Rapport de validation des modèles de la vitrine

Thèmes (CAP) et partis politiques : données, entraînement, performance, calibration

30 août 2026

Résumé

Ce rapport documente les 30 modèles de classification servis à la vitrine (21 têtes thématiques et 9 têtes de partis politiques) ainsi que le modèle de sentiment. Pour chaque modèle il donne le jeu d'entraînement exact (volume, équilibre des classes, équilibre linguistique, empreinte), la procédure suivie, la performance mesurée en validation interne puis sur 2 273 phrases annotées à la main sans pré-annotation, et la calibration du seuil de décision. Les 30 modèles sont des encodeurs mDeBERTa **distillés à partir de `gpt-5.4` (OpenAI)**, qui a produit les annotations du corpus d'entraînement.

Ce qu'il faut retenir. La performance des modèles tels qu'ils sont servis, seuils calibrés compris, est la suivante :

— **Têtes thématiques** : F1 de classe positive de **0,656** en moyenne, avec une dispersion très large (`housing`, `environment` et `health` au-dessus de 0,84 ; `domestic_commerce` et `public_lands` sous 0,45). 13 têtes sur 21 dépassent 0,65.

— **Têtes de partis** : 4 partis fédéraux (BQ, CPC, LPC, NDP) disposent d'un effectif de référence suffisant et obtiennent de 0,934 à 0,994. Les cinq autres (CAQ, GPC, PLQ, PQ, QS) comptent de 1 à 8 cas de référence et ne sont *pas* validés.

— **Sentiment** : macro-F1 de 0,653, à comparer à un accord entre annotateurs de 0,669 — soit 98 % du plafond humain.

— **Plafond humain** : l'accord entre annotateurs vaut 0,726 en moyenne sur les thèmes ; les modèles en atteignent 0,93 fois. Une part importante de l'écart à la perfection est une ambiguïté du codebook (corrélation accord/performance : 0,644).

— **À ne pas confondre** : la validation interne (0,913 de macro-F1 pour les thèmes, 0,954 pour les partis) mesure l'apprentissage sur la distribution d'entraînement.

Où lire le F1 des modèles servis. Une seule colonne fait foi par famille de modèles ; elle est identifiée en gras dans les tableaux et récapitulée en section 1.3. Les tableaux sont produits automatiquement à partir des journaux d'entraînement versionnés dans ce même dossier ; aucun chiffre n'est saisi à la main.

Table des matières

1 Périmètre et lecture du rapport

1.1 Ce qui est documenté

La vitrine s'appuie sur 30 classifieurs de phrases, tous servis par l'API d'inférence `gate.llm-tool.org` :

— **21 têtes thématiques**, une par catégorie majeure du *Comparative Agendas Project*, identifiées `cap_theme_<thème>` ;

— **9 têtes de partis politiques**, identifiées `party_<code>` : BQ, CAQ, CPC, GPC, LPC, NDP, PLQ, PQ, QS.

Chaque tête est un classifieur **binaire indépendant** (« ce thème est-il présent dans cette phrase, oui ou non ? »), et non une sortie parmi d'autres d'un modèle multi-étiquette unique. Tous sont des encodeurs mDeBERTa entraînés par distillation d'annotations produites par `gpt-5.4` (OpenAI), comme l'expose la section 2. Ce choix d'architecture, discuté en section 2.2, est la décision structurante dont découlent le protocole d'entraînement, la calibration et le mode d'appel.

Pour chaque modèle, le rapport donne : le jeu d'entraînement exact et sa composition, la procédure suivie et ses paramètres, la performance en validation interne, la performance sur des phrases annotées à la main, enfin le seuil de décision et sa calibration.

1.2 Deux niveaux de performance

Le rapport rapporte deux familles de chiffres qui ne mesurent pas la même chose et qui ne sont pas comparables entre elles.

La validation interne (section 4) est celle que produit l'entraînement : une part de 20 % du corpus d'entraînement, tirée de la même distribution que les exemples appris, jamais vue pendant l'optimisation. Elle mesure si le modèle a appris ce que ses données lui montraient. Les valeurs y sont élevées (macro-F1 moyen de 0,913 pour les thèmes et 0,954 pour les partis) et servent au suivi de l'entraînement, pas à prédire le comportement en production.

La validation externe (section 5) confronte les modèles à 2 273 phrases annotées par des humains, sur des textes de presse et parelementaires qui ne proviennent pas du corpus d'entraînement. C'est le seul chiffre qui dit ce que vaut le modèle sur les documents que traite réellement la vitrine. Il est nettement plus bas, ce qui est attendu : le changement de domaine et le désaccord entre annotateurs y pèsent tous les deux.

L'écart entre les deux chiffre le coût du transfert hors domaine, tête par tête.

1.3 Où lire le F1 des modèles servis

Le rapport contient plusieurs colonnes de F1, qui ne mesurent pas la même chose. Une seule fait foi pour chaque famille de modèles : celle qui donne la performance du modèle *tel qu'il est servi*, au seuil de décision effectivement appliqué en production. Elle est mise **en gras** dans les tableaux, sous l'intitulé « **F1 final** ». Le tableau ci-dessous dit, pour chaque famille, où la trouver.

Les colonnes à ne pas citer comme performance de production.

— Les **macro-F1 de la section 4** (tables 5 et 6) sont mesurés sur la part de validation du corpus d'entraînement : ils décrivent l'apprentissage, pas la production.

— La colonne « **F1 servi (in-éch.)** » des tables 8 et 12 donne la performance au seuil servi, mais mesurée sur les phrases qui ont servi à choisir ce seuil : elle est optimiste de 0,013 en moyenne pour les thèmes.

— La colonne « **F1 à 0,5** » de la table 12 n'est qu'un point de comparaison servant à mesurer l'apport de la calibration — sauf pour les partis, où 0,5 *est* le seuil servi.

— La dernière colonne de la table 10, « F1 h.-é. (seuil non déployé) », correspond à un seuil calibré qui a été *écarté* au déploiement parce qu'il ne faisait pas mieux que 0,5.

Table 1 – La colonne à lire pour connaître la performance réelle de chaque famille de modèles servis.

Famille	Tableau et colonne	Pourquoi celle-là
21 têtes thématiques	tab. 8, colonne **« F1 final »** (identique à « F1 hors-éch. » de tab. 12)	Ces têtes sont servies avec un *seuil calibré*. Le seuil ayant été réglé sur les mêmes phrases annotées, seule l'estimation hors échantillon (seuil appris sur une moitié, mesuré sur l'autre, 200 fois) est honnête. Moyenne : **0,656**.
9 têtes de partis	tab. 10, colonne **« F1 final »** (= « F1 servi » et « F1 à 0,5 » de tab. 12)	Ces têtes sont servies *au seuil de 0,5*, aucun seuil calibré n'ayant été déployé : la colonne « F1 servi » est donc bien le chiffre final, sans biais de calibration. Seuls BQ, CPC, LPC et NDP ont assez de cas de référence pour que le chiffre ait un sens.
Modèle de sentiment	tab. 11, ligne **« Macro-F1 du modèle »**	Modèle à trois classes, évalué à travers l'API elle-même, sans seuil à calibrer : **0,653**.

— Les macro-F1 mélangent classe positive et classe négative ; sur des données déséquilibrées la classe négative est triviale et tire la moyenne vers le haut. Toutes les colonnes « F1 final » portent sur la **seule classe positive**, qui est le chiffre informatif.

1.4 Reproductibilité

Le dossier est autoportant. Les journaux d'entraînement et les jeux de données de toutes les sessions concernées y sont copiés (`logs/`, avec `MANIFEST.json` qui donne la taille et l'empreinte SHA-256 de chaque fichier). Les deux jeux d'annotation humaine sont dans `data/gold/`. Les scripts de `scripts/` reconstruisent l'intégralité des tableaux à partir de ces sources :

1. `01_extract_training.py` consolide les journaux d'entraînement et vérifie leur cohérence avec les poids déployés ;
2. `02_evaluate_fleet.py` recalcule les métriques depuis les scores archivés et rejoue un modèle si son cache est incompatible ;
3. `03_collect_logs.py` copie les journaux et écrit le manifeste ;
4. `04_make_tables.py` produit les tableaux LaTeX.

Aucun nombre de ce document n'a été saisi à la main : chaque tableau, et même les valeurs citées dans le corps du texte, proviennent de `data/extract/`. La section A.1 rend compte des 124 contrôles automatiques qui vérifient cette chaîne.

2 Comment les modèles ont été produits

2.1 Le corpus d'entraînement et son annotation

Tous les modèles de la vitrine descendent d'un même corpus de **99 997 phrases** bilingues, extrait du fichier maître `data/cap-model/ALL_ANNOTATED.csv` (287 774 phrases annotées au total). Sa composition est stable d'une session à l'autre : 46 226 phrases en anglais, 48 319 en français, et 5 452 dont la langue n'est pas renseignée dans la source. Les phrases sont courtes (143 caractères et 36 sous-mots en moyenne, médiane à 123 caractères), avec une queue longue qui va jusqu'à 3 757 caractères, ce qui justifie la troncature à 512 sous-mots retenue à l'entraînement comme à l'inférence.

Les étiquettes de ce corpus **ne sont pas des annotations humaines** : elles ont été produites par le module d'annotation par grand modèle de langue de LLM_Tool, puis distillées dans des classifieurs légers. Le modèle enseignant est **`gpt-5.4` d'OpenAI**, interrogé via l'API Batch au cours de quatre sessions d'annotation menées du 6 au 9 mai 2026 (sessions `cap-model`,

`cap-model-2`, `cap-model-3` et la reprise du 9 mai). Les lignes de journal qui l'attestent sont recopiées dans `logs/annotation_teacher.txt`. Les 30 modèles servis sont donc, tous sans exception — têtes thématiques comme têtes de partis, qui dérivent du même fichier maître —, des distillations de `gpt-5.4` dans des encodeurs mDeBERTa. C'est le principe même de la chaîne : un modèle génératif annote massivement, un encodeur spécialisé apprend de ces annotations et sert ensuite à l'échelle, pour un coût d'inférence sans commune mesure. Cette généalogie a une conséquence directe sur la lecture des résultats : la validation interne mesure la fidélité à l'annotateur automatique, tandis que la validation externe de la section 5, elle, mesure l'accord avec des humains.

Trois jeux dérivés alimentent les sessions, tous de 99 997 lignes : `themes_good_multilabel.csv` pour les thèmes des vagues de juin et juillet, `parties_good_multilabel.csv` pour les partis, et `themes_weak_binary.csv` pour le ré-entraînement d'août. Ce ne sont pas eux qui sont archivés dans `logs/`, mais ce qui en a été tiré : le jeu binaire effectivement présenté à chaque tête, un fichier par tête, avec son empreinte.

2.2 Une tête binaire par catégorie

Chaque catégorie est traitée par un classifieur binaire indépendant (approche « un contre tous ») plutôt que par une sortie d'un modèle multi-étiquette unique. Ce choix a un coût réel : 30 modèles à entraîner, à stocker et à servir, là où un seul modèle multi-étiquette suffirait, et une inférence plus lente puisqu'une phrase doit traverser autant d'encodeurs qu'il y a de catégories interrogées. Il apporte en échange trois propriétés décisives pour la vitrine :

— **un seuil de décision par catégorie**, calibrable indépendamment (section 6). Un modèle multi-étiquette impose en pratique un compromis global, alors que les catégories n'ont ni la même fréquence ni la même difficulté ;

— **une réparation ciblée, donc meilleure maintenabilité** : une catégorie décevante se ré-entraîne seule, sans toucher aux autres. Les 21 têtes thématiques actuelles proviennent d'ailleurs de trois sessions différentes, précisément parce que les corrections ont été successives ;

— **un service parallèle** : les têtes étant indépendantes, les requêtes se répartissent sur autant de verrous distincts côté serveur.

2.3 Le protocole d'entraînement

Toutes les têtes de thèmes et de partis reposent sur **mDeBERTa-v3-base**, encodeur multilingue qui traite l'anglais et le français sans modèle séparé par langue. Trois éléments sont communs aux quatre sessions : des lots de 16 exemples, un découpage stratifié **80 / 20** avec graine 42 et sans part de test (la sélection se fait sur la part de validation de 20 %), et la sélection du meilleur checkpoint par métrique combinée avec arrêt anticipé.

Le reste varie, et le tableau 2 donne le détail. Le **taux d'apprentissage** vaut 2×10^{-5} pour les huit têtes de la vague de juin et 1×10^{-5} pour les vingt-deux autres, c'est-à-dire `immigration` et `labor` en juillet, les onze têtes ré-entraînées en août et les neuf têtes de partis. Le **préchauffage** suit la même partition : nul en juin, fixé à 10 % du parcours ensuite. Ce réglage n'est pas anodin, puisque c'est précisément l'absence de préchauffage qui avait rendu instable l'entraînement de `labor` en juin. Le plafond d'époques et la patience de l'arrêt anticipé valent enfin 12 et 3 pour les vagues de juin et d'août, 15 et 5 pour juillet et pour les partis.

Sur la session des partis, la patience de 5 s'est révélée coûteuse : plusieurs têtes ont continué à tourner des heures après leur meilleur checkpoint sans progresser. Deux d'entre elles (NDP et PLQ) ont donc été arrêtées manuellement en cours de route, à la frontière d'une époque, au moyen du mécanisme d'interruption interactive de LLM_Tool, qui conserve le meilleur checkpoint déjà enregistré. Les journaux ne distinguent pas cette interruption d'un arrêt automatique : le fait est consigné ici parce qu'il est connu de l'opérateur, non parce qu'il serait déductible des fichiers.

2.4 Deux régimes de données

Les sessions d'entraînement ne présentent pas les mêmes distribution de données dû notamment à la rareté de certaines catégories. Deux régimes ont été utilisés.

Le régime à prévalence naturelle (vagues de juin, juillet, et partis) entraîne chaque tête sur les 99 997 phrases, avec la fréquence réelle de la catégorie. Pour les thèmes fréquents c'est confortable (`international_affairs` compte 16 327 positifs), pour les catégories rares beaucoup moins : `party_QS` apprend à partir de 148 exemples positifs noyés dans 99 849 négatifs, soit 0,1 % de signal.

Le régime équilibré (ré-entraînement d'août) construit au contraire, pour chaque tête, un jeu ne contenant que les phrases où la catégorie a été explicitement tranchée, en retenant tous les positifs et cinq négatifs par positif. Le résultat est un jeu plus petit mais dense : 8 646 lignes à 16,7 % de positifs pour `agriculture`, contre 99 997 lignes à 1,4 % dans l'ancien régime. Seule `governments_governance` échappe à la règle, ses 24 781 positifs interdisant d'atteindre le rapport de un pour cinq à l'intérieur du corpus disponible.

Ce changement de régime est la principale explication du gain hors domaine mesuré en section 5 sur les onze têtes concernées. La validation interne, elle, ne le voit presque pas : les deux régimes donnent des macro-F1 comparables, chacun étant évalué sur une part de validation issue de sa propre distribution. Autrement dit, le progrès réel de ces onze têtes était invisible à la métrique d'entraînement et n'apparaît que face à des annotations humaines.

2.5 Historique des incidents

L'état actuel de la flotte est le produit de quatre sessions successives, dont trois sont des corrections. Le rapport le documente.

1. **Mai 2026 : lots de 256.** La première vague de têtes thématiques a été entraînée avec des lots de 256 exemples, contre 16 ensuite. Les têtes de classification obtenues étaient dégradées et ont dû être refaites le 26 mai par un ré-entraînement de la seule couche de classification, sur des représentations gelées. Ce sauvetage a restauré des métriques internes correctes, mais il ne pouvait pas réparer l'encodeur : c'est précisément ce déficit que la validation externe a fini par révéler, et que le ré-entraînement complet d'août a corrigé.
2. **Juin 2026 : deux têtes perdues.** Les têtes `labor` et `immigration` sont sorties inutilisables de la vague de juin, `labor` par instabilité d'optimisation et `immigration` par export corrompu.
3. **13 juillet 2026 : reprise et garde-fou.** Les deux têtes ont été refaites, et un garde-fou d'export a été ajouté à LLM_Tool : après chaque sauvegarde, le checkpoint est rechargé puis comparé au modèle en mémoire, et re-sauvegardé en cas de divergence. Le champ `save_verification` des métadonnées porte le résultat de ce contrôle. Les huit têtes de juin, antérieures au garde-fou, ne possèdent pas ce champ, ce que la section A.1 signale explicitement plutôt que de le passer sous silence.
4. **19 au 21 août 2026 : ré-entraînement des onze têtes faibles**, en régime équilibré, puis déploiement le 30 août après validation sur annotations humaines.

Table 2 – Les quatre sessions d'entraînement dont proviennent les modèles servis.

Session	Têtes	Période	Taux appr.	Warmup	Ép. / pat.	Jeu source
août (ré-entraînement)	11	2026-08-19 au 2026-08-21	1e-5	0,1	12 / 3	themes_weak_binary
juin (thèmes)	8	2026-06-26 au 2026-07-11	2e-5	0,0	12 / 3	themes_good_multilabel
juillet (reprise)	2	2026-07-13 au 2026-07-15	1e-5	0,1	15 / 5	themes_good_multilabel
août (partis)	9	2026-08-08 au 2026-08-19	1e-5	0,1	15 / 5	parties_good_multilabel

3 Les données d'entraînement, tête par tête

Les deux tableaux qui suivent donnent, pour chaque modèle servi, le jeu exact sur lequel il a appris. Ces jeux sont archivés dans le dossier (`logs/<session>/jeux/`, compressés, avec l'empreinte SHA-256 de l'original et de l'archive dans `logs/MANIFEST.json`) : les colonnes ci-dessous ont été recalculées ligne à ligne à partir de ces fichiers, et non recopiées d'un rapport de session.

La colonne « Lignes » est le nombre d'exemples présentés à la tête, « Positifs » le nombre d'exemples de la classe visée, et « Taux » leur proportion. La part de validation qui sert à mesurer la section 4 en représente 20 %, tirée de façon stratifiée avec la graine 42.

Table 3 – Jeux d'entraînement des 21 têtes thématiques. Les onze premières lignes relèvent du régime équilibré d'août (un positif pour cinq négatifs), les suivantes du régime à prévalence naturelle.

Tête	Session	Lignes	Positifs	Taux	EN / FR	Positifs EN / FR
agriculture	août 2026	8 646	1 441	16,7 %	4 258 / 3 964	901 / 523
culture_nationalism	août 2026	18 966	3 161	16,7 %	8 451 / 9 488	1 274 / 1 814
defense	août 2026	24 474	4 079	16,7 %	11 596 / 11 672	2 211 / 1 827
domestic_commerce	août 2026	27 786	4 631	16,7 %	13 316 / 13 052	2 706 / 1 845
foreign_trade	août 2026	10 878	1 813	16,7 %	5 321 / 5 009	1 143 / 637
governments_gov.	août 2026	99 997	24 781	24,8 %	46 226 / 48 319	12 532 / 11 570
public_lands	août 2026	7 992	1 332	16,7 %	3 678 / 3 962	624 / 691
rights_liberties_min.	août 2026	35 574	5 929	16,7 %	16 525 / 17 286	3 001 / 2 841
social_welfare	août 2026	25 380	4 230	16,7 %	11 697 / 12 449	2 004 / 2 184
technology	août 2026	14 892	2 482	16,7 %	6 858 / 7 307	1 137 / 1 300
transportation	août 2026	17 694	2 949	16,7 %	8 286 / 8 487	1 481 / 1 411
education	juin 2026	99 997	3 711	3,7 %	46 226 / 48 319	1 598 / 2 080
energy	juin 2026	99 997	2 582	2,6 %	46 226 / 48 319	1 416 / 1 115
environment	juin 2026	99 997	3 756	3,8 %	46 226 / 48 319	1 942 / 1 749
health	juin 2026	99 997	6 649	6,6 %	46 226 / 48 319	2 878 / 3 698
housing	juin 2026	99 997	3 214	3,2 %	46 226 / 48 319	1 690 / 1 473
international_affairs	juin 2026	99 997	16 327	16,3 %	46 226 / 48 319	8 213 / 7 740
law_and_crime	juin 2026	99 997	9 487	9,5 %	46 226 / 48 319	4 851 / 4 490
macroeconomics	juin 2026	99 997	11 037	11,0 %	46 226 / 48 319	5 573 / 5 201
immigration	juillet 2026	99 997	2 685	2,7 %	46 226 / 48 319	1 002 / 1 655
labor	juillet 2026	99 997	5 352	5,4 %	46 226 / 48 319	2 682 / 2 611

Deux lectures s'imposent. D'abord, **le déséquilibre des classes est la difficulté centrale** de cette famille de modèles : en régime naturel, une tête comme `energy` apprend sur 2,6 % de positifs, et le modèle qui répondrait « non » systématiquement obtiendrait déjà 97,4 % d'exactitude. C'est pourquoi ce rapport ne cite jamais l'exactitude seule et s'appuie sur le F1 de la classe positive.

Ensuite, **les jeux sont linguistiquement équilibrés, mais pas leurs positifs**. Le corpus est à peu près moitié anglais, moitié français dans toutes les sessions. Pour les thèmes, les positifs le sont aussi, et aucune tête thématique n'est structurellement monolingue. Pour les partis, en revanche, la répartition des positifs suit la réalité politique et non le corpus : les partis québécois sont massivement francophones (99 % des positifs de `QS`, 81 % de ceux du `PQ` et de la `CAQ`, 74 % de ceux du `PLQ`), tandis que `GPC` est aux quatre cinquièmes anglophone. Ce déséquilibre n'est pas un défaut d'échantillonnage et il se lit directement dans les F1 par langue de la section 4 : la tête `QS`, qui ne dispose que d'un seul positif anglais, y affiche un F1 anglais nul qui traduit une absence de données et non un échec du modèle.

Les partis illustrent le cas limite. Les trois partis fédéraux dominants disposent de milliers d'exemples, tandis que QS, PQ et PLQ en comptent moins de 250 chacun, soit deux dixièmes de pour cent du corpus. Que des modèles entraînés dans ces conditions atteignent malgré tout les scores de la section 5 tient à la nature de la tâche : reconnaître la mention d'un parti est largement

Table 4 – Jeux d'entraînement des 9 têtes de partis. Toutes relèvent du régime à prévalence naturelle : le déséquilibre y est extrême.

Tête	Session	Lignes	Positifs	Taux	EN / FR	Positifs EN / FR
BQ	août 2026	99 997	779	0,8 %	46 226 / 48 319	333 / 430
CAQ	août 2026	99 997	451	0,5 %	46 226 / 48 319	77 / 366
CPC	août 2026	99 997	3 480	3,5 %	46 226 / 48 319	2 098 / 1 335
GPC	août 2026	99 997	232	0,2 %	46 226 / 48 319	180 / 42
LPC	août 2026	99 997	4 588	4,6 %	46 226 / 48 319	2 870 / 1 612
NDP	août 2026	99 997	1 671	1,7 %	46 226 / 48 319	1 168 / 449
PLQ	août 2026	99 997	221	0,2 %	46 226 / 48 319	50 / 164
PQ	août 2026	99 997	210	0,2 %	46 226 / 48 319	36 / 171
QS	août 2026	99 997	148	0,1 %	46 226 / 48 319	1 / 147

lexical, là où reconnaître un thème demande une inférence sémantique.

Deux partis sans modèle. Le codebook comporte onze partis, la vitrine n'en sert que neuf. Le **PPC** n'a aucun exemple positif dans le corpus annoté : aucun modèle n'a pu être entraîné. Le **PCQ** en compte dix-huit, dont un examen manuel a montré qu'une majorité désigne en réalité les conservateurs de Colombie-Britannique ; ces données ont été jugées inutilisables et la tête écartée. Une phrase mentionnant l'un de ces deux partis ne recevra donc aucune étiquette de parti, ce qui est un silence assumé et non une erreur du système.

4 Validation interne : ce que l'entraînement a mesuré

Les chiffres de cette section proviennent de la part de validation de 20 %, tirée du même corpus que l'entraînement et jamais vue pendant l'optimisation. Ils répondent à la question « le modèle a-t-il appris ce que ses données lui montraient ? », et à elle seule. Ils ne disent rien de la performance sur des textes de presse réels, que la section 5 mesure séparément et qui est nettement plus basse.

Avertissement de lecture. Aucun chiffre de cette section n'est le F1 des modèles servis. Les macro-F1 ci-dessous sont mesurés sur la distribution d'entraînement et ne doivent jamais être cités comme performance de production. Le F1 des modèles servis se lit dans la colonne « **F1 final** » des tables 8 (thèmes) et 10 (partis) — voir la section 1.3.

Chaque ligne correspond au checkpoint effectivement déployé, c'est-à-dire à l'époque retenue par la sélection sur métrique combinée. La colonne « Époque » indique cette époque et le nombre total d'époques exécutées : un écart important signale un arrêt anticipé, ou une interruption manuelle dans le cas des partis NDP et PLQ.

Comment lire le macro-F1. Il moyenne le F1 de la classe positive et celui de la classe négative. Sur des données très déséquilibrées, la classe négative est facile et tire la moyenne vers le haut : c'est pourquoi les tableaux donnent aussi la précision, le rappel et le F1 de la seule classe positive, qui sont les chiffres informatifs. L'écart entre les deux colonnes est d'autant plus grand que la catégorie est rare.

Le cas des partis. Leurs macro-F1 (moyenne 0,954) dépassent ceux des thèmes, alors même que leurs données sont bien plus déséquilibrées. La tâche est en effet plus lexicale : le nom d'un parti, ses sigles et ceux de ses chefs sont des indices explicites, tandis qu'un thème doit être inféré. Le F1 de classe positive reste élevé y compris pour les partis à faible effectif, ce que la validation externe confirmera pour ceux d'entre eux disposant d'assez de cas de référence.

Table 5 – Validation interne des 21 têtes thématiques, au checkpoint déployé. Macro-F1 moyen : 0,913.

Tête	Époque	Exact.	Macro-F1	Préc.$_1$	Rapp.$_1$	F1$_1$	F1$_1$ EN	F1$_1$ FR	Support$_1$
agriculture	9/9	0,977	0,959	0,931	0,934	0,932	0,938	0,939	288
culture_nationalism	3/6	0,950	0,910	0,856	0,845	0,850	0,867	0,853	632
defense	1/1	0,957	0,921	0,883	0,853	0,868	0,867	0,870	815
domestic_commerce	4/7	0,938	0,889	0,808	0,822	0,815	0,824	0,812	926
foreign_trade	4/7	0,976	0,956	0,940	0,912	0,926	0,926	0,928	363
governments_gov.	3/6	0,872	0,833	0,726	0,780	0,752	0,764	0,744	4 956
public_lands	12/12	0,962	0,934	0,871	0,910	0,890	0,886	0,906	266
rights_liberties_min.	4/7	0,936	0,883	0,818	0,791	0,804	0,823	0,790	1 185
social_welfare	5/8	0,942	0,898	0,818	0,842	0,830	0,824	0,837	846
technology	6/9	0,955	0,922	0,837	0,909	0,871	0,861	0,883	496
transportation	11/12	0,966	0,939	0,876	0,924	0,899	0,898	0,902	590
education	6/9	0,988	0,917	0,810	0,872	0,840	0,817	0,858	742
energy	12/12	0,991	0,915	0,811	0,859	0,834	0,828	0,833	516
environment	12/12	0,986	0,896	0,836	0,765	0,799	0,823	0,775	750
health	6/9	0,982	0,925	0,879	0,843	0,861	0,864	0,864	1 329
housing	10/12	0,990	0,919	0,871	0,818	0,844	0,849	0,839	643
international_affairs	2/5	0,966	0,939	0,881	0,916	0,898	0,901	0,902	3 264
law_and_crime	2/5	0,964	0,890	0,835	0,768	0,800	0,823	0,780	1 897
macroeconomics	3/4	0,959	0,895	0,814	0,813	0,813	0,821	0,808	2 207
immigration	10/15	0,992	0,919	0,843	0,842	0,842	0,829	0,856	537
labor	7/8	0,981	0,910	0,813	0,848	0,830	0,857	0,803	1 070

Table 6 – Validation interne des 9 têtes de partis, au checkpoint déployé. Macro-F1 moyen : 0,954.

Tête	Époque	Exact.	Macro-F1	Préc.$_1$	Rapp.$_1$	F1$_1$	F1$_1$ EN	F1$_1$ FR	Support$_1$
BQ	8/13	1,000	0,992	0,987	0,981	0,984	0,983	0,984	156
CAQ	3/8	1,000	0,971	0,988	0,900	0,942	0,970	0,933	90
CPC	12/15	0,998	0,984	0,981	0,958	0,969	0,973	0,962	696
GPC	14/15	0,999	0,931	0,837	0,891	0,863	0,880	0,778	46
LPC	5/10	0,995	0,969	0,936	0,946	0,941	0,951	0,924	918
NDP	4/6	0,999	0,987	0,988	0,961	0,974	0,989	0,939	334
PLQ	6/10	0,999	0,815	0,793	0,523	0,630	0,632	0,615	44
PQ	8/8	1,000	0,969	0,974	0,905	0,938	1,000	0,930	42
QS	6/10	1,000	0,968	0,906	0,967	0,935	0,000	0,935	30

5 Validation externe : confrontation à des annotations humaines

5.1 Les jeux de référence

Deux jeux annotés à la main, totalisant 2 273 phrases, servent de référence indépendante. Aucun ne provient du corpus d'entraînement, et tous deux portent sur le type de texte que traite la vitrine.

Table 7 – Les deux jeux d'annotation humaine retenus comme référence.

Jeu	Phrases	Annotateurs	Thèmes	Partis
benchmark (4 annot.)	1 549	shdin, Jeremy, jdrouin, BenjaminCarignan	21	les 9 + PCQ
doccano 57	724	junior, m-a-martel	21	aucun

Le premier est le benchmark historique de LLM_Tool : 1 549 phrases annotées indépendamment par quatre personnes, avec une colonne de consensus qui sert de vérité.

Le second est le projet doccano 57 : 724 phrases annotées à l'aveugle par deux personnes sur les 21 thèmes. **Chaque phrase y a été explicitement confirmée par les deux annotateurs**, si bien que l'absence d'étiquette vaut négatif déclaré et non simple omission. Pour chaque thème, la vérité retenue est leur consensus, c'est-à-dire leur accord.

Les métriques poolées reposent ainsi sur 2 273 phrases et 1 792 occurrences positives de référence au total.

Ce qui n'est pas validé. Les 9 têtes de partis ne disposent que du benchmark à quatre annotateurs. Le projet doccano dédié (« Vitrine - Partis & Sentiments », 885 phrases chargées, les dix étiquettes de partis définies) existe mais sa campagne n'a pas démarré : 48 phrases seulement y sont confirmées, par un seul annotateur, pour trois étiquettes de partis au total. Ce jeu est donc écarté, et les partis à faible effectif restent, à ce stade, non validés de façon robuste.

5.2 Protocole de mesure

Chaque modèle a été rejoué sur l'intégralité des phrases de référence, et non sur un échantillon. Trois points de fonctionnement sont rapportés : le seuil naïf de 0,5, le **seuil effectivement servi par l'API**, et un seuil calibré sur la référence poolée.

L'estimation honnête est celle dite « hors échantillon » : le seuil est choisi sur une moitié tirée au hasard des phrases, stratifiée pour préserver la proportion de positifs, et évalué sur l'autre moitié ; l'opération est répétée 200 fois et l'on rapporte la moyenne des F1 obtenus sur les moitiés non vues. Cette précaution évite l'illusion d'optique du seuil choisi et mesuré sur les mêmes données. Les têtes comptant moins de douze positifs ne sont pas calibrées du tout : leur seuil reste à 0,5 et leur estimation hors échantillon est laissée vide plutôt que d'afficher un chiffre que l'effectif ne soutient pas.

5.3 Résultats des têtes thématiques

Quelle colonne lire. Pour les 21 têtes thématiques, le F1 du modèle servi — c'est-à-dire au seuil calibré effectivement appliqué par les clients de l'API — est la colonne « **F1 final** », en gras et en dernière position du tableau 8. C'est l'estimation hors échantillon décrite ci-dessus, et c'est la seule à citer. Les colonnes « Bench. » et « Doc. 57 » donnent le même F1 jeu par jeu, et « F1 servi (in-éch.) », « Préc. », « Rapp. » les métriques sur la référence poolée au seuil servi, mesurées sur les phrases mêmes qui ont servi à régler ce seuil : elles sont optimistes.

Les mêmes modèles qui atteignent 0,913 de macro-F1 sur leur propre distribution tombent à 0,656 de F1 sur la classe positive face à des humains. La dispersion entre têtes est considérable. En haut du tableau, `environment`, `housing` et `health` dépassent 0,84 et sont utilisables sans

Table 8 – Performance des têtes thématiques sur les phrases annotées à la main. « Sup. » est le nombre de positifs de référence cumulés. Les deux colonnes par jeu donnent le F1 au seuil servi ; les suivantes, les métriques sur la référence poolée. Moyennes : 0,669 au seuil actuellement servi, 0,656 en estimation hors échantillon.

Tête	Sup.	Bench.	Doc. 57	F1 servi (in-éch.)	Préc.	Rapp.	**F1 final**
agriculture	20	0,491	0,625	0,522	0,367	0,900	0,508
culture_nationalism	30	0,524	0,500	0,517	0,536	0,500	0,460
defense	42	0,609	0,529	0,583	0,492	0,714	0,540
domestic_commerce	40	0,344	0,424	0,371	0,316	0,450	0,339
foreign_trade	56	0,817	0,778	0,804	0,843	0,768	0,758
governments_gov.	244	0,555	0,301	0,523	0,457	0,611	0,508
public_lands	26	0,508	0,250	0,437	0,311	0,731	0,387
rights_liberties_min.	75	0,685	0,737	0,702	0,634	0,787	0,675
social_welfare	72	0,727	0,583	0,703	0,699	0,708	0,669
technology	34	0,603	0,533	0,581	0,458	0,794	0,571
transportation	49	0,733	0,462	0,651	0,525	0,857	0,647
education	38	0,571	0,718	0,632	0,526	0,789	0,684
energy	58	0,788	0,632	0,745	0,646	0,879	0,741
environment	75	0,903	0,706	0,861	0,819	0,907	0,855
health	140	0,834	0,831	0,833	0,846	0,821	0,845
housing	77	0,857	0,870	0,859	0,889	0,831	0,856
international_affairs	241	0,822	0,731	0,799	0,730	0,884	0,793
law_and_crime	180	0,775	0,703	0,757	0,743	0,772	0,762
macroeconomics	191	0,798	0,558	0,749	0,678	0,838	0,730
immigration	31	0,711	0,690	0,703	0,605	0,839	0,677
labor	73	0,763	0,605	0,724	0,624	0,863	0,780

réserve. En bas, `domestic_commerce` et `public_lands` restent sous 0,45 : ces deux catégories sont à considérer comme indicatives et non comme des filtres fiables.

5.4 Le plafond humain

Comparer un modèle à une performance parfaite n'a pas de sens quand les humains eux-mêmes ne s'accordent pas. Le tableau 9 donne, par thème et par jeu, l'accord entre annotateurs mesuré exactement comme la performance des modèles.

Table 9 – Accord entre annotateurs, par thème et par jeu de référence. Cet accord est le plafond réaliste : aucun modèle ne peut durablement le dépasser sur une catégorie que les humains eux-mêmes ne tranchent pas de la même façon.

	Benchmark		Doccano 57	
Thème	Sup.	Accord	Sup.	Accord
agriculture	14	0,575	6	0,800
culture_nationalism	23	0,349	7	0,368
defense	31	0,508	11	0,786
domestic_commerce	26	0,275	14	0,667
foreign_trade	38	0,734	18	0,878
governments_gov.	226	0,470	18	0,514
public_lands	20	0,378	6	0,444
rights_liberties_min.	49	0,635	26	0,743
social_welfare	62	0,538	10	0,513
technology	25	0,552	9	0,562
transportation	36	0,687	13	0,743
education	21	0,623	17	0,872
energy	44	0,625	14	0,800
environment	59	0,830	16	0,821
health	106	0,777	34	0,944
housing	65	0,827	12	0,828
international_affairs	180	0,612	61	0,884
law_and_crime	137	0,676	43	0,804
macroeconomics	156	0,652	35	0,569
immigration	19	0,728	12	0,923
labor	57	0,732	16	0,780

L'accord moyen vaut 0,609 sur le benchmark à quatre annotateurs et 0,726 sur le projet 57. L'écart tient au nombre d'annotateurs autant qu'à la difficulté des phrases : accorder quatre personnes est mécaniquement plus difficile qu'en accorder deux.

Deux constats se dégagent. D'abord, **la performance des modèles suit celle des humains** : la corrélation entre l'accord inter-annotateurs du projet 57 et le F1 hors échantillon est de 0,644. Les catégories où le modèle échoue sont très largement celles que les annotateurs ne tranchent pas de la même façon, `public_lands` (accord 0,44) et `culture_nationalism` (0,37) en tête. Une bonne part de ce que l'on prendrait pour une faiblesse des modèles est en réalité une ambiguïté du codebook.

Ensuite, rapporté à ce plafond, **le modèle atteint en moyenne 0,93 fois l'accord humain**, et quelques têtes le dépassent nominalement (`macroeconomics`, `social_welfare`). Il faut se garder d'y lire une performance surhumaine : le modèle est évalué contre le *consensus* des annotateurs, cible plus propre que ne l'est la prédiction d'un annotateur individuel pris au hasard. Un ratio supérieur à un signale une catégorie où les annotateurs divergent souvent mais où leur zone d'accord reste, elle, bien caractérisée.

5.5 Résultats des têtes de partis

Quelle colonne lire. Pour les 9 têtes de partis, le F1 du modèle servi est la colonne « **F1 final** » du tableau 10. Ces têtes sont servies *au seuil de 0,5* : aucun seuil calibré n'a été déployé pour elles (voir section 6), si bien que ce F1 est mesuré au point de fonctionnement réel et n'est entaché d'aucun biais de calibration. La dernière colonne, « F1 h.-é. (seuil non déployé) », est l'estimation qu'aurait donnée un seuil calibré : elle explique pourquoi il a été écarté, mais ne décrit pas ce qui est servi.

Table 10 – Performance des têtes de partis sur le benchmark à quatre annotateurs, seul jeu de référence disponible pour cette famille. Le F1 hors échantillon n'est calculé que pour les partis comptant au moins douze positifs.

Parti	Support	Accord	**F1 final**	Précision	Rappel	F1 h.-é. (seuil non déployé)
BQ	16	0,908	0,968	1,000	0,938	0,967
CAQ	8	0,780	0,667	0,714	0,625	n. d.
CPC	73	0,915	0,993	0,987	1,000	0,994
GPC	5	0,857	0,667	0,750	0,600	n. d.
LPC	97	0,840	0,959	0,950	0,969	0,951
NDP	28	0,918	0,949	0,903	1,000	0,934
PLQ	8	0,394	0,222	1,000	0,125	n. d.
PQ	1	1,000	1,000	1,000	1,000	n. d.
QS	1	1,000	1,000	1,000	1,000	n. d.

Les quatre partis fédéraux disposant d'un effectif suffisant obtiennent des scores remarquables, entre 0,93 et 0,99, très au-dessus de ce qu'atteignent les thèmes, et au niveau de l'accord entre annotateurs (0,84 à 0,92). La reconnaissance d'une mention de parti est une tâche largement lexicale, sur laquelle des encodeurs multilingues excellent.

Les cinq autres partis ne peuvent pas être évalués sérieusement : CAQ et PLQ comptent huit positifs de référence, GPC cinq, PQ et QS un seul. Les valeurs figurant au tableau sont données pour information et ne doivent pas être citées comme des mesures de performance. Le cas de `PLQ` mérite une attention particulière : son F1 de 0,222 sur huit cas est cohérent avec la faiblesse déjà observée à l'entraînement, où il est de loin le plus bas du lot, ce que la confusion entre les libéraux fédéraux et provinciaux explique naturellement.

5.6 Le modèle de sentiment

Quelle valeur lire. Pour ce modèle, la performance servie est la ligne « **Macro-F1 du modèle** » du tableau 11 (0,653). Il n'y a pas de seuil à calibrer : la décision est un *argmax* sur les trois classes, donc le chiffre mesuré est directement celui du modèle en production.

La vitrine sert également un modèle de sentiment à trois classes, distinct des précédents à plus d'un titre : il repose sur XLM-RoBERTa et non sur mDeBERTa, il est multi-classe et non binaire, et il est antérieur aux autres (entraîné le 2 février 2026, avec des lots de 256). Ses journaux de session ne se trouvent dans aucun dépôt local : seule la courbe par époque embarquée dans le dossier du modèle a pu être récupérée, ce qui constitue une lacune de traçabilité à combler.

Il a été évalué à travers l'API elle-même, ce qui présente l'avantage de valider l'artefact en production plutôt qu'une copie locale, sur la colonne de sentiment consensuel du benchmark (1 549 phrases : 863 neutres, 534 négatives, 152 positives).

Le résultat est le plus net du rapport : **le modèle est à la hauteur du plafond humain**. Son macro-F1 de 0,653 et son exactitude de 0,699 sont à comparer à un accord entre annotateurs de 0,669 et 0,713 respectivement, soit 98 % du plafond. Sur ce type de tâche, où le jugement humain est lui-même instable, il n'y a plus de marge de progression significative à attendre d'un meilleur modèle : le gain viendrait d'une définition plus stricte des classes. La classe positive, la plus rare,

Table 11 – Modèle de sentiment face au consensus des quatre annotateurs, et accord de ces annotateurs entre eux sur la même échelle.

Classe	Support	Précision	Rappel	F1
negative	534	0,599	0,869	0,709
neutral	863	0,819	0,640	0,718
positive	152	0,670	0,441	0,532
Macro-F1 du modèle		**0,653**		
Exactitude du modèle		0,699		
Accord entre annotateurs (macro-F1)		0,669		
Accord entre annotateurs (exactitude)		0,713		

reste la plus difficile pour le modèle comme pour les annotateurs.

6 Calibration des seuils de décision

6.1 Pourquoi ce seuil

Chaque tête produit une probabilité d'appartenance à la classe positive. La transformer en décision demande un seuil, et le seuil naturel de 0,5 n'a aucune raison d'être le bon. Deux effets s'y opposent.

D'une part, les têtes sont entraînées sur des données très déséquilibrées : elles apprennent à être prudentes, et concentrent leurs probabilités près de zéro. D'autre part, elles sont appliquées à un domaine différent de leur domaine d'entraînement. Un seuil réglé pour le corpus d'origine ne transporte pas tel quel sur les textes de la vitrine.

L'effet est loin d'être marginal : le tableau 12 montre des têtes dont le F1 passe de 0,34 à 0,58 par le seul déplacement du seuil, sans rien changer au modèle (c'est le cas de `technology`).

6.2 Méthode

Le seuil retenu est la **médiane de 200 seuils optimaux**, chacun calculé sur une moitié des phrases de référence tirée au hasard et stratifiée pour conserver la proportion de positifs. Cette médiane est préférée au seuil qui maximiserait directement le F1 sur toute la référence, lequel surapprend les particularités de l'échantillon, d'autant plus que les effectifs sont faibles.

La qualité de la calibration se juge sur l'**estimation hors échantillon** : pour chacun des 200 tirages, le seuil appris sur une moitié est évalué sur l'autre. La moyenne de ces F1 est l'estimation rapportée, et l'écart interquartile des 200 seuils mesure la stabilité de la procédure. Un écart interquartile serré indique un seuil robuste ; un écart large signalerait un réglage fragile, à ne pas déployer.

Un garde-fou est inscrit dans la procédure : les têtes de moins de douze positifs de référence ne sont pas calibrées du tout, leur seuil reste à 0,5, faute de quoi l'on ajusterait un paramètre sur une poignée de cas.

Une seconde règle, celle-là appliquée à la main au moment du déploiement et non automatisée dans le script, veut qu'un seuil calibré ne soit déployé que s'il fait mieux que 0,5 en estimation hors échantillon. C'est ce contrôle qui a conduit à laisser les neuf têtes de partis à 0,5 : leurs seuils calibrés dégradaient le résultat (0,933 contre 0,968 pour `BQ`, par exemple). Le tableau 12 permet de refaire cette comparaison pour chaque tête.

Comment lire ce tableau tête par tête. Trois cas se présentent. Quand seuil servi et seuil calibré coïncident (les onze têtes d'août), la colonne « F1 hors-éch. » est bien l'estimation honnête du point de fonctionnement servi. Quand ils diffèrent (`education`, `energy`, `health`, `housing`,

Table 12 – Seuils servis et seuils calibrés sur la référence poolée. Les colonnes de F1 permettent de mesurer l'apport de la calibration : « F1 à 0,5 » est le point de départ, « F1 servi » ce que produit le seuil actuellement en production mais mesuré sur les phrases qui ont servi à le régler, « F1 hors échantillon » l'estimation honnête du seuil calibré. **Pour les thèmes, le F1 final servi est la colonne « F1 hors-éch. » ; pour les partis, servis à 0,5, c'est la colonne « F1 servi » (identique à « F1 à 0,5 »).** Les têtes de partis figurent après le trait de séparation.

Tête	Seuil servi	Seuil calibré	Écart interq.	F1 à 0,5	F1 servi in-éch.	F1 hors-éch.
agriculture	0,998	0,998	[0,996 ; 0,999]	0,474	0,522	0,508
culture_nationalism	0,992	0,992	[0,989 ; 0,994]	0,301	0,517	0,460
defense	0,928	0,928	[0,901 ; 0,966]	0,533	0,583	0,540
domestic_commerce	0,977	0,977	[0,968 ; 0,982]	0,254	0,371	0,339
foreign_trade	0,998	0,998	[0,997 ; 0,998]	0,681	0,804	0,758
governments_governance	0,957	0,958	[0,943 ; 0,967]	0,446	0,523	0,508
public_lands	0,998	0,998	[0,998 ; 0,999]	0,353	0,437	0,387
rights_liberties_minorities_discrimination	0,992	0,992	[0,992 ; 0,992]	0,559	0,702	0,675
social_welfare	0,990	0,990	[0,988 ; 0,991]	0,549	0,703	0,669
technology	0,999	0,999	[0,999 ; 0,999]	0,337	0,581	0,571
transportation	0,999	0,999	[0,999 ; 0,999]	0,537	0,651	0,647
education	0,945	0,974	[0,974 ; 0,974]	0,618	0,632	0,684
energy	0,654	0,137	[0,038 ; 0,654]	0,757	0,745	0,741
environment	0,981	0,808	[0,092 ; 0,963]	0,864	0,861	0,855
health	0,966	0,143	[0,143 ; 0,854]	0,853	0,833	0,845
housing	0,993	0,896	[0,894 ; 0,989]	0,863	0,859	0,856
international_affairs	0,913	0,978	[0,864 ; 0,982]	0,796	0,799	0,793
law_and_crime	0,245	0,361	[0,361 ; 0,649]	0,774	0,757	0,762
macroeconomics	0,623	0,822	[0,604 ; 0,880]	0,734	0,749	0,730
immigration	0,006	0,019	[0,003 ; 0,178]	0,686	0,703	0,677
labor	0,603	0,998	[0,998 ; 0,999]	0,724	0,724	0,780
BQ	0,500	0,500	[0,500 ; 0,995]	0,968	0,968	0,967
CAQ	0,500	0,500	–	0,667	0,667	n. d.
CPC	0,500	0,500	[0,500 ; 0,500]	0,993	0,993	0,994
GPC	0,500	0,500	–	0,667	0,667	n. d.
LPC	0,500	0,002	[0,002 ; 0,004]	0,959	0,959	0,951
NDP	0,500	0,394	[0,394 ; 0,998]	0,949	0,949	0,934
PLQ	0,500	0,500	–	0,222	0,222	n. d.
PQ	0,500	0,500	–	1,000	1,000	n. d.
QS	0,500	0,500	–	1,000	1,000	n. d.

`international_affairs`, `law_and_crime`, `macroeconomics`, `immigration`, `labor`), le seuil servi date d'une calibration antérieure : la colonne « F1 hors-éch. » reste la meilleure estimation non biaisée de la performance de ces têtes, l'écart entre les deux seuils étant du même ordre que la dispersion interquartile. Enfin, pour les neuf têtes de partis, aucun seuil calibré n'a été déployé : leur F1 final est la colonne « F1 servi ».

6.3 Lecture des seuils

Les seuils des têtes thématiques ré-entraînées en août sont très élevés, de 0,93 à 0,999. Cela ne traduit pas une exigence particulière du déploiement mais la forme de la distribution des scores : ces modèles sont extrêmement confiants, et l'essentiel de la discrimination utile se joue dans le dernier centième de l'échelle de probabilité. Leurs écarts interquartiles restent serrés, ce qui indique un réglage stable malgré des valeurs extrêmes.

Un seuil très bas doit en revanche attirer l'attention. La tête `immigration` est servie à 0,006 : ce réglage est justifié par une distribution de scores fortement bimodale, mais il est intrinsèquement fragile, puisqu'une faible dérive du modèle ou du domaine suffirait à le rendre inadapté. Il devra être recalibré à la prochaine campagne d'annotation.

Un point de vigilance de méthode. Les seuils servis par l'API ont été calibrés sur des jeux de référence qui sont, pour partie, ceux-là mêmes sur lesquels ce rapport mesure la performance. La colonne « F1 hors échantillon » corrige ce biais par construction, et c'est elle qu'il faut citer ; la colonne « F1 servi », elle, est optimiste de 0,013 en moyenne. Une campagne d'annotation indépendante, calibrée sur un jeu et mesurée sur un autre, reste la façon propre de trancher, et devrait être la priorité de la prochaine vague.

7 Limites et recommandations

7.1 Trois niveaux de vigilance

Les 30 modèles servis n'appellent pas la même prudence. Le tableau 13 les répartit en trois niveaux. Le classement est produit par le script à partir de règles fixées d'avance, et non d'une appréciation portée au cas par cas : une tête relève du niveau 1 si elle compte moins de douze positifs de référence ou si son F1 final est inférieur à 0,45 ; du niveau 2 si ce F1 reste sous 0,70 ; du niveau 3 si, la performance étant acquise, sa calibration est instable — écart interquartile supérieur à 0,10, ou seuil servi distant de plus de 0,05 du seuil calibré. Les têtes de partis sont exclues de ce dernier critère, puisqu'elles sont servies à 0,5 et qu'aucun seuil calibré n'a été déployé pour elles.

Niveau 1 : 7 modèles à ne pas exposer comme filtres. Deux motifs distincts s'y croisent. Cinq têtes de partis (CAQ, GPC, PLQ, PQ, QS) n'ont pas assez de cas de référence pour qu'on sache quoi que ce soit de leur comportement : les 1,000 affichés par PQ et QS reposent sur un seul cas et ne sont pas des mesures. Le cas de `PLQ` est le plus préoccupant du rapport : son rappel de 0,125 signifie qu'il manque sept mentions du Parti libéral du Québec sur huit, et sa faiblesse était déjà la plus marquée à l'entraînement, ce que la confusion avec les libéraux fédéraux explique. Trois partis québécois sur quatre sont ainsi hors d'état d'être utilisés, ce qui pèse lourd pour un corpus québécois. Les deux thèmes du niveau, `domestic_commerce` et `public_lands`, remontent près de sept phrases fausses sur dix (précision 0,316 et 0,311) : ils peuvent signaler, non sélectionner.

Niveau 2 : 10 modèles utilisables sous réserve explicite. Leur F1 final se situe entre 0,46 et 0,69, et leur précision souvent sous 0,55 : tout dénombrement fondé sur eux doit être présenté comme un ordre de grandeur. Deux cas y méritent mieux qu'une lecture de tableau. D'abord `governments_governance`, dont le F1 de 0,508 est médiocre alors qu'il s'agit de **la catégorie la plus fréquente** du corpus (244 positifs de référence, 4 956 exemples d'entraînement) : à précision

Table 13 – Les modèles appelant une vigilance particulière, par niveau. « F1 final » est la colonne de référence définie en section 1.3 ; « Préc. » est la précision au seuil servi, qui dit quelle part des phrases remontées est un faux positif. Les modèles absents du tableau ne relèvent d'aucun de ces trois niveaux.

Modèle	Sup.	F1 final	Préc.	Motif
Niveau 1 — ne pas exposer comme filtre				
PLQ	8	0,222	1,000	non validé (8 cas de référence), rappel 0,125
domestic_commerce	40	0,339	0,316	F1 0,339, précision 0,316
public_lands	26	0,387	0,311	F1 0,387, précision 0,311
CAQ	8	0,667	0,714	non validé (8 cas de référence)
GPC	5	0,667	0,750	non validé (5 cas de référence)
PQ	1	1,000	1,000	non validé (1 cas de référence)
QS	1	1,000	1,000	non validé (1 cas de référence)
Niveau 2 — usage sous réserve explicite				
culture_nationalism	30	0,460	0,536	F1 0,460, précision 0,536
agriculture	20	0,508	0,367	F1 0,508, précision 0,367
governments_governance	244	0,508	0,457	F1 0,508, précision 0,457
defense	42	0,540	0,492	F1 0,540, précision 0,492
technology	34	0,571	0,458	F1 0,571, précision 0,458
transportation	49	0,647	0,525	F1 0,647, précision 0,525
social_welfare	72	0,669	0,699	F1 0,669, précision 0,699
rights_liberties_minorities_discrimination	75	0,675	0,634	F1 0,675, précision 0,634
immigration	31	0,677	0,605	F1 0,677, précision 0,605
education	38	0,684	0,526	F1 0,684, précision 0,526
Niveau 3 — performance acquise, calibration fragile				
macroeconomics	191	0,730	0,678	écart interq. 0,277, seuil servi 0,623 ≠ calibré 0,822
energy	58	0,741	0,646	écart interq. 0,617, seuil servi 0,654 ≠ calibré 0,137
law_and_crime	180	0,762	0,743	écart interq. 0,288, seuil servi 0,245 ≠ calibré 0,361
labor	73	0,780	0,624	seuil servi 0,603 ≠ calibré 0,998
international_affairs	241	0,793	0,730	écart interq. 0,118, seuil servi 0,913 ≠ calibré 0,978
health	140	0,845	0,846	écart interq. 0,711, seuil servi 0,966 ≠ calibré 0,143
environment	75	0,855	0,819	écart interq. 0,872, seuil servi 0,981 ≠ calibré 0,808
housing	77	0,856	0,889	seuil servi 0,993 ≠ calibré 0,896

0,457, c'est elle qui injectera le plus grand *nombre absolu* de faux positifs dans la vitrine, un mauvais F1 sur une catégorie fréquente coûtant davantage que le même F1 sur une catégorie rare. Ensuite `agriculture`, qui sur-prédit massivement (rappel 0,900 pour une précision de 0,367).

Niveau 3 : 8 modèles performants dont la calibration est fragile. Ici le risque ne porte pas sur la performance mesurée, qui va de 0,73 à 0,86, mais sur sa tenue dans le temps. Les seuils de `energy`, `health` et `environment` varient énormément d'un tirage de calibration à l'autre (écarts interquartiles de 0,617, 0,711 et 0,872) et diffèrent nettement du seuil effectivement servi : leur bon F1 tient à une distribution de scores qu'une dérive du domaine suffirait à déplacer. Le cas de `immigration`, servi à 0,006, relève de la même fragilité et figure au niveau 2 pour son F1. Deux têtes enfin, `labor` et `education`, sont servies à un seuil *moins bon* que leur seuil calibré (0,724 contre 0,780 ; 0,632 contre 0,684) : les recalibrer est un gain immédiat et sans coût.

7.2 Deux limites qui ne tiennent pas aux modèles

Une partie de l'écart au parfait est une ambiguïté du codebook, mais pas toute. La corrélation entre l'accord inter-annotateurs et le F1 des modèles (0,644) dit que les catégories difficiles pour les modèles sont largement celles que les humains eux-mêmes ne tranchent pas de la même façon — `public_lands` et `culture_nationalism` en sont les exemples nets. Mais la règle souffre des exceptions qui désignent, elles, un vrai défaut de modèle : `domestic_commerce`, `agriculture`, `defense`, `immigration` et `education` portent sur des catégories que les annotateurs s'accordent à reconnaître (0,67 à 0,92 d'accord) et où le modèle décroche pourtant. Ce sont les meilleures candidates à un ré-entraînement ; `culture_nationalism`, elle, ne gagnerait rien sans une définition plus stricte de la catégorie.

Le sentiment sous-détecte la classe positive. Son rappel de 0,441 sur les phrases positives signifie qu'il en manque plus d'une sur deux. Le modèle est au niveau du plafond humain (section 5.6) et il n'y a donc guère de marge à attendre d'un meilleur modèle, mais tout comptage de phrases positives issu de la vitrine sera systématiquement sous-estimé.

7.3 Reste à faire

1. **Terminer la campagne d'annotation des partis** (projet doccano 58 : 885 phrases chargées, 48 confirmées). Deux annotateurs sur l'ensemble suffiraient à faire passer cinq têtes de l'état « non validé » à l'état mesuré, et donneraient enfin une référence propre pour PLQ, la tête la plus fragile de la famille.

A Annexes

A.1 Contrôles automatiques

L'extraction des données d'entraînement vérifie sa propre sortie. Pour chacun des 30 modèles, quatre contrôles sont exécutés : que le macro-F1 annoncé par les métadonnées des poids déployés correspond exactement à celui consigné dans le journal de la session à l'époque retenue, que le garde-fou d'export a été exécuté, que les métriques par langue sont disponibles pour l'époque retenue, et que le jeu d'entraînement référencé est lisible et complet. Un cinquième contrôle, par session, vérifie que les têtes d'une même vague ont été entraînées sur des volumes cohérents.

Table 14 – Résultat des 124 contrôles automatiques.

Contrôle	Résultat	Occurrences
garde-fou d'export	ABSENT	8
garde-fou d'export	OK	22
jeu d'entrainement lisible	OK	30
macro-F1 du checkpoint == journal de session	OK	30
métriques par langue disponibles	OK	30
volume identique pour toutes les tetes	ECART	1
volume identique pour toutes les tetes	OK	3

Trois résultats méritent une explication, car ils ne sont pas des erreurs.

Le garde-fou d'export est signalé *absent* pour les huit têtes de la vague de juin : ce mécanisme a été ajouté à LLM_Tool le 13 juillet 2026, après leur entraînement. Leur checkpoint n'a donc pas été vérifié automatiquement au moment de la sauvegarde. C'est un fait de traçabilité, non un défaut constaté : ces têtes obtiennent par ailleurs de bons résultats en validation externe, ce qui exclut une corruption du type de celle qui avait frappé `immigration` en juin.

Le contrôle de volume signale un *écart* sur la session d'août : c'est attendu, puisque le régime équilibré construit un jeu de taille différente pour chaque tête (section 3). L'écart est ici le comportement correct, et son absence aurait signalé un problème.

Le contrôle des métriques par langue signale enfin que deux têtes, `labor` et `macroeconomics`, ne portent pas le fichier d'historique par langue à côté de leurs poids : elles ont été déposées depuis un dossier d'époque intermédiaire, qui ne le contient pas. Leurs valeurs par langue sont donc reprises du journal de session, qui porte exactement les mêmes colonnes. Le champ `source_metriques_langue` de `training.json` indique, pour chaque tête, laquelle des deux sources a été utilisée.

A.2 Inventaire des journaux

Le dossier `logs/` contient les journaux des quatre sessions d'entraînement et les jeux de données correspondants, soit 126 fichiers pour environ 150 Mo. Chaque session y suit la même organisation :

— `session_metadata.json` : configuration complète de la session (modèle de base, hyperparamètres, découpage, machine, version de l'outil, statistiques du corpus) ;

— `metriques/<tête>/training.csv` : le journal par époque, avec pertes, exactitude, précision, rappel et F1 par classe et par langue. C'est la source de toutes les courbes citées ;

— `donnees/` : les synthèses de découpage et de distribution produites par l'outil (complètes pour les sessions d'août, réduites au journal d'avertissements pour celles de juin et juillet, qui sont antérieures à cette fonctionnalité) ;

— `jeux/<tête>.csv.gz` : le jeu d'entraînement exact de chaque tête, compressé.

Le fichier `logs/MANIFEST.json` donne pour chacun sa taille et son empreinte SHA-256, ainsi que celle du fichier original avant compression : un lecteur peut donc vérifier n'importe quel élément sans accès à la machine d'entraînement.

A.3 Données de référence et résultats bruts

— `data/gold/` : les deux jeux annotés, dans leur forme exportée (le benchmark en CSV avec les quatre annotations individuelles et le consensus ; le projet doccano en JSON avec, pour chaque phrase, les étiquettes de chaque annotateur et la liste de ceux qui l'ont confirmée) ;

— `data/extract/training.json` : toutes les données d'entraînement consolidées, y compris les courbes complètes par époque ;

— `data/extract/evaluation.json` : toutes les métriques, par modèle, par jeu et par point de fonctionnement ;

— `data/extract/scores_bruts.npz` : les probabilités brutes produites par chaque modèle sur chaque phrase de référence, avec les empreintes des poids qui les ont produites, de sorte que toute nouvelle métrique puisse être recalculée sans réexécuter les modèles ;

— `data/extract/sentiment_*.json` : métadonnées, prédictions et métriques du modèle de sentiment ;

— `data/extract/server_metadata.json` : l'état déclaré par l'API pour chaque modèle servi, dont les seuils.

A.4 Reproduire les résultats

Depuis la racine du dossier, avec un interpréteur disposant de NumPy :

```
python3 scripts/01_extract_training.py
python3 scripts/02_evaluate_fleet.py
python3 scripts/04_make_tables.py
latexmk -pdf main.tex
```

La deuxième étape recalcule les métriques en quelques secondes depuis les scores bruts archivés avec l'empreinte des poids qui les ont produits. Si ce cache est absent ou incompatible, elle rejoue les modèles, ce qui demande environ une heure ainsi que *torch*, *transformers* et les poids locaux. L'option `-verify-weights` permet en plus de comparer les empreintes archivées aux copies locales et de rejouer tout modèle modifié. Le modèle de sentiment n'existant que sur le serveur, son évaluation passe par les scripts 05 et 06, qui interrogent l'API.

A.5 Glossaire des métriques

Précision
Parmi les phrases où le modèle annonce la catégorie, la part où elle est effectivement présente. Une précision basse produit du bruit.

Rappel
Parmi les phrases où la catégorie est présente, la part que le modèle retrouve. Un rappel bas produit des silences.

F1
Moyenne harmonique des deux, qui ne peut être élevée que si aucune des deux n'est basse. Sauf mention contraire, le F1 de ce rapport est celui de la classe positive.

Macro-F1
Moyenne du F1 de la classe positive et de celui de la classe négative. Sur des données déséquilibrées, il est systématiquement plus flatteur que le F1 de la classe positive.

Support
Nombre de phrases où la catégorie est réellement présente. C'est l'indicateur de fiabilité d'une ligne : en dessous d'une dizaine, aucune conclusion n'est solide.

Accord entre annotateurs
Même calcul que le F1, appliqué à deux annotateurs humains l'un contre l'autre, moyenné sur toutes les paires. Il donne le plafond réaliste de la tâche.

F1 hors échantillon
F1 obtenu quand le seuil est réglé sur une moitié des données et mesuré sur l'autre, moyenné sur 200 tirages. C'est l'estimation honnête, et celle à citer.