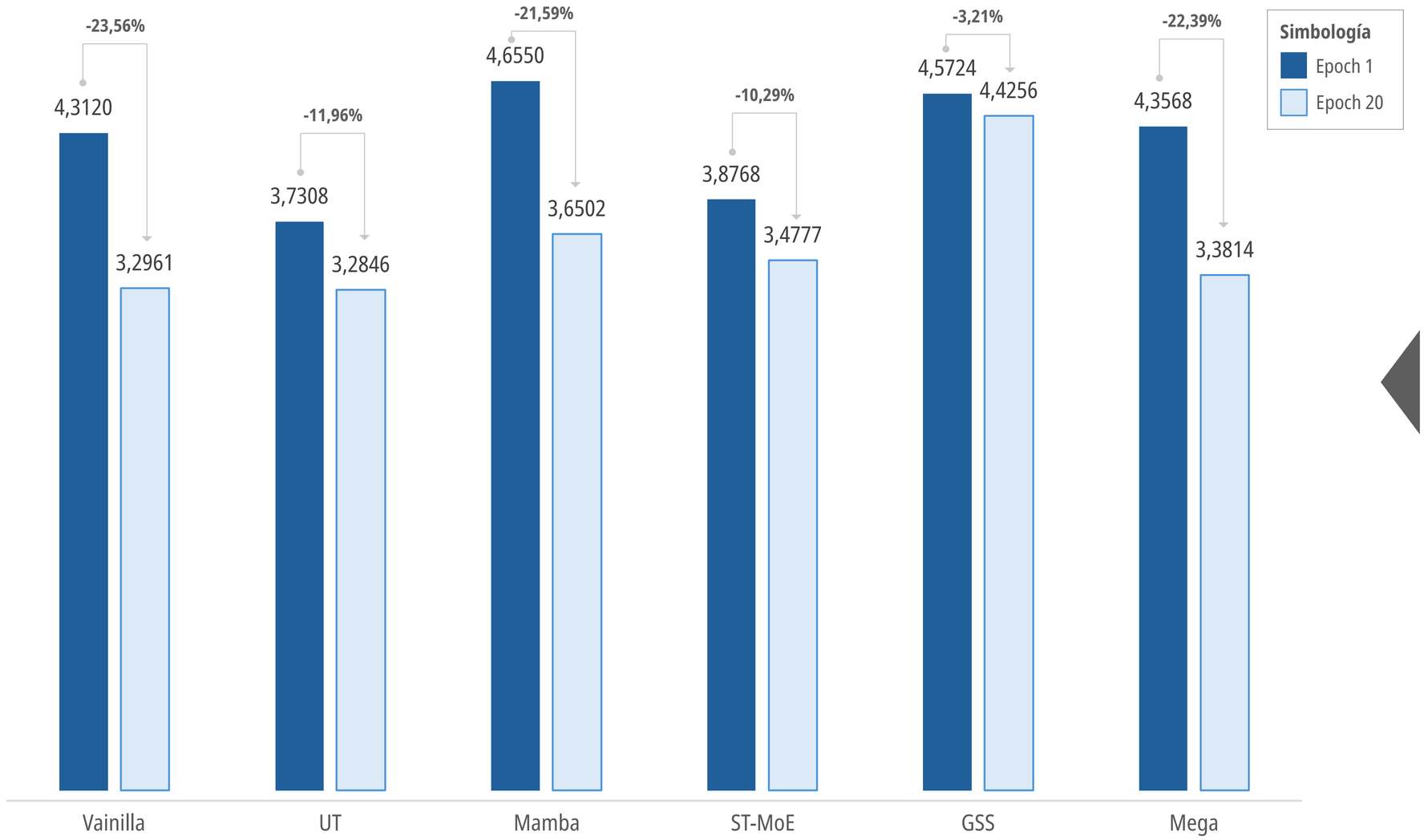


- Mega destaca con la mayor mejora porcentual (-38%), terminando con una pérdida muy baja. UT y Mamba también muestran reducciones significativas y estables.
- MoE tiene una reducción más modesta, pero logra una de las menores pérdidas absolutas.
- Vainilla, aunque mejora, se mantiene en niveles altos de pérdida.
- Lo esperado es que la pérdida disminuya con el entrenamiento. GSS es el único que incrementa su pérdida en validación lo que indica sobreajuste. ento en GSS sugiere que no generaliza bien, a pesar de entrenarse.

Nota: Una menor pérdida significa que el modelo está haciendo predicciones más precisas o está más cerca de la solución óptima. Fuente: Elaboración propia del estudiante.

Pérdida - Validación (OSPC)

Comparación de pérdida en las épocas 1 y 20 de la fase de validación de los modelos.

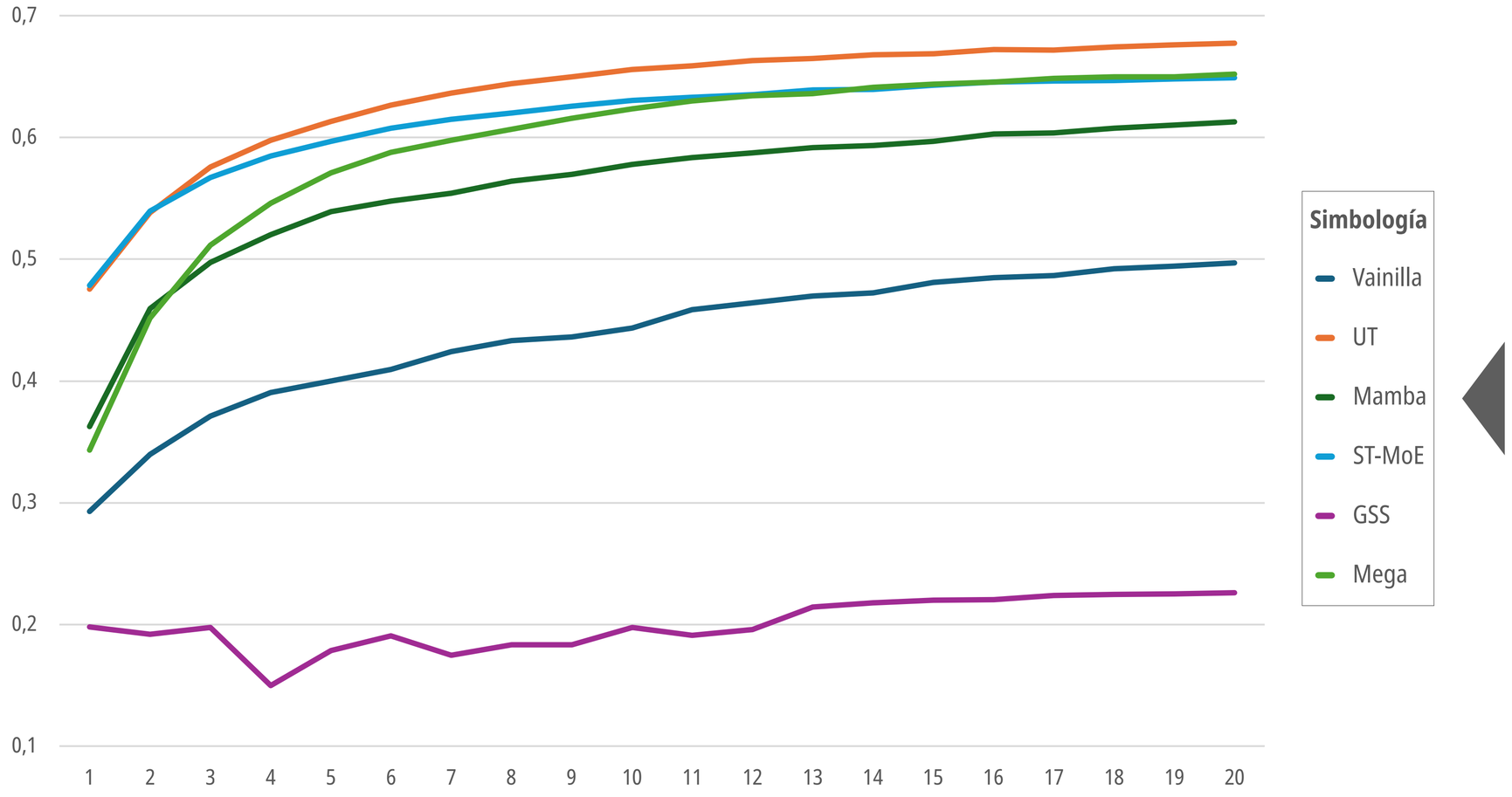


- Vainilla, con una reducción del -23,56 %, mejorara de forma sostenida y generalizar adecuadamente. Mega, con un descenso del -22,39 %, lo que respalda su competitividad en tareas prácticas más allá del entrenamiento.
- Mamba, con una mejora del -21,59 %, confirma su solidez no solo como arquitectura eficiente, sino también robusta en términos de generalización
- El modelo UT, que había mostrado un excelente desempeño en pérdida de entrenamiento, presenta en este caso una mejora más modesta de -11,96 %, lo cual podría indicar un grado mayor de sobreajuste

Nota: Una menor pérdida significa que el modelo está haciendo predicciones más precisas o está más cerca de la solución óptima.
Fuente: Elaboración propia del estudiante.

Curvas de exactitud - Validación (UNPC)

Comportamiento de la medida de exactitud de los modelos durante la validación.

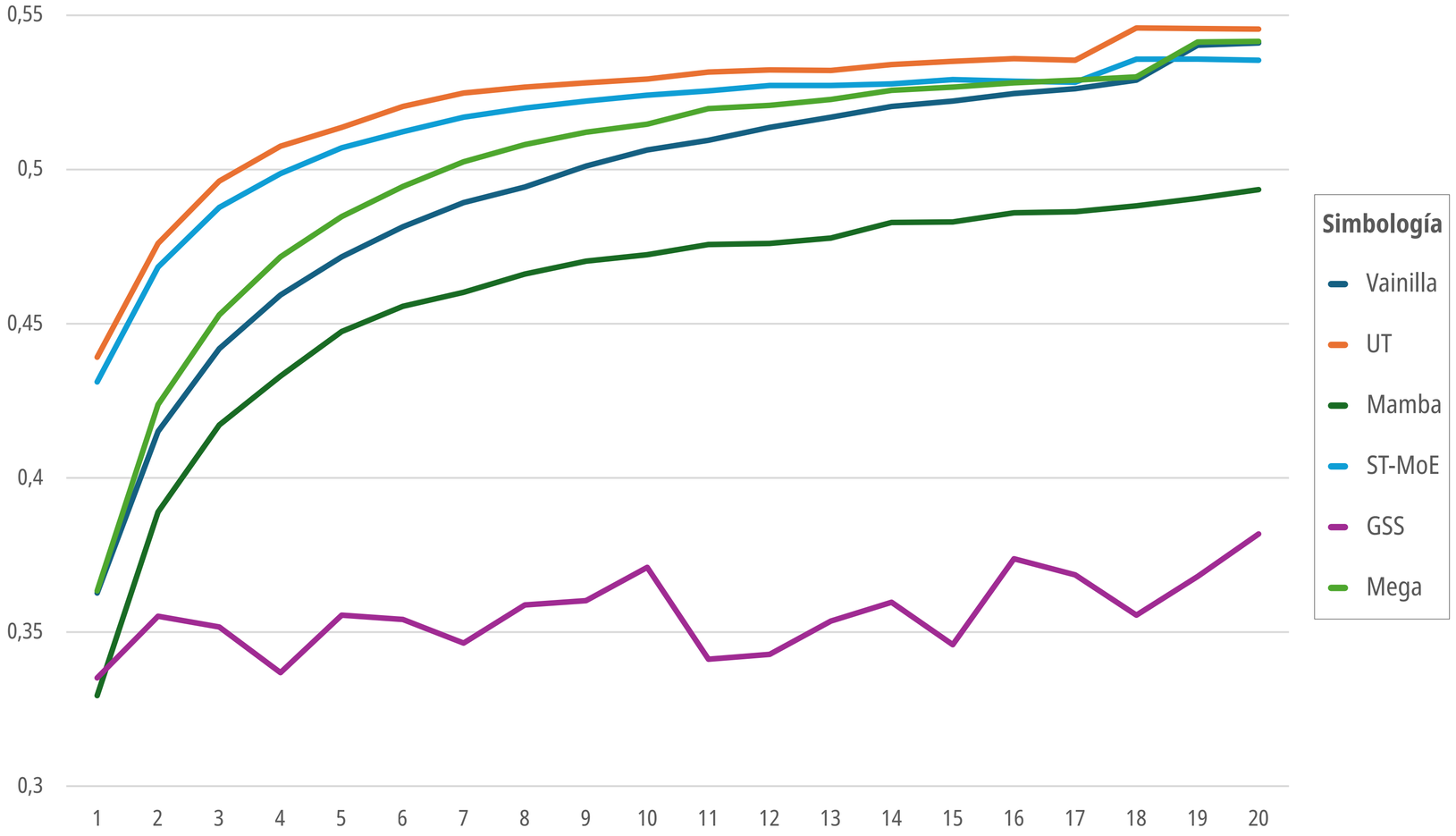


- UT, MoE y Mamba destacan claramente por sus curvas ascendentes sostenidas, alcanzando niveles de exactitud superiores al 0,60 al finalizar el proceso.
- UT lidera la comparación, seguido de cerca por MoE, cuya curva es ligeramente más estable.
- Mamba mantiene una evolución constante y competitiva, confirmando su buena capacidad de generalización en entornos formales.
- MEGA se sitúa ligeramente por debajo, pero su desempeño es consistente a lo largo de las épocas.

Nota: (1) El gap surge de comparar la métrica de exactitud del modelo durante el entrenamiento versus la validación (ejemplos no vistos).
Fuente: Elaboración propia del estudiante.

Curvas de exactitud - Validación (OSPC)

Comportamiento de la medida de exactitud de los modelos durante la validación.

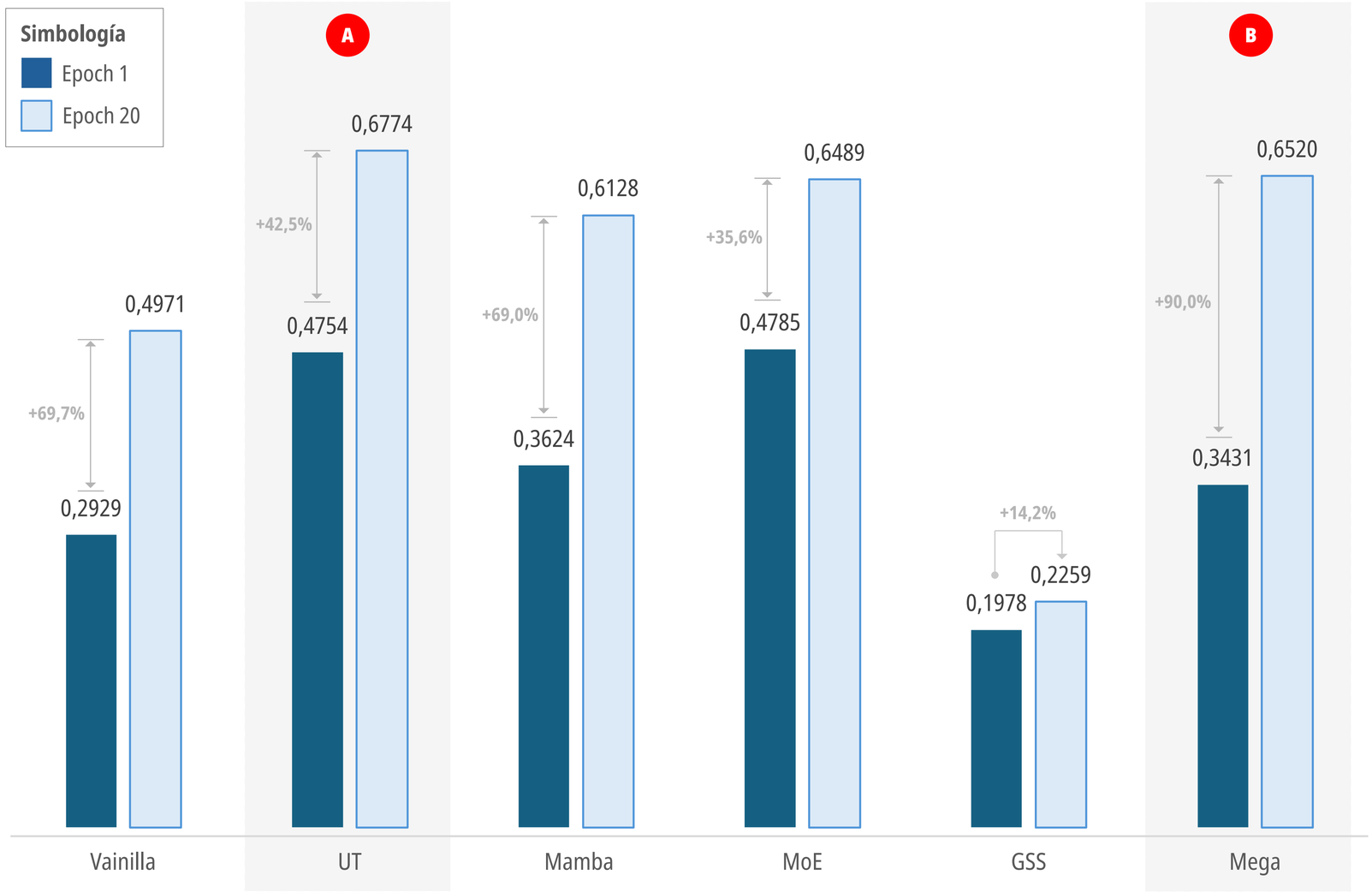


- UT se posiciona como el más destacado, con una rápida y sostenida mejora que lo lleva a superar el 0,55 de exactitud, manteniéndose como el líder a lo largo de todo el entrenamiento.
- MoE le sigue con una curva ascendente y estable, que alcanza valores cercanos al 0,48, confirmando su solidez incluso en un corpus ruidoso.
- Mamba y MEGA también muestran una evolución progresiva, logrando niveles competitivos entre 0,45 y 0,46 hacia el final de la validación

Nota: (1) El gap surge de comparar la métrica de exactitud del modelo durante el entrenamiento versus la validación (ejemplos no vistos).
Fuente: Elaboración propia del estudiante.

Exactitud - Validación (UNPC)

Comparación entre las épocas 1 y 20 de los modelos de la comparativa.

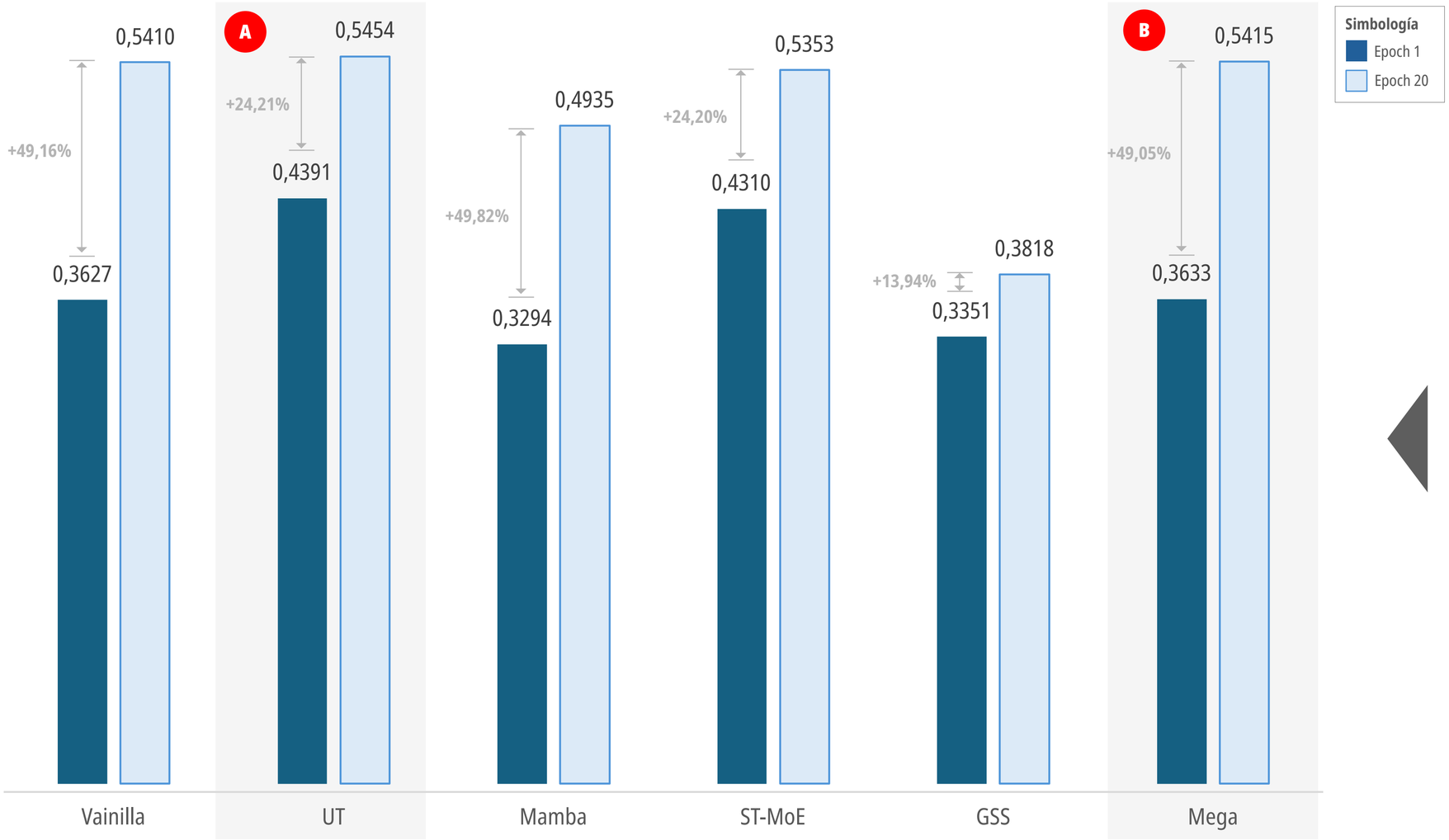


- Modelos como UT (42,49 %) y MoE (35,61 %), aunque robustos en precisión final, muestran una menor velocidad de mejora en validación, lo que sugiere una curva de aprendizaje más gradual.
- El modelo con mayor mejora relativa en validación fue MEGA, con un incremento del 90,03 %, seguido por Vainilla (69,72 %) y Mamba (69,09 %).

Nota: El gap surge de comparar la métrica de exactitud del modelo durante la época 1 y 20 de la fase de validación (ejemplos no vistos).
Fuente: Elaboración propia del estudiante.

Exactitud - Validación (OSPC)

Comparación entre las épocas 1 y 20 de los modelos de la comparativa.

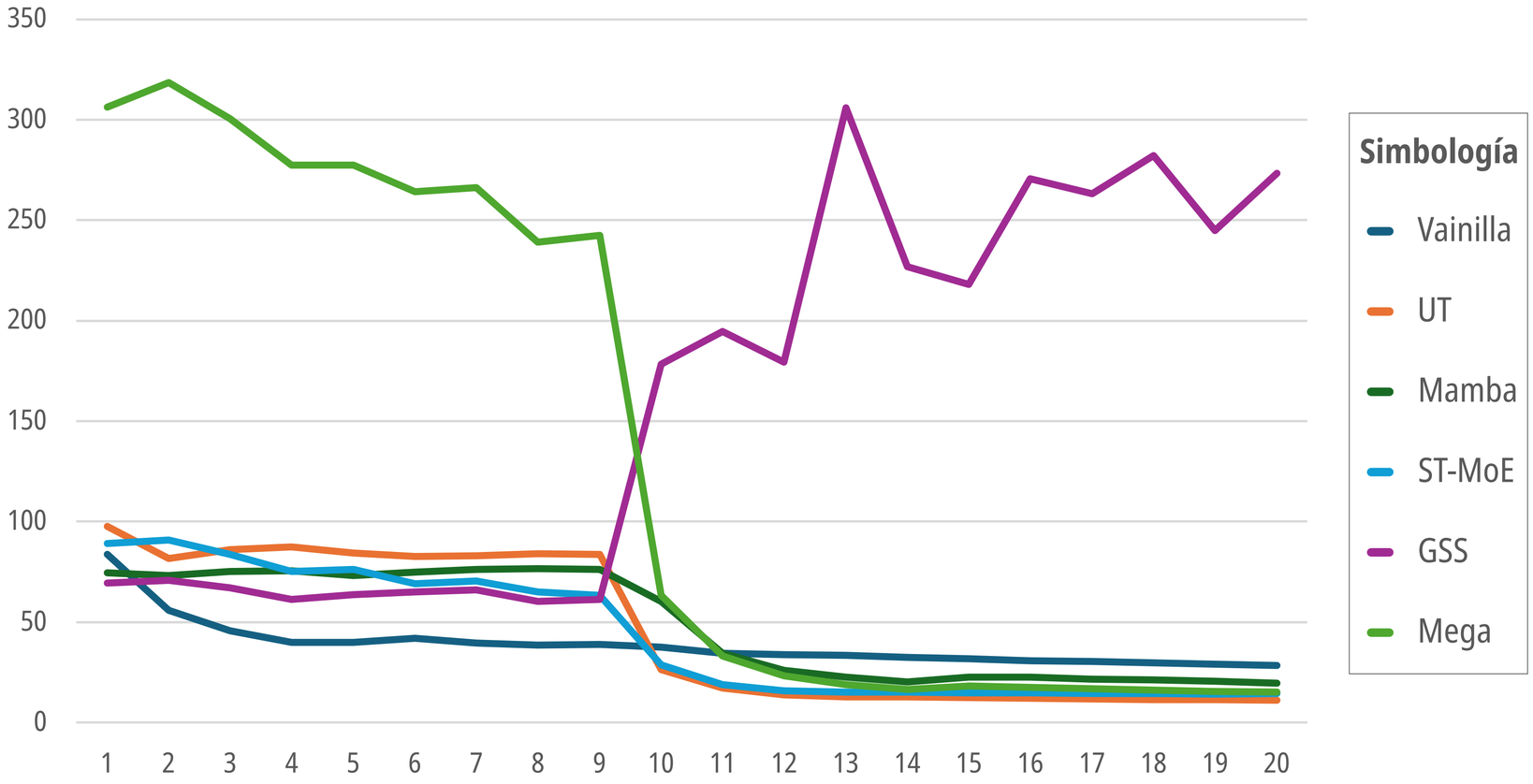


- Mamba, Vainilla y Mega muestran los mayores incrementos relativos en exactitud por encima del 49%.
- El modelo UT, si bien presenta la exactitud más alta al final de la validación (0,5454), muestra una mejora más contenida en términos relativos (+24,21 %), al igual que ST-MoE con +24,20 %.
- Aunque UT y ST-MoE lideran en valores absolutos de exactitud, Mamba, Mega y Vainilla logran avances porcentuales sobresalientes.

Nota: El gap surge de comparar la métrica de exactitud del modelo durante la época 1 y 20 de la fase de validación (ejemplos no vistos).
Fuente: Elaboración propia del estudiante.

Perplejidad - Validación (UNPC)

La diferencia entre la perplejidad de entrenamiento y validación es crucial para evaluar la capacidad de generalización del modelo y para identificar problemas como el sobreajuste o el subajuste.

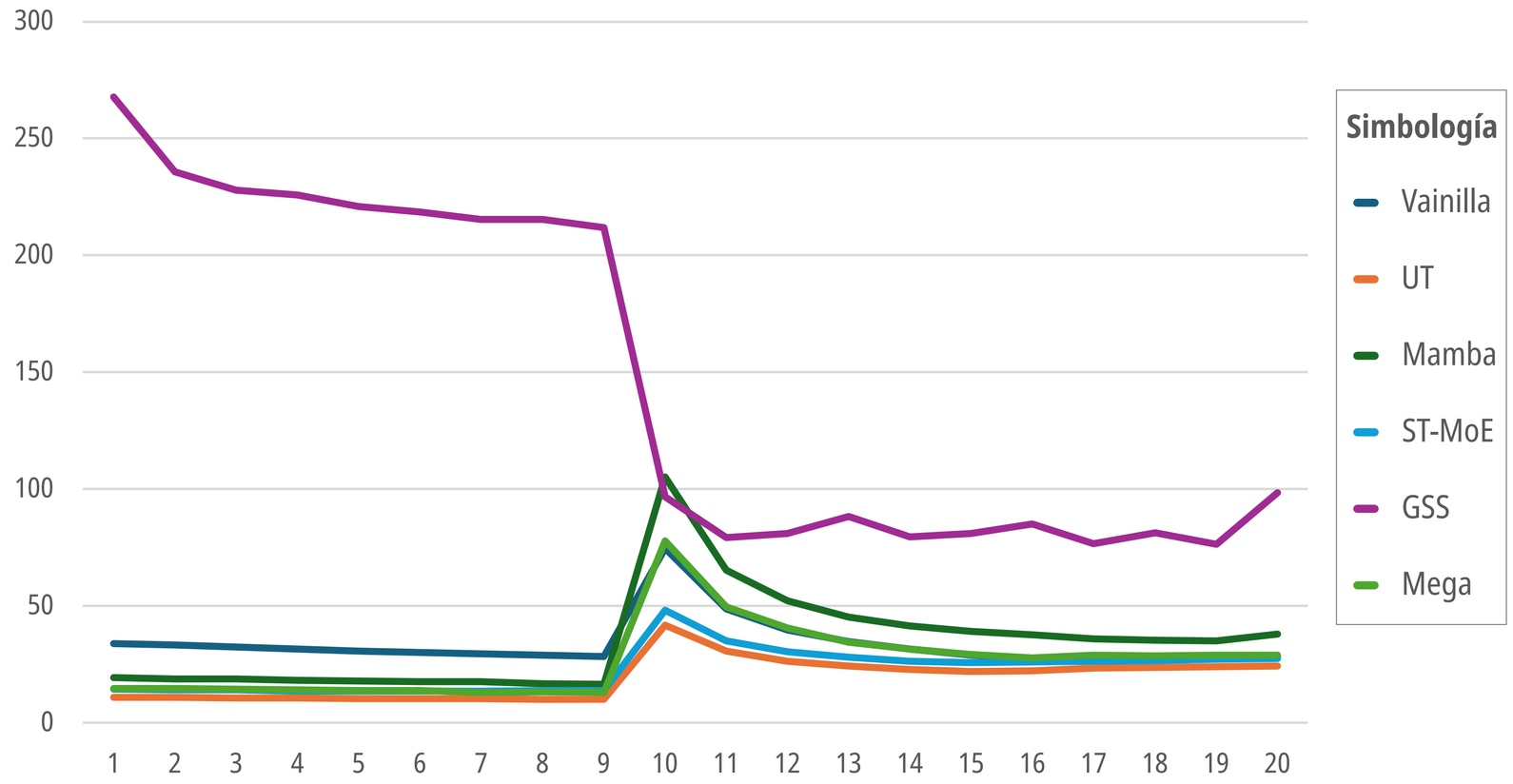


- Los modelos UT, MoE, Mamba y MEGA presentan los valores de perplejidad más bajos y estables hacia el final del entrenamiento, lo que indica un buen equilibrio entre aprendizaje y generalización.
- MEGA, a pesar de un inicio más inestable, logra reducir abruptamente su perplejidad a partir de la época 10.
- Vainilla mantiene una curva más elevada (alrededor de 40), lo que indica una menor capacidad de predicción en validación, aunque con un comportamiento relativamente estable.

Nota: Una perplejidad baja indica que el modelo puede predecir con confianza las palabras siguientes en una secuencia, mientras que una perplejidad alta indica incertidumbre en estas predicciones. Fuente: Elaboración propia del estudiante.

Perplejidad - Validación (OSPC)

La diferencia entre la perplejidad de entrenamiento y validación es crucial para evaluar la capacidad de generalización del modelo y para identificar problemas como el sobreajuste o el subajuste.



- MoE, UT, Mamba y MEGA logran mantener la perplejidad bajo control, con oscilaciones puntuales pero dentro de márgenes razonables, y tienden a estabilizarse hacia el final. Este comportamiento sugiere una adaptabilidad razonable a los datos ruidosos del corpus OSPC.
- Vainilla sigue una trayectoria similar, con una perplejidad ligeramente superior de manera sostenida, lo que indica menor capacidad de predicción bajo incertidumbre
- Por su parte, GSS presenta una curva dramáticamente inestable.

Nota: Una perplejidad baja indica que el modelo puede predecir con confianza las palabras siguientes en una secuencia, mientras que una perplejidad alta indica incertidumbre en estas predicciones. Fuente: Elaboración propia del estudiante.

Inferencia - Ejemplos de resultados (UNPC)

Origen (EN)	GT (ES)	Modelo	Resultado de la inferencia	BLEU-4
the households which estimated their economic situation as almost poor were those with unemployed members and without any working or pensionreceiving members per cent households maintained on unearned income other than pensions per cent and mothers raising children per cent	los hogares que estimaron que su situacion economica era casi pobre eran los hogares con miembros desempleados y sin ningun miembro que recibiese pension los mantenidos mediante ingresos no provenientes del trabajo ni de pensiones y las madres con hijos a su cargo	**Vainilla**	los ingresos no utilizados en el trabajo ni en el programa de familias que viven en condiciones de vida dificiles son menos del de los ingresos y el de los ingresos que se encuentran en situacion economica y social	0,4127
		Universal transformer	los hogares que estiman que su situacion economica era casi pobres era de los desempleados con miembros desempleados y sin ningun miembro que de los hogares con ingresos no y con ingresos ni os	0,6378
		Mamba	los familias que mostraban su situacion economica como practicamente practicamente son los pobres y sin ningun miembro o los familias que se encontraban en el caso de los familias no electricas el y la licencia de los ni os de la casa	0,4870
		ST-MoE	los hogares que su situacion economica era casi siempre que casi siempre eran los miembros desempleados y sin ningun miembro que recibiese pension los hogares habian mantenido en los ingresos no que no estan en condiciones de vida del y de las madres con	**0,7780**
		GSS	en algunos casos de los ni de la hogares de los ni de la casos de los ni de la hogares de los ni de los ni de la hogares de los ni de la hogares de los ni de la caso de los ni de	0,2295
		MEGA	los hogares que se han reducido su situacion economica como casi los ni os que han de ser miembros de su familia o que no estan en condiciones de favor de los ingresos no utilizados por las pensiones y las madres	0,5575

Inferencia - Ejemplos de resultados (UNPC)

Origen (EN)	GT (ES)	Modelo	Resultado de la inferencia	BLEU-4
the board also recommends that the administration review the project management report sheet and include the specific outputs in it in addition to ensuring that these are updated on an ongoing basis to serve as an effective project management instrument	la junta recomienda tambien que la administracion examine el resumen analitico de gestion de proyectos e incluya en el los productos especificos ademas de velar por que se actualicen a titulo continuo para que sirvan como instrumento eficaz de gestion de proyectos	**Vainilla**	ademas la junta recomienda que se siga examinando la cuestion de la gestion de los recursos humanos y que se incluya en el proyecto de presupuesto por programas para el bienio un nuevo sistema de gestion de los bienes que se le	0,5644
		Universal transformer	la junta recomienda tambien que la administracion examine el informe del proyecto y incluya los resultados concretos en el documento de gestion de los proyectos y que se incluya los resultados concretos a fin de asegurar que estas son actualizada	0,6221
		Mamba	la junta recomienda tambien que se establezcan el examen de la administracion el informe de gestion de proyectos y incluyen los productos concretos en ademas de garantizar que estas son un instrumento eficaz de gestion de proyectos	**0,6718**
		ST-MoE	la junta recomienda tambien que la administracion examinara el informe sobre gestion de los proyectos y se incluya los productos concretos en el curso de manera que se <unk> de manera permanente para prestar servicios de gestion	0,5901
		GSS	la junta tomo que se servicios de la secretaria de las partes de la convencion de la secretaria de los programas de la secretaria de la secretaria de la secretaria de la secretaria de las partes y y de la secretaria de las	0,3714
		MEGA	la junta recomienda tambien que el informe sobre gestion del proyecto y incluya los productos concretos de la administracion y incluya los productos concretos de ese tipo y ademas de asegurar que esos datos se <unk> de manera permanente para que puedan ser	0,5688

Inferencia - Ejemplos de resultados (UNPC)

Origen (EN)	GT (ES)	Modelo	Resultado de la inferencia	BLEU-4
it is true that the conference on disarmament has its merits and it has a universal calling	es indudable que la conferencia de desarme tiene sus meritos y su labor tiene un caracter universal	**Vainilla**	la conferencia de desarme esta dispuesta a hacer una declaracion y a hacer todo lo posible por hacer	0,4849
		Universal transformer	es cierto que la conferencia de desarme tiene sus fondo y que se le ha solicitado	0,6700
		Mamba	es probable que la conferencia de desarme tiene su fondo y ha sido un	0,6136
		ST-MoE	es verdad que la conferencia de desarme tiene sus objetivos y tiene una cuestion universal	**0,6991**
		GSS	en ese contexto el orador de la india de material fisible de armas nucleares de armas nucleares de armas nucleares	0,0000
		MEGA	es cierto que la conferencia de desarme tiene su fondo y tiene un derecho universal	**0,7121**

Inferencia - Ejemplos de resultados (OSPC)

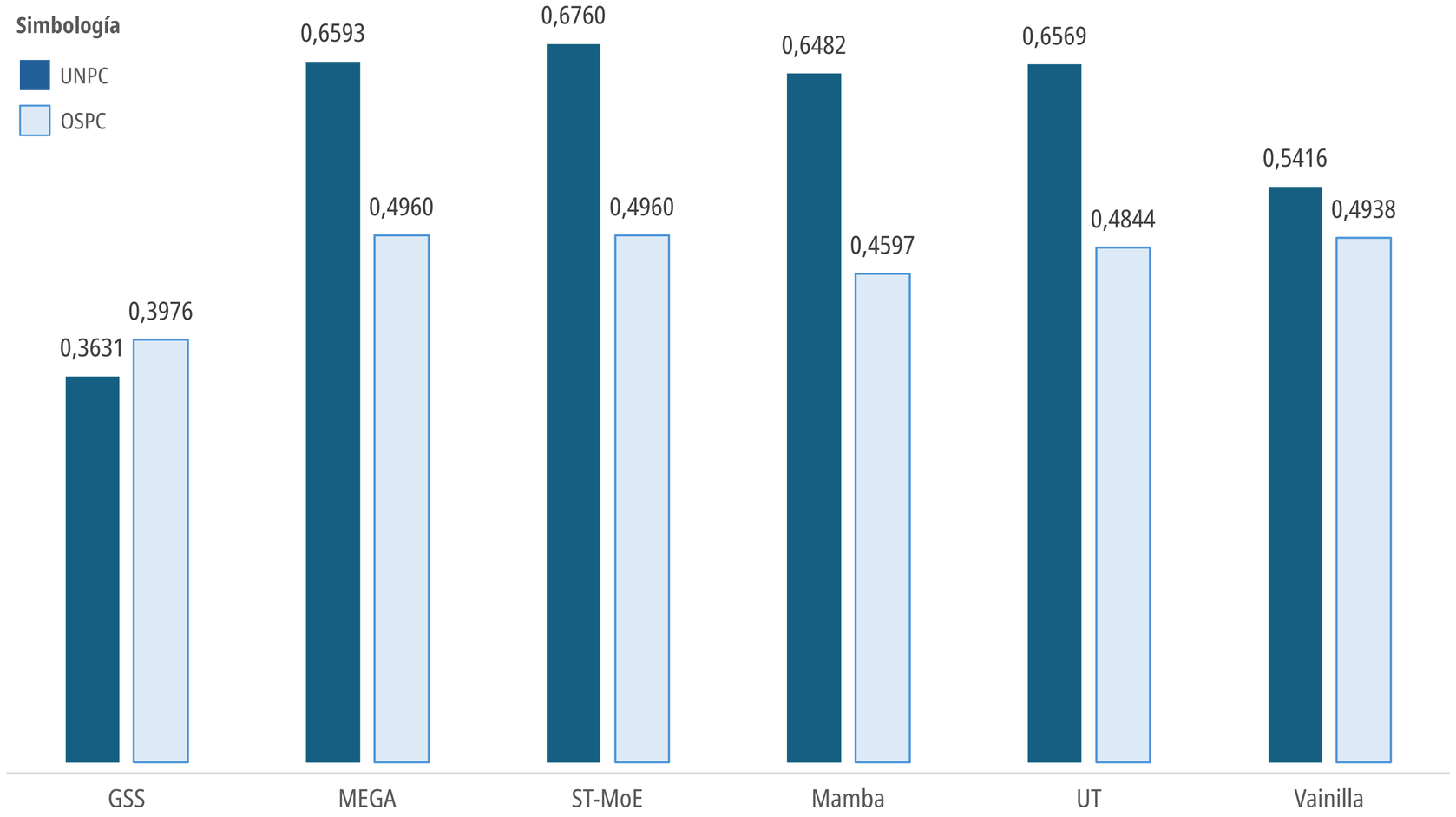
Origen (EN)	GT (ES)	Modelo	Resultado de la inferencia	BLEU-4
you could offer me the entire casino the amusement park and the rothschilds fortune and i would still refuse you hear	podria ofrecerme el casino entero el parque de atracciones y la fortuna de los rothschilds y todavia me negaria me oye	**Vainilla**	podria ofrecerle las cuentas y el casino y todavia me niego a oir la fortuna	0,4442
		Universal transformer	podria ofrecerle el casino de todo el casino y la fortuna aun asi te niego a oir una cosa	0,4816
		Mamba	podrias prestarme mi este casino el parque y la fortuna de la fortuna de la fortuna y yo te dejaria decir	**0,5199**
		ST-MoE	podrias comprar toda la fortuna del casino park y la fortuna del casino y aun asi te niego	0,3587
		GSS	podria darme la historia de la que me he hecho y me me me me de la	0,3096
		MEGA	podria darme el casino y la fortuna de la se ora	0,3144

Inferencia - Ejemplos de resultados (OSPC)

Origen (EN)	GT (ES)	Modelo	Resultado de la inferencia	BLEU-4
he does not want anybody to get fired on by the soldiers but if any of us were in there he had go in and get us soldiers or no soldiers	no quiere que los soldados disparen a nadie pero si fuera uno de nosotros entraria y lo sacaria	**Vainilla**	no quiere que nadie nos de soldados pero si estuviera en soldados y nos ha despedido	0,5510
		Universal transformer	no quiere que nadie nos la disparo por los soldados pero si alguno de nosotros tuviera soldados	0,6598
		Mamba	no quiere que nadie se aleje en los soldados pero si ninguno de nosotros estaban ahi	0,6232
		ST-MoE	no queria que nadie se disparo con soldados pero si alguno de nosotros estuviera ahi dentro y que nos de soldados	0,5053
		GSS	no quiere que nadie que el y o de los ciudad pero de los hombres de la calle pero si de la ciudad de la ciudad	0,4912
		MEGA	no quiere que nadie se lo haya despedido por los soldados pero si alguno de nosotros estuviera ahi y nos podria matar soldados o no	0,5881

Inferencia - BLEU4

Medida cuantitativa de qué tan cercanas son las traducciones automáticas generadas por el modelo a las traducciones de referencia realizadas por humanos.

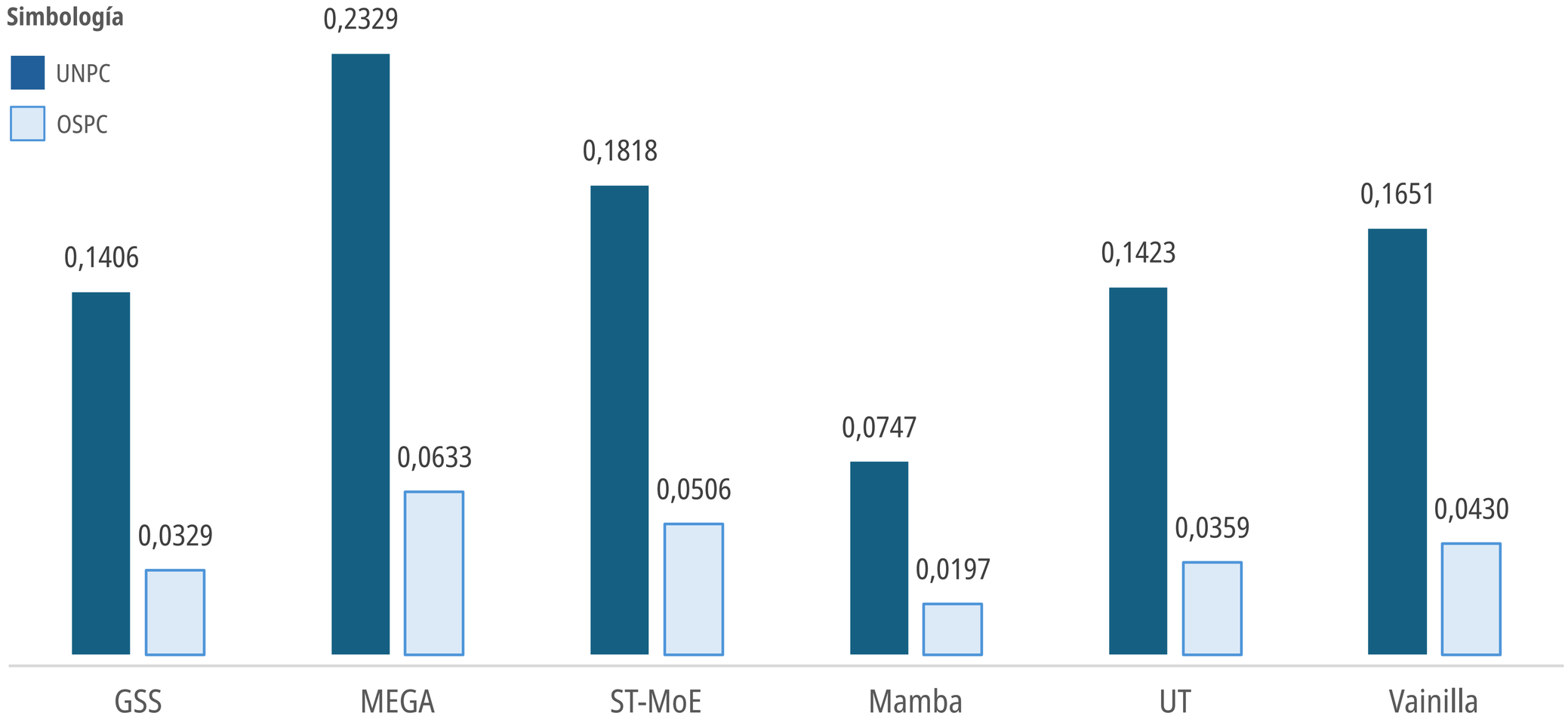


- Los modelos con mayor puntuación BLEU son MoE (0,6760), MEGA (0,6593) y UT (0,6569).
- Esto confirma su capacidad mejorada para generar traducciones precisas y naturales en un contexto de lenguaje formal.
- Sin embargo, Mamba también alcanza un desempeño sobresaliente con un BLEU de 0,6482, superando al modelo Transformer base (Vainilla).

Fuente: Elaboración propia del alumno.

Inferencia - Tiempo promedio por consulta

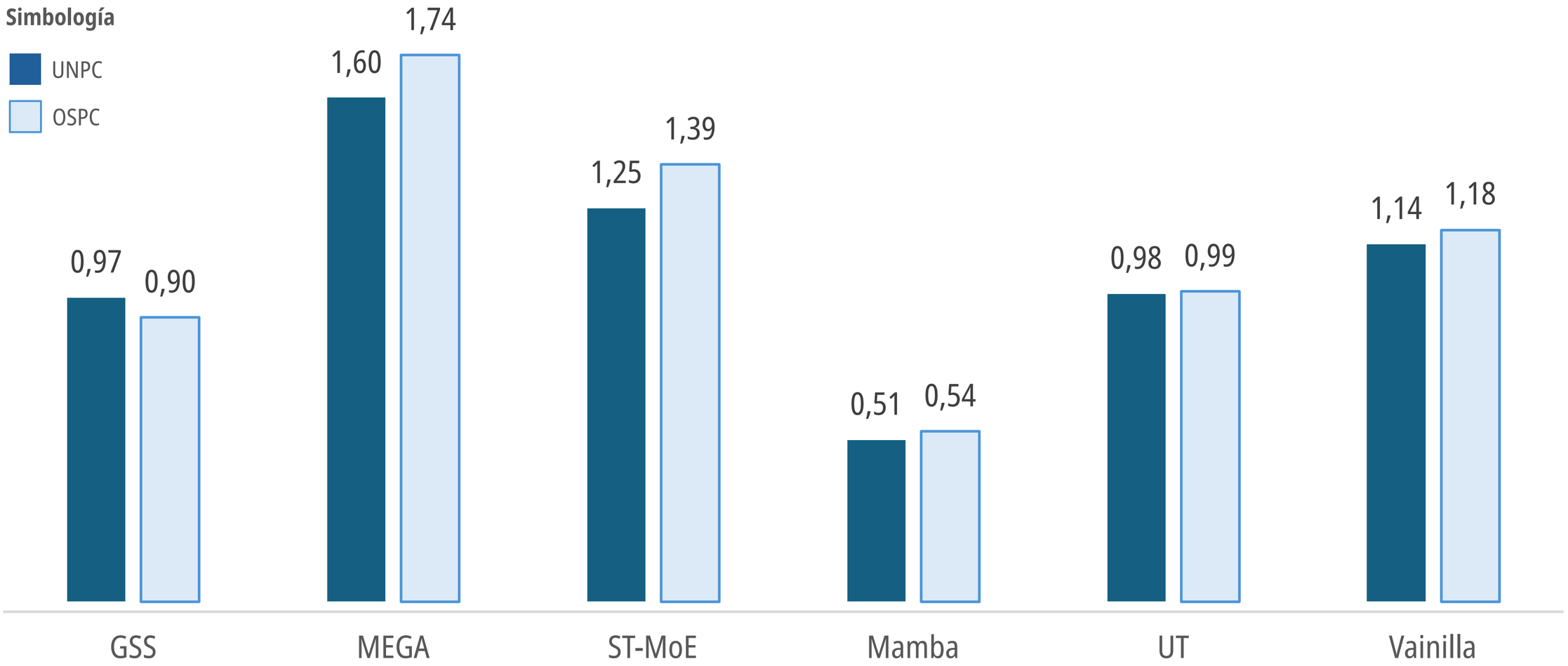
Comparar el tiempo de inferencia entre diferentes modelos puede ayudar a seleccionar el modelo más adecuado para una aplicación específica.



Nota: Valores absolutos dados en segundos, tiempo promedio para cada ejemplo utilizando una muestra de 10000 ejemplos seleccionados aleatoriamente una vez y aplicada por igual a todos los modelos. (1) y (2) Se componen dos modelos ST independientes en una solución de NMT. Fuente: Elaboración propia del alumno.

Inferencia - Tiempo total

Comparar el tiempo de inferencia entre diferentes modelos puede ayudar a seleccionar el modelo más adecuado para una aplicación específica.



Nota: Valores absolutos datos en horas, para procesar 10000 ejemplos seleccionados aleatoriamente una vez y aplicados por igual a todos los modelos.
(1) y (2) Se componen dos modelos ST independientes en una solución de NMT. Fuente: Elaboración propia del alumno.

Conclusiones

Las variantes más recientes superan su rendimiento en precisión, capacidad de generalización y eficiencia computacional, especialmente cuando se trata de adaptarse a distintos dominios lingüísticos.

Las conclusiones más importantes:

- Modelos como Universal Transformer, MEGA, MoE y Mamba introducen mecanismos de reutilización de capas, atención jerárquica, gating y mezcla de expertos, que optimizan el uso de recursos computacionales (menor tiempo de entrenamiento, validación e inferencia), y reducen la cantidad de parámetros entrenables, manteniendo o incluso mejorando la calidad de la traducción.

- Mientras el Transformer original muestra un rendimiento aceptable en corpus estructurados (como UNPC), su precisión y capacidad de generalización se degradan en corpus ruidosos o informales (como OSPC). En contraste, modelos como MoE, UT y Mamba mantienen resultados más consistentes en ambos entornos, demostrando una mayor robustez transversal.

- Las curvas de exactitud muestran que las variantes convergen más rápido y alcanzan mayores niveles de acierto tanto en entrenamiento como en validación. Asimismo, la perplejidad es sustancialmente menor y más estable en modelos como MoE, UT y Mamba, lo que indica que generan predicciones más confiables y menos inciertas.

- Entre las innovaciones recientes, Mamba ha demostrado ser especialmente eficaz, combinando tiempos de entrenamiento e inferencia reducidos, consumo moderado de memoria y precisión competitiva.

Referencias documentales

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). **Neural machine translation by jointly learning to align and translate**. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

Bishop, C., & Bishop, H. (2024). **Deep Learning: Foundations and Concepts**. Switzerland: Springer Cham. doi:https://doi.org/10.1007/978-3-031-45468-4

Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., & Kaiser, Ł. (2019). **Universal Transformers**. https://arxiv.org/abs/1807.03819

Eisenstein, J. (2019). I**ntroduction to Natural Language Processing**. MIT Press. https://books.google.co.cr/books?id=72yuDwAAQBAJ

Goodfellow, I., & Bengio, Y. (2016). **Deep Learning**. London, England: MIT Press.

Gu, A., & Dao, T. (2024). **Mamba: Linear-Time Sequence Modeling with Selective State Spaces.** https://arxiv.org/abs/2312.00752

Kamath, U., Graham, K., & Emara, W. (2022). **Transformers for Machine Learning: A deep dive**. CRC Press.

Koehn, P. (2020). **Neural Machine Translation**. Cambridge University Press. doi:https://doi.org/10.1017/9781108608480

Ma, X., Zhou, C., Kong, X., He, J., Gui, L., Neubig, G., May, J., & Zettlemoyer, L. (2022). **Mega: Moving Average Equipped Gated Attention.**

Mehta, H., Gupta, A., Cutkosky, A., & Neyshabur, B. (2022). **Long Range Language Modeling via Gated State Spaces.**

Prince, S. (2023). **Understanding Deep Learning**. The MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). **Attention is all you need**. Advances in Neural Information Processing Systems, 2017-December.

Yang, S., Wang, Y., & Chu, X. (2020). **A Survey of Deep Learning Techniques for Neural Machine Translation**. doi:https://doi.org/10.48550/arXiv.2002.07526

Zoph, B., Bello, I., Kumar, S., Du, N., Huang, Y., Dean, J., Shazeer, N., & Fedus, W. (2022). **ST-MoE: Designing Stable and Transferable Sparse Expert Models.**

¡Muchas gracias!