INSACERMO

MemGuard 0.4 Two-Door

Continuer, arrêter ou restaurer un entraînement IA

Dossier public de présentation et validation prospective V10a

3/3 apprentissages bénéfiques préservés	**3/3** sorties après entrée détectées	**3/3** trajectoires No-Entry détectées
6/6 restaurations exactes	**0 %** regret test médian	**+8,33 pts** gain pondéré vs early stopping

Version publique - recherche indépendante - Rennes, France

github.com/benblak/Memguard | insacermo.netlify.app

Document public. Les archives privées, checkpoints et logs bruts ne sont pas inclus.

1. Le problème : une mauvaise courbe ne dit pas toujours la même chose

Un early stopping standard attend généralement une succession de validations sans amélioration. Cette logique est utile, mais elle ne distingue pas toujours la géométrie du problème : un modèle peut apprendre correctement, quitter ensuite un régime bénéfique, ou ne jamais produire de bénéfice par rapport à son état initial.

Idée INSACERMO

Lire le régime qui porte les valeurs, pas seulement une valeur isolée. MemGuard transpose cette idée aux trajectoires formées par step, train loss et validation loss.

2. Les trois régimes observés

Apprentissage bénéfique

La validation s'améliore réellement. MemGuard s'abstient et laisse l'entraînement continuer.

Entrée bénéfique puis sortie

Le modèle a connu un état viable, puis la validation se dégrade durablement pendant que la géométrie train/validation se sépare. La porte Orange-Red recommande un arrêt et un rollback.

Aucune entrée bénéfique

Le fine-tuning ne dépasse jamais la référence initiale alors que la loss train continue à se spécialiser. La porte No-Entry recommande un arrêt et un retour au meilleur checkpoint.

3. Action validée

Le rollback fait partie de la méthode. Quand une porte confirme un arrêt, MemGuard recommande de restaurer le meilleur checkpoint déjà observé. Conserver les poids dégradés du moment de l'alarme ne correspond pas au protocole validé.

4. Comment utiliser MemGuard

MemGuard peut fonctionner en analyse rétrospective de logs, en supervision incrémentale ou comme callback Hugging Face. Le mode prudent recommandé sur un nouveau workload est le mode shadow : les décisions sont enregistrées sans arrêter automatiquement l'entraînement.

Entrées minimales

- step : numéro d'étape ou point d'évaluation ;
- train_loss : loss d'entraînement ;
- validation_loss ou eval_loss : loss de validation ;
- une première ligne représentant une véritable référence avant fine-tuning pour la sémantique exacte de No-Entry.

Sorties principales

- CONTINUE : aucun régime nuisible n'est confirmé ;
- STOP_AND_ROLLBACK + ORANGE_RED : sortie persistante après entrée bénéfique ;
- STOP_AND_ROLLBACK + NO_ENTRY : aucune entrée bénéfique confirmée ;
- recommended_rollback_step : checkpoint recommandé.

Exemple de commande

```
insacermo-memguard two-door training_log.csv --steps 10000 --json
```

Règles de déploiement

- Commencer en shadow mode sur tout nouveau modèle, corpus ou cadence d'évaluation.
- Ne pas modifier silencieusement les seuils gelés : toute modification doit être versionnée et revalidée prospectivement.
- Vérifier la qualité et la stabilité de la métrique de validation.
- Associer l'arrêt à une restauration effective du meilleur checkpoint.

Le paquet Python public, les exemples, les tests et la politique gelée sont disponibles sur GitHub. Le paquet PyPI nommé “memguard” appartient à un autre projet ; utiliser le dépôt INSACERMO.

5. Validation prospective V10 et audit V10a

Politique gelée avant exécution. V10 a utilisé neuf nouveaux runs, un ordre mélangé, deux branches modèle/corpus, des évaluations test ouvertes seulement après gel des décisions, et des vérifications SHA-256 des checkpoints restaurés. V10a a corrigé une seule étiquette de bord sans réentraînement, sans changement de seuil et sans changement de décision.

Protocole couvert

Élément	V10a
Entraînements prospectifs	9
Modèles	DistilGPT2, Pythia-70M
Corpus	WikiText-2, TinyStories
Régimes corrigés	3 bénéfiques, 3 sorties, 3 No-Entry
Comparateur	Early stopping standard : patience 5, min_delta 0,001
Cadence	Évaluation toutes les 20 étapes dans ce protocole

Chaîne expérimentale honnête

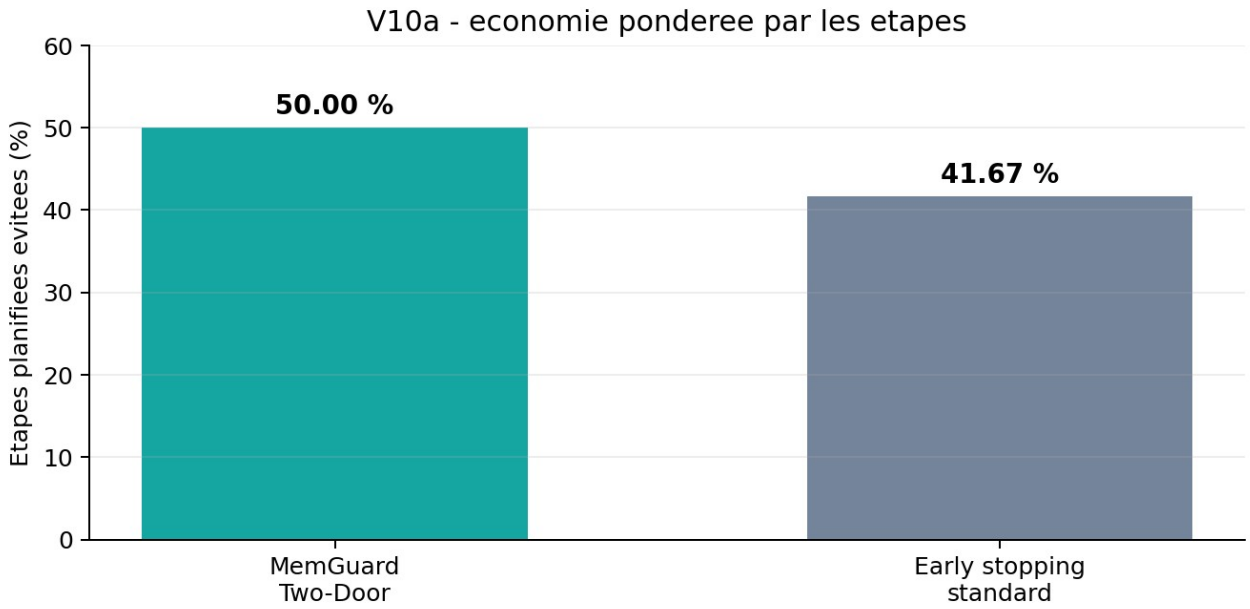
Les résultats négatifs font partie du dossier : V05 restait moins économe que l'early stopping ; V08 a révélé le régime No-Entry manquant ; V09 restait 0,69 point sous le comparateur sur sa partition de confirmation. Ces résultats ont conduit à la politique V09b, gelée avant V10.

Identités gelées

Politique V09b SHA-256 : 9b0b8ab801d2c6b52efdf38835960240dbf27d6a057f3a14dc0f9af3328c64af
Protocole V10 SHA-256 : 0ea7c2ac6252e4206ab2099033056b8dd2f977f1984d861ca3a0f6c4e1848fc0

6. Résultats V10a

3/3 contrôles bénéfiques préservés	**3/3** sorties détectées	**3/3** No-Entry détectés
0 actions avant le meilleur checkpoint	**6/6** rollbacks exacts par SHA-256	**0,0000 %** regret test médian



Comparaison du calcul planifié évité

Mesure	MemGuard	Early stopping
Économie macro	43,83 %	35,19 %
Économie pondérée par les étapes	50,00 %	41,67 %
Avantage MemGuard	+8,64 points	macro
Avantage MemGuard	+8,33 points	pondéré

Robustesse : en retirant entièrement le run T01 dont l'étiquette de bord a été auditée, le résultat reste positif : +6,94 points macro et +7,07 points pondérés.

Conclusion bornée

Sur ce protocole prospectif léger, le contrôleur à deux portes a préservé les entraînements bénéfiques, détecté tous les régimes nuisibles réalisés, restauré exactement les meilleurs checkpoints et économisé davantage de calcul planifié que le comparateur défini.

7. Ce que ce résultat permet - et ne permet pas

Ce que V10a soutient

- Une supervision causale interprétable de trois géométries d'entraînement.
- Une abstention correcte sur les trois contrôles bénéfiques planifiés.
- Une détection complète des sorties et des No-Entry réalisés dans ce protocole.
- Un rollback exact et un regret test médian nul sur les actions effectuées.
- Un gain de calcul face au comparateur early stopping défini.

Ce que V10a ne prouve pas

- Une supériorité universelle sur tous les early stopping.
- Une sécurité de production sur toutes les tailles de modèles.
- Une invariance aux optimiseurs, tâches, métriques ou cadences d'évaluation.
- Une puissance statistique générale à partir de neuf runs.

Règle de prudence

Sur tout nouveau workload : shadow mode d'abord, observation des décisions, audit des faux positifs et faux négatifs, puis activation éventuelle du contrôle automatique.

8. Deux audits externes gratuits

INSACERMO recherche une ou deux équipes acceptant une analyse confidentielle de logs indépendants. Aucun accès aux poids, au modèle ou à l'infrastructure n'est nécessaire.

- Données : step, train_loss, validation_loss ou eval_loss, et éventuellement identifiant de checkpoint.
- Livrable : régime détecté, point d'arrêt recommandé, checkpoint à conserver, comparaison avec l'early stopping et estimation du calcul évité.
- Publication : aucune étude de cas sans validation écrite ; anonymisation possible.

Benjamin Lenoir - INSACERMO - Rennes

benjamin.professionnel@gmail.com | GitHub MemGuard | Site INSACERMO