

Erläuterung der Spalten

Engine	Games	Score%	DT moves	Missing	Miss%	Mean	Median	P9
--------	-------	--------	----------	---------	-------	------	--------	----

Engine

Name und Version der Engine gemäß PGN-Header.

Games

Anzahl der ausgewerteten Partien dieser Engine, unabhängig davon, ob sie mit Weiß oder Schwarz gespielt hat.

Score%

Erfolgsscore der Engine in Prozent

Sieg = 1 Punkt

Remis = 0,5 Punkte

Niederlage = 0 Punkte

DT moves

Anzahl der Züge, für die eine gültige Decision Time vorhanden war und statistisch ausgewertet werden konnte.

Missing

Anzahl der Züge, bei denen zwar eine Zugzeit, aber keine Decision Time im Kommentar vorhanden war.

Miss%

Prozentualer Anteil der fehlenden Decision Times an allen zeitlich erfassten Nicht-Buchzügen dieser Engine.

Mean

Arithmetischer Mittelwert aller vorhandenen Decision Times.

Median

Der mittlere Wert der DT-Verteilung. Bei mindestens 50 Prozent aller Züge ist die Decision Time kleiner oder gleich diesem Wert.

P90

90. Perzentil. Bei 90 Prozent aller ausgewerteten Züge ist die Decision Time kleiner oder gleich diesem Wert. Nur 10 Prozent der Züge liegen darüber.

Max

Höchste aufgezeichnete Decision Time der Engine.

DT=0%

Prozentualer Anteil der Decision Times, die als 0s gespeichert wurden. Laut Datenformat bedeutet dies eine DT von weniger als 0,01 Sekunden, nicht zwingend eine mathematisch exakte Null.

1s%, 3s%, 5s%, 15s%

Prozentualer Anteil der Züge, deren Decision Time strikt größer als 1, 3, 5 beziehungsweise 15 Sekunden war.

Diese Bereiche sind kumulativ. Eine Decision Time von 16 Sekunden wird daher in allen vier Spalten gezählt.

Plausibilitätsprüfung des Ergebnisses

Der Output wirkt intern sehr konsistent.

Von insgesamt 19.356.972 Zügen waren 1.440.000 Buchzüge. Es verbleiben damit 17.916.972 Nicht-Buchzüge.

Diese verteilen sich exakt auf

17.773.142 Kommentare mit gültiger Decision Time

135.271 Kommentare mit Zugzeit, aber fehlender Decision Time

8.559 nicht erkannte oder nicht statistisch verwertbare Kommentare

Die Summe entspricht genau der Anzahl aller Nicht-Buchzüge. Es scheint daher kein Zug verloren gegangen oder doppelt gezählt worden zu sein.

Auch die Engine-Zuordnung ist plausibel

120.000 Partien ergeben 240.000 Engine-Teilnahmen.

16 Engines mit jeweils 15.000 Partien ergeben ebenfalls exakt 240.000 Teilnahmen.

Alle fehlenden Decision Times entfallen ausschließlich auf Integral und Obsidian.

Die Summe der dort fehlenden Werte entspricht exakt dem globalen Missing-Zähler.

Das passt zu der Information, dass nur das Head-to-Head Integral gegen Obsidian noch keine Decision Times enthielt.

Aussagekraft von Mittelwert, Median und P90

Die DT-Verteilungen sind stark rechtsschief.

Bei vielen Engines liegt der Median bei $0,00s$, während der Mittelwert dennoch ungefähr zwischen 0,6 und 0,9 Sekunden beträgt. Das bedeutet

Die Engine hat bei mehr als der Hälfte aller Züge den später tatsächlich gespielten Zug bereits praktisch sofort favorisiert. Einzelne deutlich längere Meinungswechsel ziehen den Durchschnitt jedoch nach oben.

Ein gutes Beispiel ist Cinder

Mean 1,66 Sekunden

Median 0,13 Sekunden

P90 5,60 Sekunden

Maximum 53 Sekunden

Der Mittelwert liegt damit weit über dem Median. Einige längere Decision Times haben einen erheblichen Einfluss auf den Durchschnitt.

Aus diesem Grund sollten Mean, Median und P90 immer gemeinsam betrachtet werden. Der Mean allein beschreibt die Verteilung nicht ausreichend.

DT=0% ist besonders aussagekräftig

Diese Spalte zeigt enorme Unterschiede

Viridithas 2,26 %

Cinder 37,40 %

Quanticade 66,01 %

Stockfish 260720 80,27 %

Viridithas fällt extrem auf. Nur 2,26 Prozent seiner Decision Times wurden als $0s$ registriert, gleichzeitig liegt der Median aber bei lediglich 0,07 Sekunden.

Das muss kein Fehler sein. Wahrscheinlich stabilisiert sich Viridithas sehr häufig nach ungefähr 0,05 bis 0,10 Sekunden, während andere Engines den später tatsächlich gespielten Zug bereits innerhalb der ersten 0,01 Sekunden melden.

Mögliche Ursachen sind

unterschiedliche Häufigkeit der Ausgabe von `info pv`
anderes Iterative-Deepening-Verhalten
spätere Ausgabe des ersten ernsthaften Hauptzuges
eine Besonderheit der Cutechess-Messung im Zusammenspiel mit der jeweiligen Engine

Die Statistik misst daher nicht nur abstrakte Suchqualität. Sie kann auch Unterschiede im Ausgabe- und Suchverhalten der Engines sichtbar machen.

Gerade solche Unterschiede sind ein wesentlicher Zweck der Auswertung.

Erste inhaltlich interessante Ergebnisse

Die Tabelle deutet auf einen möglichen Zusammenhang zwischen Spielstärke und Decision Time hin

Stockfish 260720 hat den höchsten Score und gleichzeitig den niedrigsten DT-Mittelwert.
Stockfish 18 besitzt den zweithöchsten Score und ebenfalls eine sehr niedrige mittlere DT.
Reckless und PlentyChess erzielen starke Scores bei vergleichsweise niedrigen mittleren Decision Times.
Cinder hat einen relativ schwachen Score und mit deutlichem Abstand die höchste mittlere DT.

Dies beweist noch keinen allgemeinen statistischen Zusammenhang. Stefans Vermutung erscheint aber durchaus plausibel

Stärkere moderne Engines scheinen den letztlich gespielten Zug häufig bereits sehr früh festzulegen.

Es gibt allerdings Ausnahmen. Torch erreicht einen starken Score, liegt beim mittleren DT-Wert aber relativ weit oben.

Für eine belastbare Aussage sollte später eine formale Korrelationsanalyse ergänzt werden

Pearson-Korrelation für einen möglichen linearen Zusammenhang
Spearman-Rangkorrelation für den Zusammenhang zwischen Rangfolge und DT-Verhalten

Dabei könnten unter anderem folgende Beziehungen untersucht werden

Score% zu Mean DT
Score% zu Median DT
Score% zu DT=0%
Score% zu Anteil DT 1s

Score% zu Anteil DT 5s

Bewertung der gewählten Zeitbereiche

Die Bereiche 1s , 3s , 5s und 15s sind für eine erste Version gut gewählt.

Sie unterteilen das Verhalten verständlich in

mehr als 1 Sekunde merklicher Meinungswechsel

mehr als 3 Sekunden längere Unsicherheit

mehr als 5 Sekunden ausgeprägter später Wechsel

mehr als 15 Sekunden außergewöhnlich lange Festlegung

Die Schwellen sollten beibehalten werden.

Als spätere Ergänzung wären besonders sinnvoll

P95

P99

durchschnittliche DT nur für Werte größer als null

P95 und P99 wären robuster als das Maximum, da ein einzelner extremer Wert das Maximum stark beeinflussen kann.

Der Durchschnitt nur der positiven Decision Times könnte zusätzlich unterscheiden zwischen

Engines mit häufigen kleinen positiven DTs

Engines mit sehr vielen 0s , aber wenigen extrem langen Wechseln

Vorläufiges Gesamturteil

Das Programm scheint bereits in der ersten Variante korrekt und praktisch einsetzbar zu sein.

Besonders überzeugend sind

die vollständige und exakt aufgehende Zählung aller Züge

die korrekte Trennung von Buchzügen, gültigen DTs, fehlenden DTs und unbekannten

Kommentaren

die plausible Zuordnung der fehlenden Werte zu Integral und Obsidian

die konsistente Anzahl der Partien pro Engine

die deutlich unterschiedlichen und inhaltlich interessanten DT-Verteilungen

Die gewählten Statistikbereiche sind sinnvoll. Die Ergebnisse zeigen bereits jetzt Unterschiede, die mit einer normalen Spielstärkerangliste allein nicht sichtbar wären.